Generalized Empirical Likelihood Estimators and Tests under Partial, Weak and Strong Identification

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

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

First Draft: June 2002; This revision: October 2004
Abstract

The purpose of this paper is to describe the performance of generalized empirical likelihood (GEL) methods for time series instrumental variable models specified by nonlinear moment restrictions as in Stock and Wright (2000) when identification may be weak. The paper makes two main contributions. Firstly, we show that all GEL estimators are first-order equivalent under weak identification. The GEL estimator under weak identification is inconsistent and has a nonstandard asymptotic distribution. Secondly, the paper proposes new GEL test statistics, which have chi-square asymptotic null distributions independent of the strength or weakness of identification. Consequently, unlike those for Wald and likelihood ratio statistics, the size of tests formed from these statistics is not distorted by the strength or weakness of identification. Modified versions of the statistics are presented for tests of hypotheses on parameter subvectors when the parameters not under test are strongly identified. Monte Carlo results for the linear instrumental variable regression model suggest that tests based on these statistics have very good size properties even in the presence of conditional heteroskedasticity. The tests have competitive power properties, especially for thick tailed or asymmetric error distributions.

JEL Classification: C12, C31, C32

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

*This paper is a revision of Guggenberger’s job market paper “Generalized Empirical Likelihood Tests under Partial, Weak, and Strong Identification”. We are thankful to the Editor, P.C.B. Phillips, and three referees for very helpful suggestions on an earlier version of this paper. Guggenberger gratefully acknowledges the continuous help and support of his advisor Donald Andrews who played a prominent role in the formulation of this paper. He thanks Peter Phillips and Joseph Altonji for their extremely valuable comments. We also thank Vadim Marner for help with the simulation section and John Chao, Guido Imbens, Michael Jansson, Frank Kleibergen, Marcelo Moreira, Jonathan Wright and Motohiro Yogo for helpful comments. Aspects of this research have been presented at the 2003 Econometric Society European Meetings, York Econometrics Workshop 2004, Seminaire Malinvaud, CREST-INSEE, and seminars at Albany, Alicante, Austin Texas, Brown, Chicago, Chicago GSB, Harvard/MIT, Irvine, Konstanz, Laval, Madison Wisconsin, Mannheim, Maryland, Penn, NYU, PennState, Pittsburgh, Princeton, Rice, Riverside, Rochester, San Diego, Texas A&M, UCLA, USC and Yale and we would like to thank all the participants. Guggenberger and Smith received financial support through a Carl Arvid Anderson Prize Fellowship and a 2002 Leverhulme Major Research Fellowship, respectively.

**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

It is often the case that the instrumental variables available to empirical researchers are only weakly correlated with the endogenous variables. That is, identification is weak. Phillips (1989), Nelson and Startz (1990) and a large literature following these early contributions show that in such situations classical normal and chi-square asymptotic approximations to the finite-sample distributions of instrumental variables estimators and statistics can be very poor. For example, even though likelihood ratio and Wald test statistics are asymptotically chi-square, use of chi-square critical values can lead to extreme size distortions in finite samples, see Dufour (1997). The purpose then of this paper is to ascertain the performance of generalized empirical likelihood (GEL) methods [Newey and Smith (2004), henceforth NS, Smith (1997, 2001)] for time series instrumental variable models specified by nonlinear moment restrictions when identification may be weak [as in Stock and Wright (2000), henceforth SW]. In particular, the paper makes two principal contributions. Firstly, the asymptotic distribution of the GEL estimator is derived for a weakly identified set-up. Secondly, the paper proposes new, theoretically and computationally attractive GEL test statistics. The asymptotic null distribution of these statistics is chi-square under partial [Phillips (1989)], weak [SW] and strong identification. Thus, the size of tests formed from these statistics is invariant to the strength or weakness of identification. Importantly, we also provide asymptotic power results for the various statistics suggested in this paper.

GEL estimators and test statistics are alternatives to those based on generalized method of moments (GMM); see Hansen (1982), Newey (1985) and Newey and West (1987). GEL has received considerable attention recently due to its competitive bias properties. For example, NS show that for many models the asymptotic bias of empirical likelihood (EL) does not grow with the number of moment restrictions, while that of GMM estimators grows without bound, a finding that may imply favorable properties for GEL-based test statistics.

Similar to the findings in Phillips (1984, 1989) and SW for limited information maximum likelihood (LIML), two stage least squares (2SLS) and GMM, GEL estimators of weakly identified parameters have nonstandard asymptotic distributions and are in general inconsistent. Therefore, inference based on the classical normal approximation is inappropriate under weak identification. As in NS for strong identification, the first-order asymptotics of the GEL estimator under weak identification do not depend on the choice of the GEL criterion function. This finding is rather surprising and contrasts with 2SLS and LIML estimators whose first-order asymptotic theory differs under weak identification.

The statistics proposed here are asymptotically pivotal in contrast to classical Wald and likelihood ratio statistics no matter the strength of identification. The first statistic $GELR_\rho$ is based on the GEL criterion function and may be thought of as a nonparametric likelihood ratio statistic. Two further statistics generalize Kleibergen’s (2001) GMM-based $K$-statistic to the GEL context. Like the $K$-statistic which is a quadratic form in the first order derivative vector of the continuous updating GMM objective function, the second GEL statistic $S_\rho$ is a score-type statistic being
a quadratic form in the GEL criterion score vector. The third statistic $LM_{\rho}$ is similar in structure to a GMM Lagrange multiplier statistic [Newey and West (1987)] and is asymptotically equivalent to the score-type statistic being a quadratic form in the sample moment vector. Confidence regions constructed from the $K$- and GEL score-type statistics are never empty and contain the continuous updating estimator (CUE) and GEL estimator respectively. All forms of GEL statistics admit limiting chi-square null distributions with degrees of freedom equal to the number of instrumental variables or moment conditions for the first statistic and the dimension of the parameter vector for the second and third statistics. In over-identified situations, therefore, tests based on the latter statistics should be expected to have better power properties than those based on the former. In many cases, an applied researcher is interested in inference on a parameter subvector rather than the whole parameter vector. Modified versions of these statistics are therefore suggested for the subvector case when the remaining parameters are strongly identified.

Monte Carlo simulations for the i.i.d. linear instrumental variable model with a wide range of error distributions compare our test statistics to several others, including homoskedastic and heteroskedastic versions of the $K$-statistic of Kleibergen (2001, 2002a) and the similar conditional likelihood ratio statistic $LR_{M}$ of Moreira (2003). We find that our tests have very good size properties even in the presence of conditional heteroskedasticity. In contrast, the homoskedastic version of the $K$-statistic of Kleibergen (2002a) and the $LR_{M}$-statistic of Moreira (2003) are size-distorted under conditional heteroskedasticity. Our tests have competitive power properties, especially for thick-tailed or asymmetric error distributions. Given the nonparametric construction of the GEL estimator, robustness of GEL-based test statistics to different error distributions should be expected.

Like SW, our paper allows for both i.i.d. and martingale difference sequences (m.d.s.) but does not apply to more general time series models, see assumption $M_{\theta}$(ii) below. Allowing for m.d.s. observations covers various cases of intertemporal Euler equations applications and regression models with m.d.s. errors. Therefore, the extension from the i.i.d. linear (dealt with in the first chapter of the dissertation of the first author, Guggenberger (2003)) to the particular time series setting with nonlinear moment restrictions considered here seems worthwhile, especially since there is essentially no cost (in terms of complications of the proofs) to making this extension. The proofs for consistency and for the asymptotic distribution of the GEL estimator build on Guggenberger (2003) which adapt those given in NS for the i.i.d. strongly identified context.

Subsequent to the i.i.d. linear version of this paper, two related papers have appeared. Firstly, Caner (2003) derives the asymptotic distribution of the exponential tilting (ET) estimator [see Imbens, Spady and Johnson (1998) and Kitamura and Stutzer (1997)] under weak identification with nonlinear moment restrictions for independent observations. Caner (2003) also obtains an ET version of the $K$-statistic for nonlinear moment restrictions. Secondly, Otsu (2003) considers GEL based tests under weak identification in a more general time series setting than considered here and examines the GEL criterion function statistic $GEL_{\rho}$ and a modified version of
the $K$-statistic based on Kitamura and Stutzer's (1997) and Smith's (2001) kernel smoothed GEL estimator that is efficient under strong identification.

The remainder of the paper is organized as follows. In Section 2, the model and the assumptions are discussed, the GEL estimator is briefly reviewed and the asymptotic distribution of the GEL estimator under weak identification is derived. Section 3 introduces the GEL-based test statistics. We derive their asymptotic limiting distribution and show that it is unaffected by the degree of identification. Section 4 generalizes these results to hypotheses involving subvectors of the unknown parameter vector. Section 5 describes the simulation results. All proofs are relegated to the Appendix.

The following notation is used in the paper. The symbols “$\rightarrow_d$”, “$\rightarrow_p$” and “$\Rightarrow$” denote convergence in distribution, convergence in probability and weak convergence of empirical processes, respectively. For the latter, see Andrews (1994) for a definition. For convergence “almost surely” we write “a.s.” and “with probability approaching 1” is replaced by “w.p.a.1”.

The space $C^i(M)$ contains all functions that are $i$ times continuously differentiable on $M$. For a symmetric matrix $A$, “$A > 0$” means that $A$ is positive definite and $\lambda_{\text{min}}(A)$ and $\lambda_{\text{max}}(A)$ denote the smallest and largest eigenvalue of $A$ in absolute value, respectively. By “$A^\prime$” we denote the transpose of a matrix $A$. For a full column rank matrix $A \in R^{k \times p}$ and positive definite matrix $K \in R^{k \times k}$, we denote by $P_A(K)$ the oblique projection matrix $A(A^\prime K^{-1}A)^{-1}A^\prime K^{-1}$ on the column space of $A$ in the metric $K$ and define $M_A(K) := I_k - P_A(K)$, where $I_k$ is the $k$-dimensional identity matrix; we abbreviate this notation to $P_A$ and $M_A$ if $K = I_k$. The symbol “$\otimes$” denotes the Kronecker product. Furthermore, $vec(M)$ stands for the column vectorization of the $k \times p$ matrix $M$, i.e. if $M = (m_1, \ldots, m_p)$ then $vec(M) = (m_1^\prime, \ldots, m_p^\prime)^\prime$. By $|| \cdot ||$ we denote the Euclidean norm.

2 Estimation

This section is concerned with the asymptotic distribution of the GEL estimator when some elements of the parameter vector of interest may be only weakly identified. Intuitively, then, the moment conditions which define the model may not be particularly informative about these parameters.

2.1 Model

We consider models specified by a finite number of moment restrictions. Let $\{z_i : i = 1, \ldots, n\}$ be $R^d$-valued data and, for each $n \in N$, $g_n : G \times \Theta \rightarrow R^k$ a given function, where $G \subset R^d$ and $\Theta \subset R^p$ denotes the parameter space. The model has a true parameter $\theta_0$ for which the moment condition

$$Eg_n(z_i, \theta_0) = 0 \quad (2.1)$$

is satisfied. For $g_n(z_i, \theta)$ we will usually write $g_i(\theta)$. [3]
Example 1: (I.i.d. Linear Instrumental Variable (IV) Regression.) Guggenberger (2003, first chapter) discusses in detail GEL estimation and testing for this model under weak identification. The structural form (SF) equation is given by

\[ y = Y\theta_0 + u, \quad (2.2) \]

and the reduced form (RF) for \( Y \) by

\[ Y = Z\Pi + V, \quad (2.3) \]

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 matrix \( Y \) may contain both exogenous and endogenous variables, \( Y = (X,W) \) say, where \( X \in \mathbb{R}^{n \times p_X} \) and \( W \in \mathbb{R}^{n \times p_W} \) denote the respective observation matrices of exogenous and endogenous variables. The variables \( Z = (X,Z_w) \) constitute a set of instruments for the endogenous variables \( W \). The first \( p_X \) columns of \( \Pi \) equal the first \( p_X \) columns of \( I_k \) and the first \( p_X \) columns of \( V \) are 0. 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. Assuming the instruments and the structural error are uncorrelated, \( E(u_i|Z_i) = 0 \), it follows that \( E(g_i(\theta_0)) = 0 \), where for each \( i = 1, ..., n \), \( g_i(\theta) := (y_i - Y_i'|\theta)Z_i \). Note that in this example \( g_i(\theta) \) depends on \( n \) if the reduced form coefficient matrix \( \Pi \) is modeled to depend on \( n \), see Staiger and Stock (1997), where \( \Pi_n = n^{-1/2}C \) for a fixed matrix \( C \).

Example 2: (Conditional Moment Restrictions.) As in SW the moment conditions may result from conditional moment restrictions. Assume \( E[h(Y_i, \theta_0)|F_i] = 0 \), where \( h : H \times \Theta \rightarrow R^{k_1}, H \subset R^{k_2} \) and \( F_i \) is the information set at time \( i \). Let \( Z_i \) be a \( k_3 \)-dimensional vector of instruments contained in \( F_i \). If \( g_i(\theta) := h(Y_i, \theta) \otimes Z_i \), then \( Eg_i(\theta_0) = 0 \) follows by taking iterated expectations. In (2.1), \( k = k_1k_3 \) and \( l = k_2 + k_3 \).

2.2 Assumptions

This section is concerned with the asymptotic distribution of the GEL estimator for \( \theta \) when some components of \( \theta_0 = (\alpha_0', \beta_0')' \), \( \alpha_0 \) say, \( \alpha_0 \in A, A \subset R^{pA} \), are only weakly identified. Intuitively, this means that the moment condition (2.1) is not very informative about \( \alpha_0 \). For parameter vectors \( \theta = (\alpha', \beta_0')' \), \( Eg_n(z_i, \theta) \) may be very close to zero, not only for \( \alpha \) close to \( \alpha_0 \) but also when \( \alpha \) is far from \( \alpha_0 \). In that case, the restriction \( Eg_n(z_i, \theta_0) = 0 \) is not very helpful for making inference on \( \alpha_0 \). Assumption ID below provides a theoretical asymptotic framework for this phenomenon, which is taken from Assumption C in SW (p.1061). We refer the reader to SW (pp.1060-1) which provides substantial detailed motivation for this assumption and an explanation of why it models \( \alpha_0 \) as weakly and \( \beta_0 \) as strongly identified.

To describe the moment and distributional assumptions, we require some additional notation.

\[ \tilde{g}(\theta) := n^{-1} \sum_{i=1}^{n} g_i(\theta), \quad \tilde{G}(\theta) := n^{-1} \sum_{i=1}^{n} G_i(\theta), \]

[4]
\[
\Psi_n(\theta) := n^{1/2}(\hat{g}(\theta) - E\hat{g}(\theta)),
\]
\[
\hat{\Omega}(\theta) := n^{-1} \sum_{i=1}^{n} g_i(\theta)g_i(\theta)',
\]
where, if defined, \(G_i(\theta) := (\partial g_i/\partial \theta)(\theta) \in R^{k \times p}\). For notational convenience, a subscript \(n\) has been omitted in certain expressions. Define the \(k \times k\) matrices\(^1\)

\[
\Omega(\theta) := \lim_{n \to \infty} En^{-1} \sum_{i=1}^{n} g_i(\theta)g_i(\theta)',
\]
\[
\Delta(\theta_1, \theta_2) := \lim_{n \to \infty} E\Psi_n(\theta_1)\Psi_n(\theta_2)' \quad \text{and} \quad \Delta(\theta) := \Delta(\theta, \theta).
\]

Let \(\theta = (\alpha', \beta')',\) where \(\alpha \in A, A \subset R^p, \beta \in B, B \subset R^p\) and \(p_A + p_B = p\). Also let \(N \subset B\) denote an open neighborhood of \(\beta_0\).

**Assumption \(\Theta\):** The true parameter \(\theta_0 = (\alpha_0', \beta_0')'\) is in the interior of the compact space \(\Theta = A \times B\).

**Assumption ID:** (i) \(E\hat{g}(\theta) = n^{-1/2}m_{1n}(\theta) + m_2(\beta)\), where \(m_{1n}, m_1 : \Theta \to R^k\) and \(m_2 : B \to R^k\) are continuous functions such that \(m_{1n}(\theta) \to m_1(\theta)\) uniformly on \(\Theta, m_1(\theta_0) = 0\) and \(m_2(\beta) = 0\) if and only if \(\beta = \beta_0\); (ii) \(m_2 \in C^1(N)\); (iii) Let \(M_2(\beta) := (\partial m_2/\partial \beta)(\beta) \in R^{k \times p_B}\). \(M_2(\beta_0)\) has full column rank \(p_B\).

Next we detail the necessary moment assumptions.\(^2\)

**Assumption M:** (i) \(\max_{1 \leq i \leq n} \sup_{\theta \in \Theta} ||g_i(\theta)|| = o_p(n^{1/2})\); (ii) \(\Omega(\cdot)\) is in \(C^0(A \times \{\beta_0\})\) and bounded on \(\Theta, \Omega(\cdot)\) is nonsingular for all \(\theta \in A \times \{\beta_0\}\); \(\sup_{\theta \in \Theta} ||\hat{\Omega}(\theta) - \Omega(\theta)|| = o_p(1)\), \(\sup_{\theta \in A \times N} n^{-1} \sum_{i=1}^{n} ||g_i(\theta)g_i(\theta)'|| = O_p(1)\); (iii) \(\Psi_n \Rightarrow \Psi, \) where \(\Psi(\theta)\) is a Gaussian stochastic process on \(\Theta\) with mean zero and covariance function \(E\Psi(\theta_1)\Psi(\theta_2)' = \Delta(\theta_1, \theta_2)\). For each \(\varepsilon > 0\) there exists a \(M_\varepsilon < \infty\) such that \(\Pr(\sup_{\theta \in A \times N} ||\Psi(\theta)|| \leq M_\varepsilon) > 1 - \varepsilon\).

\(M(i)\) adapts Assumption 1(d) of NS, \(E \sup_{\beta \in B} ||g_i(\beta)||^\xi < \infty\) for some \(\xi > 2\), from the i.i.d. setting with strong identification \((p_A = 0 \text{ and thus } \theta = \beta \text{ and } \Theta = B)\) to the weakly identified set-up considered here. A sufficient condition for \(M(i)\) in the time series context and under ID is given by

\[
\sup_{i \geq 1} \sup_{\theta \in \Theta} ||g_i(\theta)||^\xi < \infty \quad \text{for some } \xi > 2.
\]

\(^2\)Weak convergence here is defined with respect to the sup-norm on function spaces and Euclidean norm on \(R^k\).

\[\texttt{[5]}\]
nonsingular for $\theta \in A \times N$. M(iii) is essentially the “high level” Assumption B of SW (p.1059) that states that $\Psi_n$ obeys a functional central limit theorem. In Assumption B’, SW provide primitive sufficient conditions for their Assumption B that can also be found in Andrews (1994). Note that the definition of weak convergence [Andrews (1994, p.2250)] and M(iii) imply that $\sup_{\theta \in \Theta} \|\Psi_n(\theta)\| \to_d \sup_{\theta \in \Theta} \|\Psi(\theta)\|$ and, thus, also that $\sup_{\theta \in \Theta} \|\tilde{\theta}(\theta) - E\tilde{\theta}(\theta)\| \to_p 0$. In the proof of Theorem 2 below, we require $\sup_{\theta \in A \times N} \|\Psi(\theta)\|$ bounded in probability.

It is interesting to note that for i.i.d. data, an application of the Borel-Cantelli Lemma shows that M(i) is implied by Assumption 1(d) of NS even if $\xi = 2$; see Owen (1990, Lemma 3) for a proof. Hence, using Lemmas 7-9 given below, their Assumption 1(d) can be weakened to $\xi \geq 2$ for the consistency and asymptotic normality of the GEL estimator under strong identification with i.i.d. data (see their Theorems 3.1 and 3.2). Therefore, for i.i.d. data, identical assumptions guarantee consistency and asymptotic normality for both GEL and two step efficient GMM estimators [Hansen (1982)].

**Example 1 (cont.):** See Guggenberger (2003). For the linear IV model (2.2) Assumption ID can be expressed as

**Assumption ID**: $\Pi = \Pi_n = (\Pi_A, \Pi_B) \in R^{k \times (p_A + p_B)}$, where $p_A + p_B = p$. For a fixed matrix $C_A \in R^{k \times p_A}$, $\Pi_A = n^{-1/2}C_A$ and $\Pi_B$ has full column rank.

Under Assumption ID’, i.i.d. data and instrument exogeneity it follows that

$$E\tilde{\theta}(\theta) = E\tilde{\theta}_i(\theta) = E(Z_iZ'_i)(n^{-1/2}C_A, \Pi_B)(\theta_0 - \theta),$$

which implies that in the notation of ID(i), $m_{1n}(\theta) = m_1(\theta) = E(Z_iZ'_i)C_A(\alpha_0 - \alpha)$ and $m_2(\beta) = E(Z_iZ'_i)\Pi_B(\beta_0 - \beta)$. Also, note that Assumption ID’ includes the partially identified model of Phillips (1989). In particular, choosing $p_A$ and setting $C_A = 0$, one obtains a model in which $\Pi$ may have any desired (less than full) rank.

We now give simple sufficient conditions that imply Assumption M. Let $U := (u, V)$.

**Assumption M**: (i) $\{(U_i, Z_i) : i \geq 1\}$ are i.i.d.; (ii) $EZ_iU'_i = 0$; (iii) $E|Z_i|^4 < \infty$, $Q_{ZZ} := E(Z_iZ'_i) > 0$, $Ev_iu_iu'_i$, $Ev_iu_iV_iZ_iZ'_i$ and $Ev_iV_i\epsilon_iZ_iZ'_i$ exist and are finite for $j, k, l = 1, ..., p$, where $V_i$ denotes the $j^{th}$ component of the vector $V_i$; (iv) $\Omega(\theta)$ is nonsingular for all $\theta \in A \times \{\beta_0\}$.

Assumptions M’(i) and (ii) state that errors and exogenous variables are jointly i.i.d. and the standard instrument exogeneity assumption is satisfied whereas M’(iii) and (iv) are technical conditions.

The following lemma shows that Assumption M’ in the linear model implies Assumption M.

**Lemma 1** Suppose that Assumptions ID’, M’ and $\Theta$ hold in the linear IV model (2.2). Then Assumption ID and M hold.
Therefore the various technical conditions of Assumption M reduce to very simple moment conditions in the linear model. Note that M’ implies $E[\sup_{\theta \in \Theta} |g_i(\theta)|^\xi] < \infty$ for $\xi = 2$. However, we do not need the assumption $E[\sup_{\theta \in \Theta} |g_i(\theta)|^\xi] < \infty$ for a $\xi > 2$ to prove $n^{1/2}$-consistency of the GEL estimator of the strongly identified parameters.

**Assumption HOM:** (Conditional Homoskedasticity) $E(U_iU_i'|Z_i) = \Sigma_U > 0$.

HOM, which is used in Staiger and Stock (1997), is sufficient for Assumption M’(iv). That is, Assumptions M’(i)-(iii) and HOM imply M’(iv) under ID’. This follows from $\Omega(\theta) = QZZ'v_\alpha \Sigma_u V_A \alpha$ for $\theta \in A \times \{\beta_0\}$, where $v_\alpha' := (1, (\alpha_0 - \alpha)')$ and $\Sigma_u V_A$ is the $(1 + p_A) \times (1 + p_A)$ upper left submatrix of $\Sigma_U$. However, M’ is more general than HOM because it allows for conditional heteroskedasticity. For example, $u_i = |Z_i|\xi_i$, where $\xi_i \sim N(0, 1)$ is independent of $Z_i \sim N(0, I_k)$, is compatible with M’.

### 2.3 The GEL Estimator

This subsection provides a formal definition of the GEL estimator of $\theta_0$.

Let $\rho$ be a real-valued function $Q \to R$, where $Q$ is an open interval of the real line that contains 0 and

$$\hat{\Lambda}_n(\theta) := \{\lambda \in R^k : \lambda' g_i(\theta) \in Q \text{ for } i = 1, \ldots, n\}. \quad (2.5)$$

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

$$\hat{\theta}_\rho := \arg \min_{\theta \in \Theta} \sup_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}_\rho(\theta, \lambda), \quad (2.6)$$

where

$$\hat{P}_\rho(\theta, \lambda) := (2\sum_{i=1}^n \rho(\lambda' g_i(\theta))/n) - 2\rho_0. \quad (2.7)$$

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

---

3For compact $\Theta$, continuous $\rho$ and $g_i$ ($i = 1, \ldots, n$), the existence of an argmin $\hat{\theta}$ may be shown. In fact, $\sup_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta, \lambda)$, viewed as a function in $\theta$, can be shown to be lower semicontinuous (ls).

A function $f(x)$ is ls at $x_0$ if, for each real number $c$ such that $c < f(x_0)$, there exists an open neighborhood $U$ of $x_0$ such that $c < f(x)$ holds for all $x \in U$. The function $f$ is said to be ls if it is ls at each $x_0$ of its domain. It is easily shown that ls functions on compact sets take on their minimum. Uniqueness of $\hat{\theta}$, however, is not implied. As a simple example, consider the i.i.d. linear IV model in (2.2) when $p = 2$ and let the two components $Y_{ij}$, ($j = 1, 2$), of $Y_i$ be independent Bernoulli random variables. Then, for each $n$, the probability that $Y_{i1} = Y_{i2}$ for every $i = 1, \ldots, n$ is positive. If $Y_{i1} = Y_{i2}$ for every $i = 1, \ldots, n$ and $\theta \in \Theta$ is an argmin of $\sup_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta, \lambda)$, then each $\theta \in \Theta$ with $\theta_1 + \theta_2 = \theta_1 + \theta_2$ is as well. To uniquely define $\hat{\theta}$, we could, for example, do the following. From the set of all vectors $\theta \in \Theta$ that minimize $\sup_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta, \lambda)$, let $\hat{\theta}$ be the vector that has smallest first component. (If that does not pin down $\hat{\theta}$ uniquely, choose from the remaining vectors according to the second component and so on.)
The definition of the GEL estimator $\hat{\theta}_\rho$ is adopted from NS. We slightly modify their definition of $\hat{P}_\rho(\theta, \lambda)$ by recentering and rescaling which simplifies the presentation. We usually write $\hat{P}(\theta, \lambda)$ for $\hat{P}_\rho(\theta, \lambda)$ and $\hat{\theta}$ for $\hat{\theta}_\rho$.

The most popular GEL estimators are the CUE, the EL and the ET estimator which correspond to $\rho(v) = -(1 + v)^2/2$, $\rho(v) = \ln(1 - v)$ and $\rho(v) = -\exp v$, respectively. The EL estimator was introduced by Imbens (1997), Owen (1988, 1990) and Qin and Lawless (1994) and the ET estimator by Imbens, Spady and Johnson (1998) and Kitamura and Stutzer (1997). For a recent survey of GEL methods see Imbens (2002).

### 2.4 First-Order Equivalence

This subsection obtains the asymptotic distribution of the GEL estimator $\hat{\theta}_\rho$ under Assumption ID. Theorem 2 shows that the weakly identified parameters of $\theta_0$ are estimated inconsistently and their GEL estimator has a nonstandard limiting distribution while the GEL estimator of the strongly identified parameters is $n^{1/2}$-consistent but no longer asymptotically normal. Analogous results are available for LIML or more generally for GMM; see Phillips (1984) and SW, Theorem 1. The rather surprising finding is that the first-order asymptotic theory under ID is identical for all GEL estimators $\hat{\theta}_\rho$, as long as $\rho$ satisfies Assumption $\rho$.

If defined, let

$$
\lambda(\theta) \text{ be such that } \hat{P}(\theta, \lambda(\theta)) = \max_{\lambda \in \Lambda_n(\theta)} \hat{P}(\theta, \lambda).
$$

For $\theta = (\alpha', \beta')' \in \Theta$ and $b \in \mathbb{R}^p$ let

$$
P(\theta, b) := [\Psi(\theta) + m_1(\theta) + M_2(\beta)b]\Omega(\theta)^{-1}[\Psi(\theta) + m_1(\theta) + M_2(\beta)b].
$$

The next theorem establishes the asymptotic behavior of $\hat{\theta} = (\hat{\alpha}', \hat{\beta}')'$ under Assumption ID.

**Theorem 2** Suppose Assumptions $\Theta$, ID, $M$ and $\rho$ are satisfied.

(i) $\hat{\alpha}$ is in general inconsistent and

$$
n^{1/2}(\hat{\beta} - \beta_0) = O_p(1).
$$
Therefore, like the i.i.d. case.

is in general nonnormal is a consequence of the inconsistent estimation of the parameter estimates under weak identification, see Caner (2003). If $P_{ab} := \Omega \cap \{b \in \mathbb{R}^p : ||b|| \leq M\}$ and define $\theta_{ab} := (\alpha', \beta_0 + n^{-1/2}b)'$. Then, for $(\alpha, b) \in A \times B_M$, $n\hat{P}(\theta_{ab}, \lambda(\theta_{ab})) \Rightarrow P_{ab} := P(\alpha', \beta_0', b)$. Assume there exists a random element $(\alpha^*, b^*) \in A \times \mathbb{R}^p$ such that a.s. $P_{\alpha^*b^*} < \inf_{(\alpha, b) \in (A \times \mathbb{R}^p) \setminus G} P_{ab}$ for every open set $G$ that contains $(\alpha^*, b^*)$. Then

$$(\hat{\alpha}, n^{1/2}(\hat{\beta} - \beta_0)) \rightarrow_d (\alpha^*, b^*).$$

**Remark 1:** Theorem 2(ii) is analogous to Theorem 1 in SW (p.1062) for GMM. Note that from (A.5) $\lambda(\hat{\theta}) = -(\Omega((\alpha', \beta_0')' + o_p(1))^{-1}\hat{g}(\hat{\theta}) = O_p(n^{-1/2})$ as $\hat{g}(\hat{\theta}) = O_p(n^{-1/2})$. Moreover, using the proof of Theorem 2 it can be shown that

$$n^{1/2}\lambda(\hat{\theta}) \rightarrow_d -\Omega((\alpha^*, \beta_0')' - 1 M_{M_2(\beta_0)}(\Omega((\alpha^*, \beta_0')' [\Psi((\alpha^*, \beta_0')') + m_1((\alpha^*, \beta_0')')].$$

Therefore, like $\hat{\beta}$, although $n^{1/2}$-consistent, $n^{1/2}\lambda(\hat{\theta})$ admits a nonstandard asymptotic distribution, see Caner (2003). If $p_A = 0$, where all parameters are strongly identified, $n^{1/2}\lambda(\hat{\theta}) \rightarrow_d N(0, \Omega^{-1} M_{M_2}(\Omega) \Delta M_{M_2}(\Omega') \Omega^{-1})$, where $M_2 := M_2(\beta_0)$, $\Omega := \Omega(\beta_0)$ and $\Delta := \Delta(\beta_0)$. The covariance matrix reduces to $\Omega^{-1} M_{M_2}(\Omega)$ in the i.i.d. case.

The proof of Theorem 2 also provides a formula (A.7) for $b^*(\alpha) := \arg\min_{b \in \mathbb{R}^p} P_{ab}$ for $\alpha \in A$. In particular, if $p_A = 0$, (A.7) shows that

$$n^{1/2}(\hat{\beta} - \beta_0) \rightarrow_d N(0, V(\beta_0)),$$

where

$$V(\beta_0) := (M_2' \Omega^{-1} M_2)^{-1} M_2' \Omega^{-1} \Delta \Omega^{-1} M_2 (M_2' \Omega^{-1} M_2)^{-1}.$$

The matrix $V(\beta_0)$ simplifies to $(M_2' \Omega^{-1} M_2)^{-1}$ in the i.i.d. case and thus the above formula coincides with Theorem 3.2 of NS. However, the asymptotic variance matrix of $n^{1/2}(\hat{\beta} - \beta_0)$ in the time series context is in general different from that in NS and the estimator $\beta$ as defined above would thus be inefficient. Block methods as in Kitamura (1997) or kernel-smoothing methods as in Smith (2001) can be used for efficient GEL estimation in a time series context with strong identification. In the case $p_A > 0$, the fact that the asymptotic distribution of the strongly identified parameter estimates is in general nonnormal is a consequence of the inconsistent estimation of $\alpha_0$.

**Remark 2:** Given the nonnormal asymptotic distribution of the GMM and GEL parameter estimates under weak identification (established in Theorem 1 in SW and Theorem 2 above, respectively) the asymptotic distribution of test statistics based on these estimators, such as $t$- or Wald statistics, will also be nonstandard and non-pivotal. Furthermore, these limiting distributions depend on quantities that cannot be consistently estimated [see Staiger and Stock (1997, p.564)] which militates against their use for the construction of test statistics or confidence regions for $\theta_0$. The next section introduces alternative approaches that overcome these difficulties.

**Example 1 (cont.):** The specialization of Theorem 2 to the i.i.d. linear IV model of Example 1 was derived in Guggenberger (2003).
3 Test Statistics

This section proposes several statistics to test the simple hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. We establish that they are asymptotically pivotal quantities and have limiting chi-square null distributions under Assumption ID. Therefore these statistics lead to tests whose size properties are unaffected by the strength or weakness of identification. For the time series set-up considered here there are at least two other statistics that share this property, namely Anderson and Rubin’s (1949) $AR$- and Kleibergen’s (2001, 2002a) $K$-statistic. The first statistic $GELR_\rho(\theta)$ that we describe may be interpreted as a likelihood ratio statistic. It has an asymptotic $\chi^2(k)$ null distribution and is first-order equivalent to the $AR$-statistic. The second set of statistics in this section, $S_\rho(\theta)$ and $LM_\rho(\theta)$, are based on the FOC of $\hat{P}_\rho(\theta, \lambda)$ with respect to $\theta$. Each has a limiting $\chi^2(p)$ null distribution and is first-order equivalent to the $K$-statistic. For a recent survey on robust inference methods with weak identification, see Stock, Wright and Yogo (2002).

To motivate the first statistic, consider an i.i.d. setting. In this case, $GELR_{EL}(\theta)$ may be thought of in terms of the empirical likelihood ratio statistic $R(\theta)$, where

$$R(\theta) := \sup_{w_1, \ldots, w_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} w_i g_i(\theta) = 0, w_i > 0, \sum_{i=1}^{n} w_i = 1 \right\}. \quad (3.1)$$

The criterion function $R(\theta)$ can be interpreted as a nonparametric likelihood ratio. Indeed, for fixed $\theta \in \Theta$ and given $g_i(\theta), (i = 1, \ldots, n)$, the numerator of $R(\theta)$ is the maximal probability of observing the given sample $g_i(\theta), (i = 1, \ldots, n)$, over all discrete probability distributions $(w_1, \ldots, w_n)$ on the sample such that the sample analogue $\sum_{i=1}^{n} w_i g_i(\theta) = 0$ of the moment condition (2.1) is satisfied. The denominator $(1/n)^n$ equals the unrestricted maximal probability. It can then be shown that $-2 \ln R(\theta_0) = n\hat{P}_{EL}(\theta_0, \lambda(\theta_0))$, where $\lambda(\theta_0)$ is the vector of Lagrange multipliers associated with the $k$ moment restrictions $\sum_{i=1}^{n} w_i g_i(\theta_0) = 0$ in the constrained maximization problem (3.1). Therefore, the renormalized criterion function of the EL estimator has an interpretation as $-2$ times the logarithm of the likelihood ratio statistic $R(\theta_0)$.

Generalizing from the i.i.d. to the time series set-up and from EL to arbitrary $\rho$, the first statistic we consider is the renormalized GEL criterion function (2.7)

$$GELR_\rho(\theta) := n\hat{P}_\rho(\theta, \lambda(\theta)). \quad (3.2)$$

Secondly, following Kleibergen’s (2001) suggestion of constructing a statistic from the FOC with respect to $\theta$ but in the GMM framework, we construct a test statistic based on the GEL FOC for $\theta$. If the minimum of the objective function $\hat{P}(\theta, \lambda(\theta))$ is obtained in the interior of $\Theta$, the score vector with respect to $\theta$ must equal 0 at $\hat{\theta}$, i.e.

$$\lambda(\hat{\theta})' \sum_{i=1}^{n} \rho_1(\lambda(\hat{\theta})'g_i(\hat{\theta})|G_i(\hat{\theta})/n = 0'. \quad (3.3)$$

\[6\)NS show that under certain conditions including $\{z_i : i \geq 1\}$ i.i.d., $\hat{\theta}_{EL} = \arg \max_{\theta \in \Theta} \ln R(\theta)$. Thus $\ln R(\theta)$ can be interpreted as the criterion function of the EL estimator.
For \( \theta \in \Theta \), define the \( k \times p \) matrix
\[
D_p(\theta) := \sum_{i=1}^{n} \rho_i(\lambda(\theta)'g_i(\theta))G_i(\theta)/n.
\tag{3.4}
\]
Thus, (3.3) may be written as \( \lambda(\hat{\theta})'D_p(\hat{\theta}) = 0' \). The test statistic is therefore given as a quadratic form in the score vector \( \lambda(\theta)'D_p(\theta) \) evaluated at the hypothesized parameter vector \( \theta \)
\[
S_\rho(\theta) := n\lambda(\theta)'D_p(\theta)\left(D_p(\theta)'\tilde{\Omega}(\theta)^{-1}D_p(\theta)\right)^{-1}D_p(\theta)'\lambda(\theta),
\tag{3.5}
\]
where \( \rho \) is any function satisfying Assumption \( \rho \) and \( \tilde{\Omega}(\theta) \) is a consistent estimator of \( \Delta(\theta) \). We also consider the following variant of \( S_\rho(\theta) \)
\[
LM_\rho(\theta) := n\tilde{g}(\theta)'\tilde{\Omega}(\theta)^{-1}D_p(\theta)\left(D_p(\theta)'\tilde{\Omega}(\theta)^{-1}D_p(\theta)\right)^{-1}D_p(\theta)'\tilde{g}(\theta)
\tag{3.6}
\]
that substitutes \(-\Delta(\theta)^{-1}\tilde{g}(\theta)\) for \( \lambda(\theta) \) in \( S_\rho(\theta) \); see (A.8) in the Appendix, where it is shown that \( n^{1/2}\lambda(\theta) = -\Delta(\theta)^{-1}n^{1/2}\tilde{g}(\theta) + o_p(1) \). The statistic \( LM_\rho(\theta) \) is similar to a GMM Lagrange multiplier statistic given in Newey and West (1987). To make the origin of the above test statistics clearer, we adopted the notation \( LM_\rho(\theta) \) and \( S_\rho(\theta) \) respectively in place of \( K_\rho(\theta) \) and \( K_p^L(\theta) \) previously given to the statistics in Guggenberger (2003). To use these statistics for hypothesis tests or for the construction of confidence regions one needs a consistent estimator \( \tilde{\Omega}(\theta) \) of \( \Delta(\theta) \). Under assumptions given below, the sample average \( \tilde{\Omega}(\theta) \) may be used for \( \tilde{\Omega}(\theta) \).\(^7\) Note that when \( \rho(v) = -(1 + v)^2/2 \), that is, in the case of a GEL CUE criterion, the GEL statistics \( S_\rho(\theta) \) (3.5) and \( LM_\rho(\theta) \) (3.6) are then identical and given in closed form by (3.6) with \( \lambda(\theta) = -\tilde{\Omega}(\theta)^{-1}\tilde{g}(\theta) \) in the definition of \( DCUE(\theta) \), where \( \tilde{\Omega}(\theta)^{-1} \) denotes any generalized inverse of \( \tilde{\Omega}(\theta) \).

As noted above the GEL and GMM CUE are numerically identical. However, although the structures of the two statistics coincide, in general, the statistic \( LM_{CUE}(\theta) \) and Kleibergen’s (2001) \( K \)-statistic based on the GMM CUE are not identical. The reason is that, in general, the first order derivatives of the GMM and GEL CUE objective functions are not equal. The \( K \)-statistic in Kleibergen (2001) is based on the FOC of the GMM CUE criterion \( \tilde{g}(\theta)'\tilde{\Omega}(\theta)^{-1}\tilde{g}(\theta) \). It replaces \( DCUE(\theta) \) in \( LM_{CUE}(\theta) \) by \( \tilde{G}(\theta) - \tilde{V}(\theta)(I_p \otimes (\tilde{\Omega}(\theta)^{-1}\tilde{g}(\theta))) \), where \( \tilde{V}(\theta) \) is an estimator for \( \lim_{n \to \infty} E\{n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} [G_i(\theta) - E G_i(\theta)](I_p \otimes g_j(\theta)') - E(I_p \otimes g_j(\theta)') \} \). The particular assumptions made on \( \Delta(\theta) \) determine the choice of estimators \( \tilde{\Omega}(\theta) \) and \( \tilde{V}(\theta) \). If the sample average \( \tilde{\Omega}(\theta) \) is used for \( \tilde{\Omega}(\theta) \) and \( \sum_{i=1}^{n} G_i(\theta)(I_p \otimes g_i(\theta)')/n \) for \( \tilde{V}(\theta) \), then the statistic \( LM_{CUE}(\theta) \) and the \( K \)-statistic coincide.

Some intuition for these test statistics is provided under strong identification. Under strong identification, NS show consistency of \( \hat{\theta} \). Therefore, if the FOC (3.3)
hold at $\hat{\theta}$, then, at least asymptotically, they also hold at the true value $\theta_0$. The statistic $S_\rho(\theta)$ can then be interpreted as a quadratic form whose criterion is expected to be small at the true value $\theta_0$. If, however, all parameters are weakly identified this argument is no longer valid. From Theorem 2, $\hat{\theta}$ is no longer consistent for $\theta_0$. Therefore, although the FOC hold at $\hat{\theta}$, this does not imply automatically that they also approximately hold at the true value $\theta_0$. However, it can be shown that under weak identification the FOC $\lambda(\theta)'D_\rho(\theta) = 0'$ not only hold at $\hat{\theta}$ w.p.a.1 but are satisfied to order $O_p(T^{-1})$ uniformly over $\theta \in \Theta$. Thus, under weak identification the FOC do not pin down the true value $\theta_0$. Consequently, the power properties of hypothesis tests for $\theta_0$ based on the statistics $S_\rho(\theta)$ or $LM_\rho(\theta)$ should be expected to be better under strong rather than weak identification. Size properties however are not affected by the strength or weakness of identification. This is corroborated by the Monte Carlo simulations reported below and theoretically by Theorem 4.

We now turn to consider the asymptotic distribution of $GELR_\rho(\theta)$ evaluated at a vector $\theta = (\alpha', \beta_0')'$, thus allowing for a fixed alternative in the weakly identified components. We need the following local version of Assumption M.

**Assumption M$_\theta$:** Let $\theta = (\alpha', \beta_0')' \in A \times \{\beta_0\}$. Suppose (i) $\max_{1 \leq i \leq n} ||g_i(\theta)|| = o_p(n^{1/2})$; (ii) $\Delta(\theta) > 0$, $\Omega(\theta) \rightarrow_p \Delta(\theta)$, $n^{-1} \sum_{i=1}^n ||g_i(\theta)g_i(\theta)'|| = O_p(1)$; (iii) $\Psi_n(\theta) \rightarrow_d \Psi(\theta)$, where $\Psi(\theta) \equiv N(0, \Delta(\theta))$.

Note that for $\theta = (\alpha', \beta_0')'$ $M_\theta$(iii) and ID imply that $\widehat{g}(\theta) \rightarrow_p 0$. Thus, under $M_\theta$(iii) and ID the assumption $\widehat{\Omega}(\theta) \rightarrow_p \Delta(\theta)$ in $M_\theta$(ii) is equivalent to the assumption $n^{-1} \sum_{i=1}^n (g_i(\theta) - \widehat{g}(\theta))(g_i(\theta) - \widehat{g}(\theta))' \rightarrow_p \Delta(\theta)$ for $\theta = (\alpha', \beta_0')'$, which is Assumption D’ in SW. The assumption rules out many interesting time series cases. However, it is more general than an i.i.d. assumption. The assumption allows for m.d.s. and thus covers various intertemporal Euler equations applications and regression models with m.d.s. errors. As in SW, a possible application is the intertemporally separable CCAPM. Without assuming $\Omega(\theta) \rightarrow_p \Delta(\theta)$, a limiting chi-square distribution would no longer obtain in the following theorems. The problem arises because the GEL estimator as defined in (2.6) is not efficient in the time series set-up considered here.

**Theorem 3** Suppose ID, $M_\theta$(i)-(iii) and $\rho$ hold for $\theta = (\alpha', \beta_0')'$. Then

$$GELR_\rho(\theta) \rightarrow_d \chi^2(k, \delta),$$

where the noncentrality parameter $\delta = m_1(\theta)\Delta(\theta)^{-1}m_1(\theta)$. In particular,

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

To describe the asymptotic distribution of the statistics $LM_\rho(\theta_0)$ and $S_\rho(\theta_0)$, we need the following additional assumptions. Write $G_i(\theta) = (G_{iA}(\theta), G_{iB}(\theta))$, where the matrices $G_{iA}(\theta)$ and $G_{iB}(\theta)$ are of column dimension $p_A$ and $p_B$, respectively.

Let $\theta = (\alpha', \beta_0') \in A \times \{\beta_0\}$ and $M \subset \Theta$ an open neighborhood of $\theta$. 

[12]
Assumption M_δ: (cont.) (iv) \( \hat{g}(\cdot) \) is differentiable at \( \overline{\theta} \) a.s. for each \( \overline{\theta} \in \mathcal{M} \), \( \hat{g}(\overline{\theta}) \) is integrable (with respect to the probability measure) for all \( \overline{\theta} \in \mathcal{M} \), sup\( _{\theta \in \mathcal{M}} \| \hat{g}(\theta) \| \) is integrable, \( m_{1n} \in C^{1}(\Theta) \) and \( M_{1n}(\cdot) := (\partial m_{1n}/\partial \theta)(\cdot) \) converges uniformly on \( \Theta \) to some function; (v) \( n^{-1} \sum_{i=1}^{n} (\text{vec}G_{iA}(\theta))g_{i}(\theta) \rightarrow_{p} \Delta A(\theta) (\Delta_{A}(\theta) \text{ is defined in (vii)}) \), \( \hat{\Theta}(\theta) \rightarrow_{p} \Delta(\theta) \), \( G_{B}(\theta) := n^{-1} \sum_{i=1}^{n} G_{iB}(\theta) \rightarrow_{p} E\hat{G}_{B}(\theta) \); (vi) \( n^{-1} \sum_{i=1}^{n} ||G_{iA}(\theta)|| ||g_{i}(\theta)|| = O_{p}(1) \), \( n^{-3/2} \sum_{i=1}^{n} ||G_{iB}(\theta)|| ||g_{i}(\theta)|| = O_{p}(1) \); (vii) \( n^{-1/2} \sum_{i=1}^{n} ((\text{vec}G_{iA}(\theta) - E\text{vec}G_{iA}(\theta)))'g_{i}(\theta) - E(\text{vec}G_{iA}(\theta))'g_{i}(\theta))' \rightarrow_{d} N(0, \Omega(\theta)) \), where \( \Omega(\theta) := \text{lim}_{\theta \rightarrow \infty} \text{var}(n^{-1/2} \sum_{i=1}^{n} ((\text{vec}G_{iA}(\theta))'g_{i}(\theta))' \in R^{p_{A}+1 \times k(p_{A}+1)} \) is positive definite.

In \( M_{\theta}(\text{vii}) \) write

\[
V(\theta) = \begin{pmatrix}
\Delta_{AA} & \Delta_{A} \\
\Delta_{A}' & \Delta
\end{pmatrix}(\theta) \text{, where } \Delta_{AA}(\theta) \in R^{p_{A} \times k(p_{A}+1)}.
\]

Assumption \( M_{\theta}(\text{iv}) \) allows the interchange of the order of integration and differentiation in Assumption ID, i.e. \( (\partial E\hat{g}/\partial \theta)(\theta) = E\hat{G}(\theta) \). It also guarantees that \( M_{1n}(\theta) \rightarrow M_{1}(\theta) := (\partial m_{1}/\partial \theta)(\theta) \). Assumption ID and \( M_{\theta} \) thus imply that

\[
E\hat{G}(\theta) = n^{-1/2} M_{1n}(\theta) + (0, M_{2}(\beta_{0})) \rightarrow (0, M_{2}(\beta_{0})),
\]

where by ID the limit matrix \( (0, M_{2}(\beta_{0})) \) is of deficient rank \( p_{B} \). Assumption \( M_{\theta}(\text{v}) \) is comparable to \( M_{\theta}(\text{ii}) \), where \( \hat{\Theta}(\theta) \rightarrow_{p} \Delta(\theta) \) was assumed and extends \( M_{\theta}(\text{ii}) \) to cross-product terms in \( \text{vec}G_{iA}(\theta) \) and \( g_{i}(\theta) \). Assumption \( M_{\theta}(\text{vi}) \) contains additional weak technical conditions that guarantee that certain expressions in the proof of Theorem 4 are asymptotically negligible.

The key assumption is \( M_{\theta}(\text{vii}) \) which is a stronger version of \( M_{\theta}(\text{iii}) \) and states that a CLT theorem holds simultaneously for the centered \( g_{i}(\theta) \) and part of the derivative matrix, namely \( \text{vec}G_{iA}(\theta) \). Write \( LM_{\theta}(\theta) = n\hat{g}Q^{-1}D(D^{-1}DQ^{-1})^{-1}D^{*}Q^{-1}(\theta) \), where \( D = D_{\rho}(\theta) \) and \( Q = Q(\theta) \). As shown in the proof of Theorem 4, for \( \theta = (\alpha', \beta_{0})' \), Assumptions ID, \( \rho \), \( M_{\theta}(\text{i}-\text{vi}) \) and \( G_{A}(\theta) := n^{-1} \sum_{i=1}^{n} G_{iA}(\theta) \rightarrow_{p} E\hat{G}_{A}(\theta) \) imply that \( D \rightarrow_{p} (0, M_{2}(\beta_{0})) \). Therefore, the probability limit of \( D^{*}Q^{-1}D \) is not invertible without renormalisation. Define \( D^{*} := DA \) where the \( p \times p \) diagonal matrix \( \Lambda := \text{diag}(n^{1/2}, ..., n^{1/2}, 1, ..., 1) \) with first \( p_{A} \) diagonal elements equal to \( n^{1/2} \) and the remainder equal to unity. Hence,

\[
LM_{\theta}(\theta) = n\hat{g}Q^{-1}D^{*}(D^{*}Q^{-1}D^{*})^{-1}D^{*}Q^{-1}(\theta).
\]

In the proof of Theorem 4 we show that under Assumptions ID, \( \rho \) and \( M_{\theta}(\text{i}-\text{vi}) \)

\[
\text{vec}D^{*} = \text{vec}(0, -M_{2}(\beta_{0})) + \begin{pmatrix}
-I_{k_{p}} & \Delta_{A}(\theta) \Delta(\theta)^{-1} \\
0 & 0
\end{pmatrix} n^{-1/2} \sum_{i=1}^{n} \begin{pmatrix}
\text{vec}G_{iA}(\theta) \\
g_{i}(\theta)
\end{pmatrix} + o_{p}(1).
\]

The additional \( M_{\theta}(\text{vii}) \), in particular the full rank assumption on \( V(\theta) \), ensures that \( D^{*}Q^{-1}D^{*} \) has full rank w.p.a.1. Assumption \( M_{\theta}(\text{vii}) \) is closely related to Kleibergen’s (2001) Assumption 1. Unlike Kleibergen (2001), however, we assume ID which, as just
shown, requires that we are specific about which part of the derivative matrix $G_i(\theta)$ together with $g_i(\theta)$ satisfies a CLT with full rank covariance matrix, namely $G_{iA}(\theta)$ which corresponds to the weakly identified parameters. Assumption ID possesses the advantage that we can obtain the asymptotic distribution of the test statistics under fixed alternatives of the form $\theta = (\alpha', \beta_0')'$ and therefore derive asymptotic power results.

**Theorem 4** Suppose ID, M (i)-(vii) and $\rho$ hold for $\theta = (\alpha', \beta_0')'$. Then,

$$S_\rho(\theta), LM_\rho(\theta) \rightarrow_d (W(\alpha) + \zeta')(W(\alpha) + \zeta),$$

where the random $p$-vector $W(\alpha)$ is defined in (A.11) of the Appendix, $\zeta \sim N(0, I_p)$ and $W$ and $\zeta$ are independent. We have $W(\alpha_0) \equiv 0$ and therefore

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

**Remark 1:** The proof of Theorem 4 crucially hinges on the fact that $n^{1/2}\lambda(\theta_0)$ and $\text{vec}D_\rho(\theta_0)$ (suitably normalized) from the FOC (3.3) are asymptotically jointly normally distributed and, moreover, are asymptotically independent. A similar result is critical also for Kleibergen’s (2001) K-statistic which generalizes Brown and Newey’s (1998) analysis of efficient GMM moment estimation to the weakly identified set-up. Therefore, by using an appropriate weighting matrix in the quadratic forms (3.5) and (3.6) that define $S_\rho(\theta_0)$ and $LM_\rho(\theta_0)$, respectively, we immediately obtain the limiting $\chi^2(p)$ null distribution of Theorem 4.

**Remark 2:** Theorems 3 and 4 provide a straightforward method to construct confidence regions or hypothesis tests on $\theta_0$. For example, a critical region for a test of the hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ 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. A $(1-r)$-confidence region for $\theta_0$ is obtained by inverting the just-described test, i.e. $\{\theta \in \Theta : GELR_\rho(\theta) \leq \chi^2_r(k)\}$. Confidence regions and hypothesis tests based on $S_\rho(\theta)$ and $LM_\rho(\theta)$ may be constructed in a similar fashion.

**Remark 3:** Theorems 3 and 4 demonstrate that $GELR_\rho(\theta_0), S_\rho(\theta_0)$ and $LM_\rho(\theta_0)$ are asymptotically pivotal statistics under weak and strong identification. Therefore, the size of tests based on these statistics should not vary much with the strength or weakness of identification in finite samples. However, these results also show that under weak identification hypothesis tests based on these statistics are inconsistent. For example, the noncentrality parameter $\delta$ does not diverge for increasing sample size and therefore the rejection rate under the alternative does not converge to 1. This is intuitively reasonable because if identification is weak one cannot learn much about $\alpha_0$ from the data.

**Remark 4:** 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. 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. Anderson and Rubin’s (1949) AR-statistic has a \( \chi^2(k) \) limiting null distribution as well. Kleibergen (2002b) shows that it equals the sum of two independent statistics, namely the \( K \)-statistic [Kleibergen (2002a)] and a \( J \)-statistic [Hansen (1982)] that test location and misspecification, respectively. Mutatis mutandis, a similar decomposition may be given for the \( GELR_\rho(\theta_0) \) statistic in terms of \( S_\rho(\theta_0) \) or \( LM_\rho(\theta_0) \).

**Remark 5:** SW (Theorem 2) derive the asymptotic distribution under weak identification of the analogue of \( GELR_\rho(\theta_0) \) for the (GMM) CUE which is also a \( \chi^2(k) \) null distribution. In the i.i.d. context, Qin and Lawless (1994, Theorem 2) propose the statistic \( 2\ln R(\hat{\theta}_{EL}) - 2\ln R(\theta_0) \) to test the hypothesis \( H_0 : \theta = \theta_0 \) which is shown to be asymptotically distributed as \( \chi^2(p) \) under strong identification. However, due to the dependence on \( \hat{\theta}_{EL} \), this statistic is no longer asymptotically pivotal and thus leads to size-distorted tests under weak identification.

**Example 1 (cont.):** Guggenberger (2003) derives the results given in Theorems 3 and 4 under Assumptions \( \Theta \), ID’, M’ and \( \rho \) allowing for alternatives \( \alpha \in A \) and Pitman drift in the data generating process for the strongly identified parameters to assess the asymptotic power properties of the tests; i.e. ID’ holds and for some fixed \( b \in R^p \), \( y = Y(\theta_0 + n^{-1/2}(0', b') + u \). To simplify our presentation here we ignore the possibility of Pitman drift. Results for the i.i.d. linear IV model follow directly from the above theorems because, as is easily shown, Assumptions ID’, M’, \( \rho \) and \( V(\theta) > 0 \) imply \( M_\theta \) for any consistent estimator \( \hat{\Omega}(\theta) \) of \( \Omega(\theta) \). In particular, \( V(\theta) \) has a simple representation. For \( \theta = (\alpha', \beta_0')', \Omega(\theta) = \Delta(\theta) \) and \( \Delta_{AA}(\theta) = E(V_{iA}V_{iA}' \otimes Z_iZ_i') \), where \( V_{iA} \) consists of the first \( p_A \) components of \( V_i \) in (2.3).

### 4 Subvector Test Statistics

We now assume that interest is focussed on the subvector \( \alpha_0 \in R^{p_A} \) of \( \theta_0 = (\alpha_0', \beta_0')' \). However, we no longer maintain Assumption ID. In particular, \( \alpha_0 \) may not necessarily be weakly identified.

To adapt the test statistics of section 3 to the subvector case, the basic idea is to replace \( \beta \) by a GEL estimator \( \hat{\beta}(\alpha) \). To make this idea more rigorous, define the GEL estimator \( \hat{\beta}(\alpha) \) for \( \beta_0 \)

\[
\hat{\beta}(\alpha) := \arg\min_{\beta \in B} \sup_{\lambda \in \Lambda_n(\alpha', \beta')'} \bar{P}( (\alpha', \beta')', \lambda).
\]  

(4.1)

We usually write \( \hat{\beta} \) for \( \hat{\beta}(\alpha) \) where there is no ambiguity. A requirement of the analysis below is that \( \hat{\beta} \rightarrow_p \beta_0 \) if \( \alpha = \alpha_0 \). Therefore, we assume that the nuisance parameters \( \beta_0 \) which are not involved in the hypothesis under test are strongly identified; see Theorem 2. On the other hand, the components of \( \alpha_0 \) can be weakly or strongly identified and in Assumption ID_\alpha below we assume the former holds for \( \alpha_{01} \)

[15]
and the latter for $\alpha_{02}$, where $\alpha_0 = (\alpha_{01}', \alpha_{02}')'$. The main advantage of the subvector test statistics introduced in this section is that asymptotically they have accurate sizes independent of whether $\alpha_0$ is weakly or strongly identified. This property is not shared by classical tests based on Wald, likelihood ratio or Lagrange multiplier statistics. In general, they have correct size only if $\theta_0$ is strongly identified. In contrast, the subvector tests in Guggenberger and Wolf (2004) based on a subsampling approach have exact asymptotic sizes without any additional identification assumption.

Let $\theta = (\alpha_1', \alpha_2', \beta')'$, where $\alpha_j \in A_j$, $A_j \subset R^{pA_j}$, $(j = 1, 2)$, $p_{A_1} + p_{A_2} = p_A$ and $\beta \in B$, $B \subset R^{p_B}$. Also let $N' \subset A_2 \times B$ be an open neighborhood of $(\alpha_{02}, \beta_0)$.

**Assumption A:** The true parameter $\theta_0 = (\alpha_{01}', \alpha_{02}', \beta_0)'$ is in the interior of the compact space $\Theta$, where $\Theta = A_1 \times A_2 \times B$.

**Assumption ID$_\alpha$:** (i) $E\tilde{g}(\theta) = n^{-1/2}m_{1n}(\theta) + m_2(\alpha_2, \beta)$, where $m_{1n}, m_1 : \Theta \to R^k$ and $m_2 : A_2 \times B \to R^k$ are continuous functions such that $m_{1n}(\theta) \to m_1(\theta)$ uniformly on $\Theta$, $m_1(\theta_0) = 0$ and $m_2(\alpha_2, \beta) = 0$ if and only if $(\alpha_2, \beta) = (\alpha_{02}, \beta_0)$; (ii) $m_2 \in C^1(\Theta)$; (iii) Let $M_2(\cdot) := (\partial m_2/\partial (\alpha_2', \beta'))(\cdot) \in R^{k \times (p_{A_2} + p_B)}$. $M_2(\alpha_{02}, \beta_0)$ has full column rank $p_{A_2} + p_B$.

Assumption ID$_\alpha$ implies $\alpha_{01}$ and $(\alpha_{02}, \beta_0)$ are weakly and strongly identified, respectively. Assumptions A and ID$_\alpha$ adapt Assumptions $\Theta$ and ID in section 3 for the subvector case.

Let

$$\hat{\theta}_a := (\alpha', \beta(\alpha)')'$$

and $\theta_{a_\beta} := (\alpha', \beta')'$.

We now introduce the subvector statistics. Recall the definition of $GELR_p(\theta)$ in (3.2). The $GELR_p$ subvector test statistic is given by

$$GELR_{p_{sub}}(\alpha) := GELR_p(\hat{\theta}_a).$$

We need the following technical assumptions for our derivation of its asymptotic distribution. To obtain theoretical power properties, we again allow a fixed alternative for the weakly identified components, $\alpha_{01}$ here.

For $a_1 \in A_1$ let $a := (a_1', \alpha_{02}')'$ be a fixed vector whose strongly identified component $\alpha_{02}$ is the same as the corresponding component of the true parameter vector $\theta_0$. Let $\mathcal{M} \subset B$ be an open neighborhood of $\beta_0$.

**Assumption M$_\alpha$:** (i) $\max_{1 \leq i \leq n} \sup_{\beta \in B} \|g_i(\theta_{a_\beta})\| = o_p(n^{1/2})$; (ii) $\sup_{\beta \in B} \|\hat{\Omega}(\theta_{a_\beta}) - \Gamma(\theta_{a_\beta})\| \to 0$ for some matrix $\Gamma(\cdot)$ that is uniformly bounded on $\{\theta_{a_\beta} : \beta \in B\}$, continuous at $\theta_{a_\beta}$ and $\Gamma(\theta_{a_\beta}) = \Delta(\theta_{a_\beta}) > 0$ and $n^{-1} \sum_{i=1}^n \|g_i(\theta_{a_\beta}) - g_i(\theta_{a_\beta})'\|^2 = O_p(1)$; (iii) $\Psi_n(\theta_{a_\beta}) \to_d \Psi(\theta_{a_\beta})$, where $\Psi(\theta_{a_\beta}) \equiv N(0, \Delta(\theta_{a_\beta}))$; (iv) $\tilde{G}_B(\cdot) := n^{-1} \sum_{i=1}^n (\partial g_i/\partial \beta)(\cdot)$ exists at $\theta_{a_\beta}$ a.s. for each $\beta \in \mathcal{M}$, $\tilde{g}(\theta_{a_\beta})$ is integrable for all $\beta \in \mathcal{M}$, $\sup_{\beta \in \mathcal{M}} \|\tilde{G}_B(\theta_{a_\beta})\|$ is integrable, $\partial m_{1n}/\partial \beta(\cdot)$ is continuous at $\theta_{a_\beta}$ a.s. for each $\beta \in \mathcal{M}$ and $\partial m_{1n}/\partial \beta(\theta_{a_\beta})$ converges uniformly over $\beta \in \mathcal{M}$ to some function; (v) $\hat{g}(\theta_{a_\beta}) \to_p E\hat{g}(\theta_{a_\beta})$ uniformly over $\beta \in B$, $\hat{G}_B(\theta_{a_\beta}) \to_p E\hat{G}_B(\theta_{a_\beta})$ uniformly over $\beta \in B$. 

[16]
\[ \beta \in M; \ (vi) \ \sup_{\beta \in M} n^{-1} \sum_{i=1}^{n} ||G_{iB}(\theta_{ab})|| = O_p(1). \]

Mutatis mutandis \( M_{\alpha} \) has the same interpretation as \( M_{\theta} \). For example \( M_{\alpha}(ii) \) guarantees that \( \lambda_{\text{max}}(\Omega(\theta_{ab})) \) is bounded and \( \lambda_{\text{min}}(\Omega(\theta_{ab})) \) is bounded away from zero w.p.a.1 while \( M_{\alpha}(iv) \) and ID\( \alpha \) imply that for \( \beta \in M \) we have \( E\tilde{G}_{B}(\theta_{ab}) = n^{-1/2}(\partial m_{1n}/\partial \beta)(\theta_{ab}) + (\partial m_{2}/\partial \beta)(\alpha_{02}, \beta) \to (\partial m_{2}/\partial \beta)(\alpha_{02}, \beta) \). By ID\( \alpha \) this last matrix has full column rank for \( \beta = \beta_0 \). If we assume that the \( G_{iB}(\theta_{ab}) \), \( (i = 1, ..., n) \), viewed as functions of \( \beta \), are continuous at \( \beta_0 \) a.s., then we can simplify \( M_{\alpha}(vi) \) to \( n^{-1} \sum_{i=1}^{n} ||G_{iB}(\theta_{a0})|| = O_p(1) \). A similar comment holds for the assumptions in the continuation of \( M_{\alpha} \) below.

**Theorem 5** Assume \( 1 \leq p_A < p \). Suppose Assumptions A, ID\( \alpha \), \( M_{\alpha}(i)-(vi) \) and \( \rho \) hold for some \( a_1 \in A_1 \) and \( a = (a_1', \alpha_{02}')' \). Then,

\[ GELR_{\rho}^{\text{sub}}(a) \rightarrow_d \chi^2(k - p_B, \delta), \]

where the noncentrality parameter \( \delta \) is given by

\[ \delta := m_1(\theta_{a0}')\Delta(\theta_{a0})^{-1}M_{M_{2\beta}(a_{02}, \beta_0)}(\Delta(\theta_{a0}))(\theta_{a0}). \]

where \( M_{2\beta}(\cdot) := (\partial m_{2}/\partial \beta)(\cdot) \in R^{k \times p_B} \). In particular,

\[ GELR_{\rho}^{\text{sub}}(\alpha_0) \rightarrow_d \chi^2(k - p_A). \]

Theorem 5 confirms that the subvector statistic \( GELR_{\rho}^{\text{sub}}(\alpha_0) \), like the full vector statistic \( GELR_{\rho}(\theta_0) \), is asymptotically pivotal. As above, this result can be used to construct hypothesis tests and confidence regions for \( \alpha_0 \).

We now generalize the statistics \( S_{\rho} \) and \( LM_{\rho} \) to the subvector case. The asymptotic variance matrices of \( n^{1/2}\hat{\varphi}(\hat{\theta}_{\alpha}) \) and \( n^{1/2}\lambda(\hat{\theta}_{\alpha}) \) differ from those of \( n^{1/2}\hat{\varphi}(\theta_{a0}) \) and \( n^{1/2}\lambda(\theta_{a0}) \). Therefore different weighting matrices are required in the quadratic forms defining these subvector statistics. In the Appendix (proof of Theorem 6) it is shown that for \( a = (a_1', \alpha_{02}')' \), \( \lambda(\hat{\theta}_{a}) = \arg \max_{\lambda \in \Lambda_n(\hat{\theta}_{a})} P(\theta_a, \lambda) \) exists w.p.a.1 and that \( n^{1/2}\lambda(\hat{\theta}_{a}) \) is asymptotically normal with covariance matrix \( M(a) \), where for \( \alpha = (\alpha_1', \alpha_{02}')' \in R^{p_A} \)

\[ M(\alpha) := \Delta(\theta_{a0})^{-1}M_{M_{2\beta}(a_{02}, \beta_0)}(\Delta(\theta_{a0})). \tag{4.2} \]

The first \( p_A \) elements of the FOC (3.3), evaluated at \( \hat{\theta}_{a} \), are

\[ \lambda(\hat{\theta}_{a})' \sum_{i=1}^{n} \rho_1(\lambda(\hat{\theta}_{a})'g_i(\hat{\theta}_{a}))(G_{iA}(\hat{\theta}_{a})/n = 0'. \tag{4.3} \]

For \( \alpha \in R^{p_A} \), let

\[ D_{\rho}(\alpha) := \sum_{i=1}^{n} \rho_1(\lambda(\hat{\theta}_{a})'g_i(\hat{\theta}_{a}))(G_{iA}(\hat{\theta}_{a})/n \in R^{k \times p_A}, \tag{4.4} \]

[17]
which coincides with the definition of $D_\rho(\theta)$ (3.4) when $\alpha$ is the full vector $\theta$. Similarly to $S_\rho(\theta)$ (3.5) the subvector test statistic $S^\text{sub}_\rho(\alpha)$ is constructed as a quadratic form in the vector $\lambda(\hat{\theta}_\alpha)'D_\rho(\hat{\theta}_\alpha)$ from (4.3) with weighting matrix given by $M(\alpha)$ in (4.2). Let $\tilde{M}(\alpha)$ be an estimator of $M(\alpha)$ that is given by replacing the expressions $\Delta(\theta_{\alpha\beta_0})$ and $M_{23}(\alpha_2, \beta_0)$ in $M(\alpha)$ by consistent estimators, $\hat{\Omega}$ and $\tilde{M}_2$ say. By Assumptions $M_\alpha(\text{ii})$ and $M_\alpha(\text{iv})-(\text{v})$ we may choose $\hat{\Omega}(\hat{\alpha})$ for $\hat{\Omega}$ and $\tilde{G}_B(\hat{\alpha})$ for $\tilde{M}_2$ when $\alpha = a = (a_1', a_0'_{02})'$. Hence,

$$S^\text{sub}_\rho(\alpha) := n\lambda(\hat{\theta}_\alpha)'D_\rho(\alpha)\left(D_\rho(\alpha)'\tilde{M}(\alpha)D_\rho(\alpha)\right)^{-1}D_\rho(\alpha)'\lambda(\hat{\theta}_\alpha).$$

The statistic $LM^\text{sub}_\rho(\alpha)$ is constructed like $S^\text{sub}_\rho(\alpha)$ but replaces $\lambda(\hat{\theta}_\alpha)$ by $-\hat{\Omega}^{-1}\tilde{g}(\hat{\theta}_\alpha)$.

Thus,

$$LM^\text{sub}_\rho(\alpha) := n\tilde{g}(\hat{\theta}_\alpha)'\hat{\Omega}^{-1}D_\rho(\alpha)\left(D_\rho(\alpha)'\tilde{M}(\alpha)D_\rho(\alpha)\right)^{-1}D_\rho(\alpha)'\hat{\Omega}^{-1}\tilde{g}(\hat{\theta}_\alpha).$$

Let $a = (a_1', a_0'_{02})'$, $\mathcal{M} \subset \mathcal{B}$ be an open neighborhood of $\beta_0$ and $\tilde{G}_{A_1}(\theta) := n^{-1}\sum_{i=1}^n (\partial g_i/\partial \alpha_j)(\theta), (j = 1, 2)$.

**Assumption $M_\alpha$: (cont.)** (vii) $\tilde{G}_{A_1}(\theta_{a\beta})$ viewed as a function in $\beta$ is continuously differentiable at $\beta$ a.s. for each $\beta \in \mathcal{M}$, $(\partial vec\tilde{G}_{A_1}/\partial \beta)(\theta_{a\beta}) \rightarrow_p E(\partial vec\tilde{G}_{A_1}/\partial \beta)(\theta_{a\beta}) = (\partial E vec\tilde{G}_{A_1}/\partial \beta)(\theta_{a\beta}), \tilde{G}_{A_1}(\theta_{a\beta}) \rightarrow_p E\tilde{G}_{A_1}(\theta_{a\beta}) = (\partial E\tilde{g}(\theta_{a\beta})/\partial \alpha)(\theta_{a\beta}), (\partial vecm_{1n}/\partial \alpha_j)(\theta_{a\beta}) \rightarrow_p (\partial vec(\partial m_{1n}/\partial \alpha_j)/\partial \beta)(\theta_{a\beta})$, where convergence is uniform over $\beta \in \mathcal{M}$ in all cases, $\partial m_{1n}/\partial \alpha_j$ is continuous at $\theta_{a\beta}$ a.s. for each $\beta \in \mathcal{M}$ and $\partial m_{1n}/\partial \alpha_j$ converges uniformly over $\beta \in \mathcal{M}$ to some function; (viii) $n^{-1}\sum_{i=1}^n (vecG_{iA_1}(\theta_{a\beta}))g_i(\theta_{a\beta}) \rightarrow_p \Phi(\theta_{a\beta})$ uniformly over $\beta \in \mathcal{M}$ for some matrix $\Phi(\cdot)$ that is continuous at $\theta_{a\beta_0}$ and satisfies $\Phi(\theta_{a\beta_0}) = \Delta_{A_1}(\theta_{a\beta_0})$ ($\Delta_{A_1}(\cdot)$ is defined in (x) below). Theorem $\tilde{\Omega}(\hat{\alpha}) \rightarrow_p \Delta(\theta_{a\beta_0})$; (ix) $n^{-1}\sum_{i=1}^n ||G_{iA_1}(\theta_{a\beta})|| ||g_i(\theta_{a\beta})|| = O_p(1), n^{-3/2}\sum_{i=1}^n ||G_{iA_2}(\theta_{a\beta})|| ||g_i(\theta_{a\beta})|| = O_p(1)$ uniformly over $\beta \in \mathcal{M}$; (x) $n^{-1/2}\sum_{i=1}^n (vec(G_{iA_1}(\theta_{a\beta}) - E\tilde{G}_{iA_1}(\theta_{a\beta})))'g_i(\theta_{a\beta}) \rightarrow_d N(0, V^\alpha(\theta_{a\beta})))$, where $V^\alpha(\theta_{a\beta})$ is the appropriate submatrix of $V(\theta_{a\beta})$ defined in $M_\alpha(\text{vii})$. $V^\alpha(\theta_{a\beta})$ is positive definite.

In $M_\alpha(\cdot)$ write

$$V^\alpha(\theta) = \left(\begin{array}{cc}
\Delta_{A_1} & \Delta_{A_1} \\
\Delta_{A_1}' & \Delta
\end{array}\right)(\theta), \text{ where } \Delta_{A_1A_2}(\theta) \in R^{pA_1 \times pA_1}.$$
Theorem 6 Assume $1 \leq p_A < p$. Suppose Assumptions A, ID, $M_\alpha(i)\sim(x)$ and $\rho$ hold for $a = (a_1', a_2')$ for $a_1 \in A_1$. Then,

$$S^a_\rho(a), L M^a_\rho(a) \to \mathcal{D} (W_\alpha(a) + \zeta_\alpha)'(W_\alpha(a) + \zeta_\alpha),$$

where the random $p_A$-vector $W_\alpha(a)$ is defined in (A.22) of the Appendix, $\zeta_\alpha \sim N(0, I_{p_A})$ and $\zeta_\alpha$ and $W_\alpha$ are independent. We have $W_\alpha(\alpha_0) \equiv 0$ and therefore

$$S^\alpha_\rho(\alpha_0), L M^\alpha_\rho(\alpha_0) \to \chi^2(p_A).$$

Remark 1: The subvector statistics are asymptotically pivotal when elements of $\alpha_0$ are arbitrarily weakly or strongly identified. This result can be used for the construction of test statistics or confidence regions that have correct size or coverage probabilities asymptotically, independent of the strength or weakness of identification of $\alpha_0$. Compared to Kleibergen’s (2001) GMM-subvector statistic the statistics $S^\alpha_\rho(\alpha_0)$ and $L M^\alpha_\rho(\alpha_0)$ are appealing due to their compact formulation.

Remark 2: Even though it is unclear how the asymptotic distribution of these test statistics might be derived without assuming strong identification of $\beta_0$, it is obvious neither $S^\alpha_\rho(\alpha_0)$ nor $L M^\alpha_\rho(\alpha_0)$ would converge to a $\chi^2(p_A)$ random variable.

In general the quantities $n^{1/2}g(\widehat{\theta}_\alpha)$ in $S^\rho_\alpha(\alpha_0)$ and $n^{1/2}g(\widehat{\theta}_\alpha)$ in $L M^\rho_\alpha(\alpha_0)$ are no longer asymptotically normal because of their dependence on the GEL estimator $\widehat{\beta}(\alpha_0)$, which as a direct consequence of Theorem 2 has a nonstandard limiting distribution if $\beta_0$ is not strongly identified. Moreover, the subvector version of Kleibergen’s (2001) $K$-statistic also experiences the same problem in these circumstances as the (GMM) CUE of $\beta_0$ has a nonnormal limiting distribution under weak identification; see SW. Somewhat surprisingly, however, Monte Carlo simulations by the authors (not reported here) for the subvector statistic $L M^\alpha_\rho(\alpha_0)$ indicate that its size properties are not much affected by the strength or weakness of identification of $\beta_0$. Startz, Zivot and Nelson (2004) report similar findings from Monte Carlo simulations for Kleibergen’s (2001) subvector test statistic.

Example 1 (cont.): Guggenberger (2003) derives the corresponding results. Note that Assumptions $\Theta$, ID, $M'$, $\rho$ and also assuming that $V^\alpha(\theta_a\beta_0)$ is full column rank imply Assumption $M_\alpha$. In the linear model the components of $V^\alpha(\theta_a\beta_0)$ can be easily calculated. For example, $\Delta_{A_1A_1} = E(V_{iA_1}V_{iA_1}' \otimes Z_iZ_i')$, where $V_{iA_1}$ is the subvector of $V_i$ that contains its first $p_A$ components. Let $Y = (X, W)$ denote the partition of the included variables of the structural equation into exogenous and endogenous variables. Partition $\theta_0 = (\theta'_X, \theta'_W)'$ and $\theta = (\theta'_X, \theta'_W)'$ conformably. Valid inference is possible on any subvector of $\theta_W$ if the appropriate assumptions given above are fulfilled. Unfortunately, if the dimension of the parameter vector not subject to test is large, then the argmin-sup problem in (4.1) is computationally very involved. Premultiplication of equation (2.2) by $M_X$ should ameliorate this problem through the elimination of the exogenous variables; i.e. $M_Xy = M_XW\theta_W + M_Xu$. If Assumption $M_\alpha$ holds for $\theta_W = (\alpha_W, \beta_W)$ and $g_i(\theta_W) := M'_{X, i}(y - W\theta_W)Z_i$, where $M_{X, i}$ denotes the $i^{th}$ row of $M_X$ written as a column vector, valid inference may be undertaken on $\alpha_W$. 

[19]
5 Simulation Evidence

To assess the efficacy of the hypothesis tests introduced in Theorems 3 and 4, we conduct a set of Monte Carlo experiments. The data generating process (DGP) is given by model (2.2) considered in Example 1 above and is similar to that in Kleibergen (2002a, p.1791); viz.

\[
\begin{align*}
y & = Y\theta_0 + u, \\
Y & = Z\Pi + V.
\end{align*}
\]

There is a single right hand side endogenous variable and no included exogenous variables, \( p = 1 \), \( Z \sim N(0, I_k \otimes I_n) \), where \( k \) is the number of instruments and \( n \) the sample size. In the just-identified case, that is, \( k = 1, \Pi = \Pi_1 \) whereas, in the over-identified case, \( k > 1, \Pi = (\Pi_1, 0)' \), i.e. irrelevant instruments are added.

Interest focuses on testing the scalar null hypothesis \( H_0 : \theta_0 = 0 \) versus the alternative hypothesis \( H_1 : \theta_0 \neq 0 \).

5.1 Error Distributions

We examine several distributions for \((u, V)\) to investigate the robustness of the test statistics to potentially different features of the error distribution. All designs are constructed from Design (I) obtained by modifying the distribution of the structural error \( u \).

- Design (I): \((u, V)' \sim N(0, \Sigma \otimes I_n)\), where \( \Sigma \in \mathbb{R}^{2 \times 2} \) with diagonal elements unity and off-diagonal elements \( \rho_{uV} \).
- Design (II): \( u_i \) in Design (I) is modified as \( u_i/(w_i/r)^{1/2} \), where \( w_i \) is a \( \chi^2(r) \) random variable independent of \( u_i \) and \( V_i \), i.e. \( u_i \) is \( t_r \)-distributed. We fix \( r = 2 \).
- Design (III): modifies Design (I) by exchanging \( u_i^2 - 1 \) for \( u_i \), i.e. \( u_i \) is a recentered \( \chi^2(1) \) random variable.
- Design (IV): \( u_i \) from Design (I) is replaced by \( B_i \left| u_i + 2 \right| - (1 - B_i) \left| u_i + 2 \right| \) where \( B_i \) is Bernoulli (.5,.5) distributed and independent of all other random variables.

Design (II) examines the robustness of the performance of the test statistics to thick tailed distributions for the structural equation error. Design (III) examines robustness with respect to asymmetric structural error distributions. In Design (IV) the structural error \( u_i \) is bimodal with peaks at \(-2\) and \(+2\).

In addition, the impact of conditional heteroskedasticity on the performance of the test statistics is examined. Designs \((I_{HET})-(IV_{HET})\) modify Designs (I)-(IV) respectively replacing \( u_i \) by \( u_i = ||Z_i||u_i \).
5.2 Test Statistics

We calculate three versions of the statistic $\text{GELR}_p(\theta)$ \((3.2)\), for \(\rho(v) = -(1 + v)^2/2\) (CUE), \(\rho(v) = \ln(1 - v)\) (EL) and \(\rho(v) = -\exp v\) (ET). We also consider the corresponding versions for each of \(S_p(\theta)\) \((3.5)\) and \(LM_p(\theta)\) \((3.6)\) with \(\hat{\Omega}(\theta)\) replaced by \(\hat{\Omega}(\theta)\). As noted above, for CUE, \(S_p(\theta)\) and \(LM_p(\theta)\) are then numerically identical. Theorems 3 and 4 present the asymptotic null distributions of these statistics.\(^8\)

Additional statistics considered are the Anderson-Rubin test statistic \((AR)\), see Anderson and Rubin \((1949)\), two versions of the \(K\)-statistic proposed by Kleibergen \((2001, 2002a)\), one assuming homoskedastic errors \(K\), the other robust to conditional heteroskedasticity \(K_{HET}\), Moreira’s \((2003)\) conditional likelihood ratio test \(LR_M\) and two versions of the two-stage least squares \((2SLS)\) Wald statistic \(2SLS\), see for example Wooldridge \((2002, \text{p. } 98, 100)\), one assuming homoskedastic errors \((2SLS_{HOM})\) and the other robust to conditional heteroskedasticity \((2SLS_{HET})\).\(^9\) Under \(H_0 : \theta = \theta_0\), \(AR(\theta_0) \rightarrow_d \chi^2(k)\) and \(K(\theta_0) \rightarrow_d \chi^2(p)\). In the just-identified case \(k = p = 1\), the \(AR\)- and \(K\)-statistics coincide. Both Wald statistics are asymptotically distributed as \(\chi^2(1)\) under \(H_0 : \theta = \theta_0\) and strong identification.

\(^8\)To calculate \(\text{GELR}_p(\theta), S_p(\theta)\) and \(LM_p(\theta)\) for EL and ET, the globally concave maximization problem \(\max_{\lambda \in \lambda_n(\theta)} P(\theta, \lambda)\) must be solved numerically. To do so we implement a variant of the Newton-Raphson algorithm. We initialize the algorithm by setting \(\lambda\) equal to the zero vector. At each iteration the algorithm tries several shrinking stepsizes in the search direction and accepts the first one that increases the function value compared to the previous value for \(\lambda\). This procedure enforces an “uphill climbing” feature of the algorithm.

\(^9\)The statistics are defined as follows:

\[
AR(\theta) := (y - Y\hat{\theta})' P_Z (y - Y\hat{\theta})/s_{uu}(\theta),
\]

where \(s_{uu}(\theta) := (y - Y\hat{\theta})' M_Z (y - Y\hat{\theta})/(n - k),\)

\[
K(\theta) := (y - Y\hat{\theta})' P_{V(\theta)} (y - Y\hat{\theta})/s_{uu}(\theta),
\]

where \(\tilde{Y} := Z\tilde{\Omega}(\theta), \tilde{\Omega}(\theta) = (Z'Z)^{-1} Z'[Y - (y - Y\hat{\theta}) s_{uu}(\theta)/s_{uu}(\theta)]\) and \(s_{uu}(\theta) := (y - Y\hat{\theta})' M_Z (y - Y\hat{\theta})/(n - k).\) The statistic \(K(\theta)\), Kleibergen \((2002a)\), is not robust to conditional heteroskedasticity. However, a version of the \(K\)-statistic in Kleibergen \((2001, \text{equation (22)})\) that uses a het-

\[2SLS_{HOM} := \hat{\theta} W^{-1} \hat{\theta}, 2SLS_{HET} := \hat{\theta} W_{HET}^{-1} \hat{\theta},\]

where \(\hat{\theta} := (Y' P_{Z2} Y)^{-1} Y' P_{Z2} y, W := \hat{\sigma}^2 (Y' P_{Z2} Y)^{-1}, \hat{\sigma}^2 := (n - k)^{-1} \sum_{i=1}^{n} \hat{u}_i^2, \hat{u}_i := y_i - Y' \hat{\theta}, (i = 1, \ldots, n)\) and \(W_{HET} := n (Y' P_{Z2} Y)^{-1} \sum_{i=1}^{n} (\hat{u}_i^2 P_{Z2} Y)^2/(n - k)\) is a conditional heteroskedasticity robust estimator for the variance of \(\hat{\theta}\).
5.3 Size Comparison

Empirical sizes are calculated using 5% asymptotic critical values for all of the above statistics for DGPs (5.1) corresponding to all 54 possible combinations of sample size $n = 50, 100, 250$, number of instruments $k = 1, 5, 10$, structural and reduced form error correlation $\rho_{uV} = 0.0, 0.5, 0.99$ and reduced form coefficient $\Pi_1 = 0.1, 1.0$ for Designs (I)-(IV) and (I_HET)-(IV_HET).\footnote{Kleibergen (2002a) generates one sample for the instrument matrix $Z$ from a $N(0, I_k \otimes I_n)$ distribution and then keeps $Z$ fixed across $R = 10,000$ samples of the DGP (5.1) using Design (I) with $n = 100$ and $\rho_{uV} = .99$. We simulate a new matrix $Z$ with each sample of the DGP (5.1). As a consequence, our results do not coincide with those reported by Kleibergen (2002a).}

We use $R = 3,000$ replications of each DGP. We also use 3,000 realizations each of $\chi^2(1)$ and $\chi^2(k-1)$ random variables to simulate the critical values of Moreira’s $LR_M$ statistic. For the results reported in the Tables 1 and 2 below we use $R = 10,000$ replications. We refer to $\Pi_1 = 0.1$ and 1.0 as the “weak” and “strong” instrument cases, respectively. The value of $\rho_{uV}$ allows the degree of endogeneity of $Y$ to be varied. While for $\rho_{uV} = 0$, $Y$ is exogenous, $Y$ is strongly endogenous for $\rho_{uV} = .99$. We include the just-identified case, $k = 1$ and two over-identified-cases, $k = 5$ and 10.

We now turn to describe the results for Designs (I) and (I_HET) given in Tables 1 and 2, respectively, which exclude those for $GELR_{EL}$, $S_{EL}$, $LM_{ET}$, $AR$ and the case $n = 100$. The qualitative features of the size results for $GELR_{EL}$, $S_{ET}$ and $LM_{ET}$ are identical to their $ET/EL$ counterparts. For $k = 1$, $AR$ coincides with $K$ and, for $k > 1$, we find that in most cases $K$ has better size properties than $AR$. We report $K$ and $2SLS_{HOM}$ for the homoskedastic and $K_{HET}$ and $2SLS_{HET}$ for the heteroskedastic design. We now discuss the results for the homoskedastic case of Design (I).

**Table 1 about here**

Firstly, we consider the separate effects of $\Pi_1$, $n$, $\rho_{uV}$ and $k$ on the size results.

The most important finding is that the empirical sizes of all statistics except 2SLS show little or no dependence on $\Pi_1$; (some additional Monte Carlo results show that this even holds true for the completely unidentified case where $\Pi_1 = 0$). However, those for 2SLS depend crucially on the strength or weakness of identification. While for $\Pi_1 = 1.0$, 2SLS has reliable size properties for many cases, with weak instruments sizes range over the entire interval, 0% to 100%.

In general, increasing $n$ leads to more accurate size across all statistics. This holds especially true for those that are poor for smaller $n$. For example, the 2SLS statistics, $GELR_{ET}$ and $S_{EL}$ severely over-reject in over-identified and strongly endogenous
cases when $n = 50$. Even though they still over-reject for $n = 250$, the rejection rates are much closer to the 5% significance level.

It is easily shown that the rejection rates under the null hypothesis for $AR$ and $GELR_\rho$ are independent of the value of $\rho_{uV}$. The slight dependence of the size results in Table 1 on $\rho_{uV}$ results from the use of different samples. For all the remaining statistics except for $2SLS$, there does not appear to be a clear pattern for how $\rho_{uV}$ affects their size properties. Moreover, there is little dependence of the results on $\rho_{uV}$. However, for $2SLS$, increasing $\rho_{uV}$ leads to severe over-rejection when combined with over-identification, especially so in the weak instrument case.

Increasing the number of instruments $k$ usually leads to over-rejection for $2SLS$, $GELR_{ET}$ and $S_{EL}$. For $2SLS$ this is especially true under weak identification and/or strong endogeneity. All the other statistics show little dependence on $k$.

We now turn to a comparison of performance across statistics. The $2SLS$ statistics should not be used with weak instruments or in strongly endogenous over-identified situations. In all other cases, $2SLS$ has competitive size properties. The statistics $GELR_{ET}$ and $S_{EL}$ severely over-reject in over-identified problems when the sample size is small. Overall, then, the statistics $LM_{EL}$, $K$ and $LR_M$ lead to the best size results. The statistics $LM_{CUE}$ and $GELR_{CUE}$ come in only as second winners because they tend to under-reject, especially in over-identified situations. Across the 36 experiments in Table 1, the sizes of $LM_{EL}$, $LM_{CUE}$, $GELR_{CUE}$, $K$ and $LR_M$ are in the intervals $[4.0,6.2]$, $[1.6,5.3]$, $[1.3,5.3]$, $[4.8,8.6]$ and $[4.3,10.3]$, respectively. The statistics $K$ and $LR_M$ usually slightly over-reject. In 22 of the 36 cases, the size of $LM_{EL}$ comes closest to the 5% significance level across all the statistics. The corresponding numbers for $LM_{CUE}$, $GELR_{CUE}$, $K$ and $LR_M$ are 8, 8, 9 and 7. Based on Design (I), $LM_{EL}$ seems to have a slight advantage over the remaining statistics.

Table 2 about here

We now discuss the size results for Design (I$_{HET}$) summarized in Table 2. As most findings are similar to those discussed for Design (I), we only describe the new features.

The statistics $2SLS_{HOM}$, $K$ and $LR_M$ perform uniformly worse than in Design (I). Tests based on these statistics severely over-reject, especially in the just-identified case. Their performance does not improve when $n$ increases. We therefore report results for the heteroskedasticity robust versions $2SLS_{HET}$ and $K_{HET}$. Their size properties and those of the statistics based on GEL methods do not appear to be negatively influenced by the presence of conditional heteroskedasticity. This is to be expected from our earlier theoretical discussion of the GEL statistics which does not assume conditional homoskedasticity. Of course, $2SLS_{HET}$ still suffers in weakly identified models and $GELR_{ET}$ and $S_{EL}$ perform poorly in over-identified situations for small $n$. Rejection rates of the test statistics $LM_{EL}$, $LM_{CUE}$, $GELR_{CUE}$, $K_{HET}$ and $LR_M$ across the 36 experiments of Table 2 are in the intervals $[3.6,6.4]$, $[1.6,5.1]$, $[1.0,5.1]$, $[4.3,9.2]$ and $[7.8,28.8]$, respectively. In 21 of the 36 cases, the size of $LM_{EL}$ comes closest to the 5% significance level across all the statistics. $K_{HET}$ wins in 18 cases.
In summary, the only statistics with accurate size properties across all experiments of Designs (I) and (I_HET) are \(LM_{EL}, LM_{CUE}, GELR_{CUE}\) and \(K_{HET}\). Based on the above results it seems that \(LM_{EL}\) enjoys a slight advantage over the other statistics. From the 72 cases in Tables 1 and 2 the empirical size of \(LM_{EL}\) is closest to the nominal 5% in 43 cases across all statistics.

The qualitative features of the size results for Designs (II)-(IV) and (II_HET)-(IV_HET) are generally very similar to their normal counterparts of Designs (I) and (I_HET). For this reason, we do not include additional tables for these designs. One striking difference however occurs for 2SLS under weak identification with \(\chi^2(1)\) (Design (III)) and bimodal errors (Design (IV)). Rejection rates across these 54 combinations for 2SLS\(_{HOM}\) are in the intervals \([0.1,7.1]\) and \([0.0,5.4]\), respectively. While with normal errors and weak identification 2SLS severely over-rejects, with these error distributions it severely under-rejects.

To summarise this size study, \(LM_{EL}, LM_{CUE}, GELR_{CUE}\) and \(K_{HET}\) have reliable size properties across all designs which appear independent of both the strength or weakness of identification and possible conditional heteroskedasticity. 2SLS performs very poorly in the presence of weak instruments. The \(LR_M\) statistic performs well in homoskedastic cases but poorly otherwise.

### 5.4 Power Comparison

Empirical power curves are calculated for the above statistics and DGPs (5.1) corresponding to all 16 possible combinations of sample size \(n = 100, 250\), number of instruments \(k = 5, 10\), structural and reduced form error correlation \(\rho_{UV} = 0.5, 0.99\) and reduced form coefficient \(\Pi_1 = 0.1, 1.0\) for each of the error distributions of Designs (I)-(III). Except for \(LR_M\), we report size-corrected power curves at the 5\% significance level, using critical values calculated in the size comparison above. We do so because size-correction of \(LR_M\) is not straightforward due to the conditional construction of \(LR_M\) and, as shown above, for Designs (I)-(III), \(LR_M\) has empirical size very close to nominal at the 5\% significance level.

We use \(R = 1,000\) replications from the DGP (5.1) with various values of the true value \(\theta_0\). The null hypothesis under test is again \(H_0: \theta_0 = 0\). For weak identification \((\Pi_1 = 0.1), \theta_0\) takes values in the interval \([-4.0, 4.0]\) while, with strong identification \((\Pi_1 = 1.0), \theta_0 \in [-0.4, 0.4]\). We use 1,000 realizations each of \(\chi^2(1)\) and \(\chi^2(k-1)\) random variables to simulate the critical values of \(LR_M\). For those results reported in the figures below, we use 10,000 replications from (5.1).

Detailed results are presented only for the statistics \(LM_{EL}, K, LR_M\) and 2SLS\(_{HET}\). The statistics \(LM_{CUE}, LM_{EL}\) and \(LM_{ET}\) display a very similar performance across almost all scenarios. We therefore only report results for \(LM_{EL}\). We do not report power results for the statistics \(S_{EL}\) and \(S_{ET}\) because, as seen above, their size properties appear to be quite poor for the sample sizes considered here. When \(k = 1\), \(AR\) and \(K\) are numerically identical. In over-identified cases, \(K\) generally performs better than \(AR\). We therefore do not report results for \(AR\); see Kleibergen (2002a) for a comparison of \(K\) and \(AR\). Similarly, \(GELR_{CUE}\) is numerically identical to \(LM_{\rho}\) for \(k = 1\) but leads to a less powerful test for \(k > 1\). Also EL and ET versions of

[24]
for positive \( \rho_u \) reaches values close to 100\% when

\[
\theta_0 = 0
\]

In Designs (I) and (II), \( 2SLS_{HET} \) is usually biased, taking on its lowest value at a negative \( \theta_0 \) value in the interval \([-0.2, 0.0]\). When \( \theta_0 \) is weakly identified, the power curves of \( LM_{EL}, K \) and \( LR_M \) are generally very flat across all \( \theta_0 \) values, often only slightly exceeding the significance level of the test. This is especially true for \( LM_{EL} \) and \( K \) but less so for \( LR_M \) which is generally more powerful than the other two statistics in this situation. There is one exception when the power of the three tests is high. In Design (I) with \( \rho_u = 0.99 \), while being flat at about 5\% for positive \( \theta_0 \) values, the power curves reach a sharp peak of almost 100\% around \( \theta_0 = -1 \). The reason for this anomaly is most easily explained in the case \( k = 1 \), where

\[
LM_{EL}(0) = GELR_{CUE}(0) = n g(0) \Omega(0)^{-1} g(0).
\]

We have \( \Omega(0) = E(u_1 + Y_i \theta_0)^2 \), which in Design (I) with \( \Pi_1 = 0.1 \) equals \( 1 + 2 \theta_0 \rho_u + (1.01) \theta_0^2 \). If \( \rho_u = .99 \) this expression is minimized at around \( \theta_0 = -.98 \) where it equals approximately .03. Therefore, this peak is caused by \( \Omega(0)^{-1} \) taking on large values for \( \theta_0 \) in the neighborhood of \(-1\).

For negative \( \theta_0 \) values with \( |\theta_0| > 1 \) power quickly falls, reaching between 20\% and 50\% across the different designs at \( \theta_0 = -4 \).

In contrast to the power curves of \( LM_{EL}, K \) and \( LR_M \), the power curve of \( 2SLS_{HET} \) retains its U-shaped form for \( \Pi_1 = 0.1 \). In many cases, the power curve reaches values close to 100\% when \( |\theta_0| \) is close to 4.

As to be expected the tests are more powerful when \( n \) is increased from 100 to 250. This holds uniformly across all statistics and designs with a more pronounced power increase in the strongly identified cases.

There does not seem to be a systematic effect due to \( \rho_u \) as it varies with the specific design. For reasons explained above, the shape of the power curves can change dramatically in Design (I) when \( \rho_u \) is increased from 0.5 to 0.99 if \( \Pi_1 = 0.1 \).

In most cases, there is only little change in the power functions when \( k \) is increased from 5 to 10. In general, if the power function changes, then power is slightly lower for larger \( k \).

We now compare the power functions across statistics. Figures (i)-(iii) display the power curves of the four statistics for Designs (I)-(III) in the case \( \Pi_1 = 1.0, n = 250, \rho_u = 0.5 \) and \( k = 5 \) (the figures for \( \Pi_1 = 0.1 \) and for the other parameter combinations are available upon request). The qualitative comparison for the other parameter combinations is very similar and we therefore focus on these representative cases.

When identification is weak, the test based on \( LR_M \) is usually more powerful than those based on \( LM_{EL} \) and \( K \). The power gain of using \( LR_M \) is quite substantial for negative \( \theta_0 \) values but less so for positive \( \theta_0 \). However, the Wald test \( 2SLS_{HET} \) is
by far the most powerful test in all three designs. Except for some small negative
\( \theta_0 \) values its power curve uniformly dominates the power curves of the other tests. 
Recall though that 2SLS\(_{HET}\) has unreliable size properties under weak identification.

When identification is strong, \( LM_{EL} \) uniformly dominates \( LR_M \) and \( K \) in Designs 
(II) and (III), see Figures (ii) and (iii). However, \( LR_M \) and \( K \) uniformly dominate 
\( LM_{EL} \) in Design (I), see Figure (i). This result is to be expected. On the one hand, 
the \( LM_{EL} \) test is based on nonparametric GEL methods. On the other hand, \( LR_M \) 
and \( K \) are motivated within the normal model framework. While the power gain 
of \( LM_{EL} \) is small in Design (III), it is substantial in Design (II). Therefore, \( LM_{EL} \) 
should be used when errors have thick tails.

With strong identification, the Wald test is the most powerful test for positive \( \theta_0 \) 
values. For negative \( \theta_0 \) values, its performance varies from being most powerful in 
Design (III) to least powerful in Design (I). These results confirm that the Wald test 
is a reasonable choice when identification is strong.

Overall, therefore, the power study does not lead to an unambiguous ranking of 
the different tests considered here. Which test is most appropriate depends on the 
particular error distribution and degree of identification. We find that with strong 
identification and errors with thick tails or asymmetric errors, \( LM_{EL} \) seems to be 
the best choice while with normal errors \( LR_M \) and \( K \) appear preferable. When 
identification is weak, \( LR_M \) generally dominates \( K \) and \( LM_{EL} \) in terms of power 
although as noted above the size properties of \( LR_M \) deteriorate substantially in the 
presence of heteroskedasticity.

### Appendix: Proofs

**Proof of Eq. (2.4).** Let \( f_i := \sup_{\theta \in \Theta} ||g_i(\theta)|| \). Define \( K := \sup_{i \geq 1} E f_i^{1/2} < \infty \). Let 
\( \varepsilon > 0 \) and choose a positive \( C \in R \) such that \( K/C < \varepsilon \). Then

\[
\Pr(\max_{1 \leq i \leq n} f_i) n^{-1/2} \leq \max_{1 \leq i \leq n} \Pr(f_i > nC) \leq \sum_{i=1}^n \frac{1}{nC} E(f_i^2) \leq K/C < \varepsilon,
\]

where the first inequality follows from \( \Pr(A \cup B) \leq \Pr(A) + \Pr(B) \) and the second 
uses the Markov inequality. It follows that \( (\max_{1 \leq i \leq n} f_i) n^{-1/2} = o_p(1) \) and thus 
\( (\max_{1 \leq i \leq n} f_i) = o_p(n^{1/2}) \) by \( \xi > 2 \). Thus (2.4) implies M(i). \( \square \)

**Proof of Lemma 1.** ID holds trivially. By (2.2) and (2.3), \( g_i(\theta) = (y_i - Y_i(\theta))Z_i = 
Z_i(Z_i'\Pi + V_i')\theta_0 - \theta + Z_iu_i \). Next \( \max_{1 \leq i \leq n} \sup_{\theta \in \Theta} ||g_i(\theta)|| = o_p(n^{1/2}) \) 
is established. 
An application of the Borel-Cantelli Lemma shows that for real-valued i.i.d. random 
variables \( W_i \) such that \( EW_i^2 < \infty, max_{1 \leq i \leq n} |W_i| = o(n^{1/2}) \); see Owen (1990, Lemma 
3) for a proof. By the definition of \( g_i(\theta) \) and the triangle inequality,

\[
\max_{1 \leq i \leq n} \sup_{\theta \in \Theta} ||g_i(\theta)|| \leq \max_{1 \leq i \leq n} \sup_{\theta \in \Theta} (||Z_iZ_i'\Pi(\theta_0 - \theta)|| + ||Z_iV_i'(\theta_0 - \theta)|| + ||Z_iu_i||).
\]

By Assumption M'(iii), we can apply the just-mentioned result to each of the three 
summands in the above equation which proves the result.
Next M(ii) is shown. By the i.i.d. assumption, \( \Omega(\theta) = \lim_{n \to \infty} E g_i(\theta) g_i(\theta)' \) and continuity and boundedness in M(ii) follow immediately from M'(iii) and compactness of \( \Theta \). The same is true for the \( O_p(1) \) statement in M(ii). Finally, uniform convergence follows from the WLLN and compactness of \( \Theta \).

Next M(iii) is proven. Because \( \sup_{\theta \in \Theta} || n^{-1} \sum_{i=1}^{n} (Z_i Z_i' - Q_{ZZ}) C_A(\alpha_0 - \alpha) || \to_p 0 \), we only have to deal with the empirical process

\[
\Psi_n(\cdot, \theta) := n^{-1/2} \sum_{i=1}^{n} [Z_i (Z_i' \Pi_B (\beta_0 - \beta) + V_i' (\theta_0 - \theta) + u_i) - Q_{ZZ} \Pi_B (\beta_0 - \beta)].
\]

Finite dimensional joint convergence follows from the CLT and M'(iii) and stochastic equicontinuity follows from the fact that \( (\theta_0 - \theta) \) enters \( \Psi_n(\cdot, \theta) \) linearly:

\[
\sup_{||\theta_1 - \theta_2|| < \delta} || \Psi_n(\cdot, \theta_1) - \Psi_n(\cdot, \theta_2) || = \sup_{||\theta_1 - \theta_2|| < \delta} ||(\beta_2 - \beta_1)' n^{-1/2} \sum_{i=1}^{n} \Pi_B (Z_i Z_i' - Q_{ZZ}) + (\theta_2 - \theta_1)' n^{-1/2} \sum_{i=1}^{n} V_i Z_i' ||,
\]

where the last expression is bounded by \( \delta O_p(1) \) by the CLT. Furthermore, \( \Theta \) is compact by assumption. The proposition in Andrews (1994, p.2251) can thus be applied which yields the desired result. □

The following proofs are straightforward generalizations of Guggenberger’s (2003) proofs for the i.i.d. linear model to the more general context considered here. We require three lemmas that are modified versions of Lemmas A1-A3 in NS for the proofs of our theorems. These modifications are necessary because unlike NS we need to work with weakly and strongly identified parameters and do not make an i.i.d. assumption.

For \( n \in \mathbb{N} \) let \( \Theta_n \subset \Theta \). Let \( c_n := n^{-1/2} \max_{1 \leq i \leq n} \sup_{\theta \in \Theta_n} ||g_i(\theta)|| \). Let \( \Lambda_n := \{ \lambda \in \mathbb{R}^k : ||\lambda|| \leq n^{-1/2} c_n^{-1/2} \} \) if \( c_n > 0 \) and \( \Lambda_n = \mathbb{R}^k \) otherwise. Write “u.w.p.a.1” for “uniformly over \( \theta \in \Theta_n \) w.p.a.1”.

**Lemma 7** Assume \( \max_{1 \leq i \leq n} \sup_{\theta \in \Theta_n} ||g_i(\theta)|| = o_p(n^{1/2}) \).

Then \( \sup_{\theta \in \Theta_n, \lambda \in \Lambda_n, 1 \leq i \leq n} |\lambda' g_i(\theta)| \to_p 0 \) and \( \Lambda_n \subset \hat{\Lambda}_n(\theta) \) u.w.p.a.1, where \( \hat{\Lambda}_n(\theta) \) is defined in (2.5).

**Proof:** The case \( c_n = 0 \) is trivial and thus wlog \( c_n \neq 0 \) can be assumed. By assumption \( c_n = o_p(1) \) and the first part of the statement follows from

\[
\sup_{\theta \in \Theta_n, \lambda \in \Lambda_n, 1 \leq i \leq n} |\lambda' g_i(\theta)| \leq n^{-1/2} c_n^{-1/2} \max_{1 \leq i \leq n} \sup_{\theta \in \Theta_n} ||g_i(\theta)|| = n^{-1/2} c_n^{-1/2} n^{1/2} c_n = c_n^{1/2} = o_p(1),
\]

which also immediately implies the second part. □

**Lemma 8** Suppose \( \max_{1 \leq i \leq n} \sup_{\theta \in \Theta_n} ||g_i(\theta)|| = o_p(n^{1/2}) \), \( \lambda_{\min}(\hat{\Omega}(\theta)) \geq \varepsilon \) u.w.p.a.1 for some \( \varepsilon > 0 \), \( \hat{g}(\theta) = O_p(n^{-1/2}) \) uniformly over \( \theta \in \Theta_n \) and Assumption \( \rho \) holds.

[27]
Then \( \lambda(\theta) \in \hat{\Lambda}_n(\theta) \) satisfying \( \hat{P}(\theta, \lambda(\theta)) = \sup_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta, \lambda) \) exists u.w.p.a.1, \( \lambda(\theta) = O_p(n^{-1/2}) \) and \( \sup_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta, \lambda) = O_p(n^{-1}) \) uniformly over \( \theta \in \Theta_n \).

**Proof:** Wlog \( c_n \neq 0 \) and thus \( \Lambda_n \) can be assumed compact. For \( \theta \in \Theta_n \), let \( \lambda_\theta \in \Lambda_n \) be such that \( \hat{P}(\theta, \lambda_\theta) = \max_{\lambda \in \Lambda_n} \hat{P}(\theta, \lambda) \). Such a \( \lambda_\theta \in \Lambda_n \) exists u.w.p.a.1 because a continuous function takes on its maximum on a compact set and by Lemma 7 and Assumption \( \rho \), \( \hat{P}(\theta, \lambda) \) (as a function in \( \lambda \) for fixed \( \theta \)) is \( C^2 \) on some open neighborhood of \( \Lambda_n \) u.w.p.a.1. We now show that actually \( \hat{P}(\theta, \lambda_\theta) = \sup_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta, \lambda) \) u.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 \( \lambda_\theta^* \) on the line segment joining \( 0 \) and \( \lambda_\theta \) such that for some positive constants \( C_1 \) and \( C_2 \)

\[
0 = \hat{P}(\theta, 0) \leq \hat{P}(\theta, \lambda_\theta) = -2\lambda_\theta^* \hat{g}(\theta) + \lambda_\theta^* \left( \sum_{i=1}^n \rho_2(\lambda_\theta^* g_i(\theta))g_i(\theta)g_i(\theta)' / n \right) \lambda_\theta
\]

\[
\leq -2\lambda_\theta^* \hat{g}(\theta) - C_1 \lambda_\theta^* \hat{\Omega}(\theta) \lambda_\theta \leq 2||\lambda_\theta|| ||\hat{g}(\theta)|| - C_2 ||\lambda_\theta||^2
\]

u.w.p.a.1, where the second inequality follows as \( \max_{1 \leq i \leq n} \rho_2(\lambda_\theta^* g_i(\theta)) < -1/2 \) u.w.p.a.1 from Lemma 7, continuity of \( \rho_2(\cdot) \) at zero and \( \rho_2 = -1 \). The last inequality follows from \( \lambda_{\min}(\hat{\Omega}(\theta)) \geq \varepsilon > 0 \) u.w.p.a.1. Now, (A.1) implies that \( (C_2/2)||\lambda_\theta|| \leq ||\hat{g}(\theta)|| \) u.w.p.a.1, the latter being \( O_p(n^{-1/2}) \) uniformly over \( \theta \in \Theta_n \) by assumption. It follows that \( \lambda_\theta \in \text{int}(\Lambda_n) \) u.w.p.a.1. To prove this, let \( \epsilon > 0 \). Because \( \lambda_\theta = O_p(n^{-1/2}) \) uniformly over \( \theta \in \Theta_n \) and \( c_n = o_p(1) \), there exists \( M_\epsilon < \infty \) and \( n_\epsilon \in \mathbb{N} \) such that \( \Pr(||n^{1/2}\lambda_\theta|| < M_\epsilon) > 1 - \epsilon/2 \) uniformly over \( \theta \in \Theta_n \) and \( \Pr(c_n^{-1/2} > M_\epsilon) > 1 - \epsilon/2 \) for all \( n \geq n_\epsilon \). Then \( \Pr(\lambda_\theta \in \text{int}(\Lambda_n)) = \Pr(||n^{1/2}\lambda_\theta|| < c_n^{-1/2}) \geq \Pr(||n^{1/2}\lambda_\theta|| \leq M_\epsilon) \wedge (c_n^{-1/2} > M_\epsilon)) > 1 - \epsilon \) for \( n \geq n_\epsilon \) uniformly over \( \theta \in \Theta_n \).

Hence, the FOC for an interior maximum \( (\partial \hat{P}/\partial \lambda)(\theta, \lambda) = 0 \) hold at \( \lambda = \lambda_\theta \) u.w.p.a.1. By Lemma 7, \( \lambda_\theta \in \hat{\Lambda}_n(\theta) \) u.w.p.a.1 and thus by concavity of \( \hat{P}(\theta, \lambda) \) (as a function in \( \lambda \) for fixed \( \theta \)) and convexity of \( \hat{\Lambda}_n(\theta) \) it follows that \( \hat{P}(\theta, \lambda_\theta) = \sup_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta, \lambda) \) u.w.p.a.1 which implies the first part of the lemma. From above \( \lambda_\theta = O_p(n^{-1/2}) \) uniformly over \( \theta \in \Theta_n \). Thus the second and by (A.1) the third parts of the lemma follow. \( \square \)

Suppose \( \Theta_1 \times \Theta_2 \subset \Theta, \Theta_i \subset \mathbb{R}^{p_i}, p_1 + p_2 = p \). Partition \( \theta_0 = (\theta_{01}', \theta_{02})' \) accordingly and assume \( \theta_{02} \in \Theta_2 \). For \( d_1 \in \Theta_1 \) define

\[
\hat{\theta}_2(d_1) := \arg \min_{d_2 \in \Theta_2} \sup_{\lambda \in \hat{\Lambda}_n((d_1', d_2)')} \hat{P}((d_1', d_2)', \lambda) \in \mathbb{R}^{p_2},
\]

\[
\hat{\theta}_1(d_1) := (d_1', \hat{\theta}_2(d_1)')' \in \mathbb{R}^p, \ \theta_{d_1} := (d_1', \theta_{02})' \in \mathbb{R}^p.
\]

By “u.w.p.a.1” we denote “uniformly over \( d_1 \in \Theta_1 \) w.p.a.1”.

**Lemma 9** Suppose \( \max_{1 \leq i \leq n} \sup_{\theta \in \Theta_1 \times \Theta_2} ||g_i(\theta)|| = o_p(n^{1/2}) \), \( \lambda_{\max}(\hat{\Omega}(\theta_{d_1})) \leq \kappa \) u.w.p.a.1 for some \( \kappa < \infty \), \( \sup_{\lambda \in \hat{\Lambda}_n(\theta_{d_1})} \hat{P}(\theta_{d_1}, \lambda) = O_p(n^{-1}) \) uniformly over \( d_1 \in \Theta_1 \) and Assumption \( \rho \) holds.

Then \( \hat{g}(\theta_{d_1}) = O_p(n^{-1/2}) \) uniformly over \( d_1 \in \Theta_1 \).
Proof: Wlog $\hat{g}(\theta_d) \neq 0$ can be assumed. Define $\Delta := -n^{-1/2}\hat{g}(\theta_d) / ||\hat{g}(\theta_d)||$. Note that $\lambda \in \Lambda_n$ and thus $\lambda \in \hat{\Lambda}(\theta)$ uniformly over $\theta \in \Theta_n$ w.p.a.1 (see Lemma 7 with $\Theta_n := \Theta_1 \times \Theta_2$). By a second order Taylor expansion around $\lambda = 0$, there is a $\lambda$ on the line segment joining 0 and $\mu$ such that for some positive constants $C_1$ and $C_2$

$$\hat{P}(\theta_d, \lambda) = -2\lambda'\hat{g}(\theta_d) + \lambda'[\sum_{i=1}^{n} \rho_2(\lambda' g_i(\theta_d)) g_i(\theta_d) g_i(\theta_d)' / n] \lambda$$

$$\geq 2n^{-1/2}||\hat{g}(\theta_d)|| - C_1 \lambda'[\sum_{i=1}^{n} g_i(\theta_d) g_i(\theta_d)' / n] \lambda$$

$$\geq 2n^{-1/2}||\hat{g}(\theta_d)|| - C_2 n^{-1} \tag{A.2}$$

c.f. w.p.a.1, where the first inequality follows from Lemma 7 which implies that $\min_{i=1,...,n} \rho_2(\lambda' g_i(\theta_d)) \geq -1.5$ u.w.p.a.1. The second inequality follows by $\lambda_{\max}(\hat{\Omega}(\theta_d)) \leq \kappa < \infty$ u.w.p.a.1. The definition of $\hat{\theta}_d$ implies

$$\hat{P}(\theta_d, \lambda) \leq \sup_{\lambda \in \hat{\Lambda}_n(\theta_d)} \hat{P}(\theta_d, \lambda) \leq \sup_{\lambda \in \Lambda_n(\theta_d)} \hat{P}(\theta_d, \lambda) = O_p(n^{-1}) \tag{A.3}$$

uniformly over $\theta_d \in \Theta_1$. Combining equations (A.2) and (A.3) implies $n^{-1/2}||\hat{g}(\theta_d)|| = O_p(n^{-1})$ uniformly over $\theta_d \in \Theta_1$. \hfill \Box

Proof of Theorem 2: (i) We first show consistency of $\hat{\beta}$. By Assumption ID and M(iii) $\sup_{\theta \in \Theta} ||\hat{g}(\theta) - (n^{-1/2} m_1(\theta) + m_2(\beta))|| \to_p 0$, where $m_2(\beta) = 0$ if and only if $\beta = \beta_0$. Therefore, $\hat{g}(\theta) = o_p(1)$ is a sufficient condition for consistency of $\hat{\beta}$. Applying Lemma 8 to the case $\Theta_n = \{\theta_0\}$, gives $\sup_{\lambda \in \hat{\Lambda}_n(\theta_0)} \hat{P}(\theta_0, \lambda) = O_p(n^{-1})$.

Assumption M(ii) implies $\lambda_{\max}(\hat{\Omega}(\beta)) \leq \kappa$ w.p.a.1 for some $\kappa < \infty$ and thus Lemma 9 (applied to the case $p_1 = 0$, $\Theta_2 = \Theta$) implies $\hat{g}(\theta) = O_p(n^{-1/2})$.

Next we establish $n^{1/2}$-consistency of $\hat{\beta}$. By consistency of $\hat{\beta}$ and Assumption M(ii) $\lambda_{\max}(\hat{\Omega}(\hat{\beta})) \geq \varepsilon$ w.p.a.1 for some $\varepsilon > 0$ and thus Lemma 8 for the case $\Theta_n = \{\hat{\beta}\}$ implies that the FOC

$$n^{-1} \sum_{i=1}^{n} \rho_1(\lambda' g_i(\theta)) g_i(\theta) = 0 \tag{A.4}$$

have to hold at $(\hat{\theta}, \hat{\lambda})$ w.p.a.1, where $\hat{\lambda} := \lambda(\hat{\theta}) = O_p(n^{-1/2})$ and $\lambda(\theta)$, for given $\theta \in \Theta$, is defined in Lemma 8. Expanding the FOC in $\lambda$ around 0, there exists a mean value $\bar{\lambda}$ between 0 and $\hat{\lambda}$ (that may be different for each row) such that

$$0 = -\hat{g}(\theta) + [\sum_{i=1}^{n} \rho_2(\hat{\lambda}' g_i(\theta)) g_i(\theta) g_i(\theta)' / n] \bar{\lambda} = -\hat{g}(\theta) - \bar{\Omega}_{\hat{\lambda}} \bar{\lambda},$$

where the matrix $\bar{\Omega}_{\hat{\lambda}}$ has been implicitly defined. Because $\bar{\lambda} = O_p(n^{-1/2})$, Lemma 7 and Assumption $\rho$ imply that $\sup_{i=1,...,n, \theta \in \Theta} |\rho_2(\hat{\lambda}' g_i(\theta)) + 1| \to_p 0$. By Assumption M(ii), it follows that $\bar{\Omega}_{\hat{\lambda}} \to_p \Omega((\hat{\alpha}', \beta_0'))$ and thus $\bar{\Omega}_{\hat{\lambda}}$ is invertible w.p.a.1 and $(\bar{\Omega}_{\hat{\lambda}})^{-1} \to_p \Omega((\hat{\alpha}', \beta_0'))^{-1}$. Therefore

$$\hat{\lambda} = -(\bar{\Omega}_{\hat{\lambda}})^{-1} \hat{g}(\theta) \tag{A.5}$$

[29]
w.p.a.1. Inserting this into a second order Taylor expansion for \( \hat{P}(\theta, \lambda) \) (with mean value \( \lambda' \) as in (A.1) above) it follows that

\[
\hat{P}(\hat{\theta}, \lambda) = 2\hat{g}(\hat{\theta})'\hat{\Omega}^{-1}_{\lambda \hat{\theta}}\hat{g}(\hat{\theta}) - \hat{g}(\hat{\theta})'\hat{\Omega}^{-1}_{\lambda \hat{\theta}}\hat{\theta}'\lambda\hat{g}(\hat{\theta}). \tag{A.6}
\]

The same argument as for \( \hat{\Omega}_{\lambda \hat{\theta}} \) proves \( \hat{\Omega}_{\lambda \hat{\theta}} \rightarrow_p \Omega((\hat{\alpha}', \hat{\beta}_0')') \). We therefore have \( \hat{P}(\hat{\theta}, \lambda) = \hat{g}(\hat{\theta})'\Omega((\hat{\alpha}', \hat{\beta}_0')')^{-1} + o_p(1)\hat{g}(\hat{\theta}) \). By the definition of \( \lambda' \),

\[
n\lambda \hat{\theta} - n\hat{\theta} = n^{1/2}\hat{g}(\hat{\theta})'(\Omega((\hat{\alpha}', \hat{\beta}_0')')^{-1} + o_p(1))n^{1/2}\hat{g}(\hat{\theta}) - n^{1/2}\hat{g}(\hat{\theta})'(\Omega(\theta_0)^{-1} + o_p(1))n^{1/2}\hat{g}(\hat{\theta}) \leq 0.
\]

By Assumption ID, we have up to \( o_p(1) \) terms that \( n^{1/2}\hat{g}(\hat{\theta}) = \Psi_n(\hat{\theta}) + m_{1n}(\hat{\theta}) + n^{1/2}m_2(\hat{\beta}) \) and \( n^{1/2}\hat{g}(\theta_0) = \Psi_n(\theta_0) \). The same analysis as in the proof of Lemma A1 in SW (p.1091, line six from the top) can now be applied to prove \( n^{1/2} \)-consistency of \( \hat{\beta} \), where the symmetric matrix \( \Omega((\hat{\alpha}', \hat{\beta}_0')')^{-1} + o_p(1) \) plays the role of \( W_T(\hat{\theta}_T) \) in SW. Note that in eq. (A.4) in SW, the assumption M(iii) of bounded sample paths w.p.a.1 is used. Finally, note that \( \lambda_{\min}(\Omega((\hat{\alpha}', \hat{\beta}_0')')^{-1} + o_p(1)) \) is bounded away from zero w.p.a.1.

(ii) By Assumption M(iii)

\[
n^{1/2} \sup_{(a,b) \in A \times B_M} ||\hat{g}(\theta_{ab}) - E\hat{g}(\theta_{ab})|| = O_p(1)
\]

and by ID we have for some mean-vector \( \overline{\beta} \) between \( \beta_0 \) and \( \beta_0 + n^{-1/2}b \) (that may differ across rows)

\[
n^{1/2}E\hat{g}(\theta_{ab}) = m_{1n}(\theta_{ab}) + n^{1/2}m_2(\beta_0 + n^{-1/2}b) = m_{1n}(\theta_{ab}) + M_2(\overline{\beta})b.
\]

Because the latter expression is bounded, it follows that \( \hat{g}(\theta_{ab}) = O_p(n^{-1/2}) \) w.p.a.1, where “u.w.p.a.1” stands for “uniformly over \( (a,b) \in A \times B_M \) w.p.a.1”. Therefore, by Lemma 8, \( \lambda(\theta_{ab}) \) such that \( \hat{P}(\theta_{ab}, \lambda(\theta_{ab})) = \sup_{\lambda \in \Lambda_u(\theta_{ab})} \hat{P}(\theta_{ab}, \lambda) \) exists u.w.p.a.1 and \( \lambda(\theta_{ab}) = O_p(n^{-1/2}) \) uniformly over \( (a,b) \in A \times B_M \). This means that the FOC

\[
n^{-1} \sum_{i=1}^n \rho_1(\lambda g_i(\theta))g_i(\theta) = 0
\]

have to hold at \( \lambda = \lambda(\theta_{ab}) \) and \( \theta = \theta_{ab} \) u.w.p.a.1. Expanding the FOC and using the same steps and notation as in part (i), it follows that \( \lambda(\theta_{ab}) = - (\hat{\Omega}_{\lambda \theta_{ab}})^{-1}\hat{g}(\theta_{ab}) \) and upon inserting this into a second order Taylor expansion of \( \hat{P}(\theta, \lambda) \) we have

\[
\hat{P}(\theta_{ab}, \lambda(\theta_{ab})) = 2\hat{g}(\theta_{ab})'\hat{\Omega}^{-1}_{\lambda \theta_{ab}}\hat{g}(\theta_{ab}) - \hat{g}(\theta_{ab})'\hat{\Omega}^{-1}_{\lambda \theta_{ab}}\hat{\theta}'\lambda\hat{g}(\theta_{ab})
\]

u.w.p.a.1. The matrices \( \hat{\Omega}_{\lambda \theta_{ab}} \) and \( \hat{\Omega}^*_{\lambda \theta_{ab}} \) converge to \( \Omega((\alpha', \beta_0')') \) uniformly over \( A \times B_M \). By M(iii), \( n^{1/2}\hat{g}(\theta_{ab}) \Rightarrow \Psi((\alpha', \beta_0')') + m_1((\alpha', \beta_0')') + M_2(\beta_0) \) and therefore that \( n\hat{P}(\theta_{ab}, \lambda(\theta_{ab})) \Rightarrow P_{ab} = P((\alpha', \beta_0')', b) \) on \( A \times B_M \).

By part (i) of the proof and Lemma 3.2.1 in van der Vaart and Wellner (1996, p.286) it follows that

\[
(\hat{\alpha}', n^{1/2}(\hat{\beta} - \beta_0')') \rightarrow_d (\alpha'', b'').
\]
For given $\alpha \in A$, one can calculate $\arg\min_{b \in R^B} P_{\alpha b}$ by solving the FOC for $b$. Writing $\Omega$ for $\Omega((\alpha', \beta_0'))$ and $M_2$ for $M_2(\beta_0)$ the result is
\[
b^*(\alpha) = -(M_2^* \Omega^{-1} M_2)^{-1} M_2^* \Omega^{-1} \Psi((\alpha', \beta_0') + m_1((\alpha', \beta_0'))]. \tag{A.7}
\]
This holds in particular for $\alpha = \alpha^*$. It follows that $\alpha^* = \arg\min_{\alpha \in A} P_{\alpha \beta^*}$. $\square$

**Proof of Theorem 3:** Applying Lemma 8 to the case $\Theta_n = \{\theta\}$, it follows that $\lambda(\theta) \in \tilde{\Lambda}_n(\theta)$ exists, such that $\hat{P}(\theta, \lambda(\theta)) = \sup_{\lambda \in \tilde{\Lambda}_n(\theta)} \hat{P}(\theta, \lambda)$. Using the same steps and notation as in the proof of Theorem 2 leads to
\[
\hat{P}(\theta, \lambda(\theta)) = 2\hat{g}(\theta)\hat{\Omega}^{-1}\hat{g}(\theta) - \hat{g}(\theta)\hat{\Omega}^{-1}\hat{\Omega}_{\lambda^*\theta}^*\hat{\Omega}^{-1}\hat{g}(\theta)
\]
w.p.a.1, where by $M_\theta$ both $\hat{\Omega}_{\lambda^\theta}$ and $\hat{\Omega}_{\lambda^*\theta}$ converge in probability to $\Delta(\theta)$. By $M_\theta$, we have
\[
n^{1/2}\hat{g}(\theta) \rightarrow_d N(m_1(\theta), \Delta(\theta)),
\]
from which the result follows. $\square$

**Proof of Theorem 4:** Using $M_\theta(i)$-iii and a similar argument that led to (A.5) we have
\[
n^{1/2}\lambda(\theta) = -\Delta(\theta)^{-1}n^{1/2}\hat{g}(\theta) + o_p(1) \tag{A.8}
\]
and therefore the statement of the theorem involving $S_\rho(\theta)$ follows immediately from the one for $LM_\rho(\theta)$. Therefore, we only deal with the statistic $LM_\rho(\theta)$ given in eq. (3.8).

First, we show that the matrix $D^*$ is asymptotically independent of $n^{1/2}\hat{g}(\theta)$. For notational convenience from now on we omit the argument $\theta$, e.g., we write $g_i$ for $g_i(\theta)$. By a mean value expansion about 0 we have $\rho_1(\lambda' g_i) = -1 + \rho_2(\xi_i) g_i' \lambda$ for a mean value $\xi_i$ between 0 and $\lambda' g_i$ and thus by (A.8) and the definition of $\Lambda$ we have
\[
D^* = -n^{-1} \sum_{i=1}^n (n^{1/2} G_{iA}, G_{iB}) - n^{-3/2} \sum_{i=1}^n [\rho_2(\xi_i)(n^{1/2} G_{iA}, G_{iB}) g_i' \Delta^{-1} n^{1/2} \hat{g}] + o_p(1)
\]
\[
= -(n^{-1/2} \sum_{i=1}^n G_{iA} - n^{-1} \sum_{i=1}^n G_{iA} g_i' \Delta^{-1} n^{1/2} \hat{g}, M_2(\beta_0)) + o_p(1),
\]
where for the last equality we use (3.7) and Assumption $M_\theta(v)$-vi. By Assumption $M_\theta(v)$ it thus follows that
\[
vec(D^*, n^{1/2} \hat{g}) = w_1 + M v + o_p(1),
\]
where $w_1 := vec(0, -M_2(\beta_0), 0) \in R^{k_{PA} + k_{PB} + k}$ and
\[
M := \begin{pmatrix} -I_{k_{PA}} & \Delta_A \Delta^{-1} \\ 0 & 0 \\ 0 & I_k \end{pmatrix}, \quad v := n^{-1/2} \sum_{i=1}^n \begin{pmatrix} vec G_{iA} \\ g_i \end{pmatrix};
\]

[31]
$M$ and $v$ have dimensions $(kpA + kpB + k) \times (kpA + k)$ and $(kpA + k) \times 1$, respectively. By Assumption ID, $M\alpha(vii)$ and (3.7) $v \rightarrow N(w_2, V(\theta))$, where $w_2 := ((vecM_1A)\cdot, m_1')'$ and $M_{1A}$ are the first $p_A$ columns of $M_1$. Therefore

$$vec(D^*, n^{1/2}\mathbf{g}) \rightarrow_d N(w_1 + Mw_2, \begin{pmatrix} \Psi & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Delta \end{pmatrix}),$$  \hspace{1cm} (A.9)

where $\Psi := \Delta_{AA} - \Delta_A\Delta^{-1}\Delta'_A$ is positive definite. Equation (A.9) proves that $D^*$ and $n^{1/2}\mathbf{g}$ are asymptotically independent.

We now derive the asymptotic distribution of $LM_\theta(\theta)$. Denote by $\mathcal{D}$ and $\mathbf{g}$ the limiting normal distributions of $D^*$ and $n^{1/2}\mathbf{g}$, respectively, see (A.9). Below we show that the function $h : R^{k \times p} \rightarrow R^{p \times k}$ defined by $h(D) := (D'\Delta^{-1}D)^{-1/2}D'$ for $D \in R^{k \times p}$ is continuous on a set $C \subset R^{k \times p}$ with $\Pr(\mathcal{D} \in C) = 1$. By the continuous mapping theorem and $M_\theta(v)$ we have

$$(D^{\ast'}\hat{\Omega}^{-1}D^{\ast})^{-1/2}D^{\ast'}\hat{\Omega}^{-1}n^{1/2}\mathbf{g} \rightarrow_d (\mathcal{D}\Delta^{-1}\mathcal{D})^{-1/2}\mathcal{D}\Delta^{-1}\mathbf{g}.$$  \hspace{1cm} (A.10)

By the independence of $\mathcal{D}$ and $\mathbf{g}$, the latter random variable is distributed as $W + \zeta$, where the random $p$-vector $W$ is defined as

$$W = W(\alpha) := (\mathcal{D}\Delta^{-1}\mathcal{D})^{-1/2}\mathcal{D}\Delta^{-1}m_1(\theta),$$  \hspace{1cm} (A.11)

$\zeta \sim N(0, I_p)$ and $W$ and $\zeta$ are independent. Note that for $\theta = \theta_0$, $W \equiv 0$. From (A.10) the statement of the theorem follows.

We now prove the continuity claim for $h$. Note that $h$ is continuous at each $D$ that has full column rank. It is therefore sufficient to show that $\mathcal{D}$ has full column rank a.s.. From (A.9) it follows that the last $p_B$ columns of $\mathcal{D}$ equal $-M_2(\beta_0)$ which has full column rank by assumption. Define $O := \{ o \in R^{kpA} : \exists \beta \in R^{kpA}, \text{s.t. } o = vec(\beta) \}$ and the $k \times p$-matrix $(\beta, -M_2(\beta_0))$ 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 $\mathcal{D}$, $\mathcal{D}_{p_A}$ say, it has been shown that $vec(\mathcal{D}_{p_A})$ is normally distributed with full rank covariance matrix $\Psi$. This implies that for any measurable set $O^* \subset R^{kpA}$ with Lebesgue-measure 0, it holds that $\Pr(\text{vec}(\mathcal{D}_{p_A}) \in O^*) = 0$, in particular, for $O^* = O$. This proves the continuity claim for $h$. $\Box$

**Proof of Theorem 5.** By Assumption $M_\alpha(v)$ and $ID_\alpha \hat{g}(\hat{\theta}_a) = m_2((\alpha_0, \hat{\beta}) + o_p(1)$ and by Lemmas $8$ and $9$ (applied to $\Theta_n = \{\theta_{a_0}\}$ and $\Theta_1 = \{a\}, \Theta_2 = B$, respectively) we have $\hat{g}(\hat{\theta}_a) = O_p(n^{-1/2})$. Assumption $ID_\alpha$ then implies consistency of $\hat{\beta}$. Applying Lemma $8$ to the case $\Theta_n = \{\hat{\theta}_a\}$ implies that the FOC for $\lambda$ must hold in the definition of $\hat{\theta}_a$, see (A.4) above. Then repeating the analysis that leads to (A.6) in the proof of Theorem 2, we have by $M_\alpha(ii)$

$$GELR_{p}^{sub}(a) = n^{1/2}\hat{g}(\hat{\theta}_a)^{'}\Delta(\theta_{a_0})^{-1}n^{1/2}\hat{g}(\hat{\theta}_a) + o_p(1).$$  \hspace{1cm} (A.12)

The next goal is to derive the asymptotic distribution of $n^{1/2}\hat{g}(\hat{\theta}_a)$. Our analysis follows NS, see their proof of Theorem 3.2. Differentiating the FOC (A.4) with
respect to \( \lambda \) yields the matrix \( n^{-1} \sum_{i=1}^{n} \rho_{2}(\lambda' g_{i}(\hat{\theta}_{a})) g_{i}(\hat{\theta}_{a})g_{i}(\hat{\theta}_{a})' \) which by \( M_{a}(ii) \) converges in probability to \(-\Delta(\theta_{a\beta_{0}})\) which is nonsingular. Therefore, the implicit function theorem implies that there is a neighborhood of \( \hat{\theta}_{a} \) where the solution to the FOC, say \( \hat{\lambda}(\theta) \), is continuously differentiable w.p.a.1. The envelope theorem then implies
\[
n^{-1} \sum_{i=1}^{n} \rho_{1}(\lambda' g_{i}(\hat{\theta}_{a})) (\partial g_{i}(\hat{\theta}_{a})/\partial \beta)'(\hat{\theta}_{a}) \hat{\lambda} = 0
\] (A.13)

w.p.a.1. Also, a mean-value expansion of (A.4) in \((\beta, \lambda)\) about \((\beta_{0}, 0)\) yields (where \(g_{i}(\hat{\theta})\) inside \(\rho_{1}\) is kept constant at \(g_{i}(\hat{\theta}_{a})\))
\[
-\bar{g}(\theta_{a\beta_{0}}) + n^{-1} \sum_{i=1}^{n} [\rho_{1}(\lambda' g_{i}(\hat{\theta}_{a}))G_{i}(\theta_{a\beta_{0}})(\hat{\beta} - \beta_{0}) + \rho_{2}(\lambda' g_{i}(\hat{\theta}_{a}))g_{i}(\theta_{a\beta_{0}})g_{i}(\hat{\theta}_{a})'\hat{\lambda}] = 0,
\] (A.14)

where \((\bar{\beta}, \bar{\lambda})\) are mean-values on the line segment that joins \((\beta_{0}', 0')\) and \((\bar{\beta}', \bar{\lambda}')\) that may be different for each row. Combining the \(p_{B}\) rows of (A.13) with the \(k\) rows of (A.14) we get
\[
\begin{pmatrix}
0 \\
-\bar{g}(\theta_{a\beta_{0}})
\end{pmatrix}
+ M \begin{pmatrix}
\hat{\beta} - \beta_{0} \\
\hat{\lambda}
\end{pmatrix} = 0,
\] (A.15)

where the \((p_{B} + k) \times (p_{B} + k)\) matrix \(M\) has been implicitly defined. By \(M_{a}(ii)\) and \(M_{a}(iv)-(vi)\) the matrix \(M\) converges in probability to \(\bar{M}\), where (writing \(M_{2\beta}\) for \(M_{2\beta}(\theta_{a\beta_{0}})\))
\[
\bar{M} := - \begin{pmatrix}
0 & M'_{2\beta} \\
M_{2\beta} & \Delta(\theta_{a\beta_{0}})
\end{pmatrix}
\] and \(\bar{M}^{-1} = - \begin{pmatrix}
-\Sigma & H' \\
H & P
\end{pmatrix}\)

and (omitting the argument \(\theta_{a\beta_{0}}\))
\[
\Sigma := (M'_{2\beta}\Delta^{-1}M_{2\beta})^{-1}, \quad H := \Sigma M'_{2\beta}\Delta^{-1} \quad \text{and} \quad P := \Delta^{-1} - \Delta^{-1}M_{2\beta}\Sigma M'_{2\beta}\Delta^{-1}.
\]

It follows that \(M\) is nonsingular w.p.a.1. Equation (A.15) implies that w.p.a.1
\[
n^{1/2}((\hat{\beta} - \beta_{0})', \hat{\lambda}')' = M^{-1}(0', n^{1/2}\bar{g}(\theta_{a\beta_{0}})'')'
\] (A.16)

An expansion of \(\bar{g}(\theta_{a})\) in \(\beta\) around \(\beta_{0}\) and the above lead to
\[
\bar{g}(\theta_{a}) = \bar{g}(\theta_{a\beta_{0}}) + \bar{G}_{B}(\bar{\theta})(\hat{\beta} - \beta_{0}) = (I_{k} - M_{2\beta}H)\bar{g}(\theta_{a\beta_{0}}) + o_{P}(n^{-1/2})
\] (A.17)

for some appropriate mean-value \(\bar{\theta}\). Note that
\[
I_{k} - M_{2\beta}H = M_{M_{2\beta}}(\Delta(\theta_{a\beta_{0}}))
\] (A.18)

which has rank \(k - p_{B}\). From (A.12), \(GELR_{p}^{sub}(a) \rightarrow_{d} \xi' \Delta(\theta_{a\beta_{0}})^{-1}M_{2\beta}(\Delta(\theta_{a\beta_{0}}))\xi\), where \(\xi \sim N(m_{1}(\theta_{a\beta_{0}}), \Delta(\theta_{a\beta_{0}}))\), which concludes the proof. □

**Proof of Theorem 6.** As in the proof of Theorem 5, \(n^{1/2} \lambda(\hat{\theta}_{a}) = -\Delta(\theta_{a\beta_{0}})^{-1}n^{1/2}\bar{g}(\theta_{a}) + o_{P}(1)\). Hence, the result for \(LM_{p}^{sub}(a)\) implies the result for \(S_{p}^{sub}(a)\).
As in the proof of Theorem 4 renormalize $D^* := D_\rho(a)\Lambda$, where the diagonal $p_A \times p_A$ matrix $\Lambda := diagon(n^{1/2}, ..., n^{1/2}, 1, ..., 1)$ has first $p_A$ diagonal elements equal $n^{1/2}$ and the remaining $p_A$ elements equal to unity. We now show that $D^*$ and $n^{1/2}\tilde{g}(\theta_0)$ are asymptotically independent. By a mean value expansion about $\theta_{a\beta_0}$ and Assumption $M_\alpha(vii)$ we have for some mean value $\tilde{\theta} = (\alpha', \beta')'$ (that may be different for each row)

$$n^{1/2} \text{vec} \hat{G}_{A_1}(\tilde{\theta}_a) = n^{1/2} \text{vec} \hat{G}_{A_1}(\theta_{a\beta_0}) + (\text{vec} \hat{G}_{A_1}/\partial \beta)(\tilde{\theta}) n^{1/2}(\tilde{\beta} - \beta_0)$$

$$= n^{1/2} \text{vec} \hat{G}_{A_1}(\theta_{a\beta_0}) - (\text{vec} \hat{G}_{A_1}/\partial \beta)(\tilde{\theta}) H n^{1/2}\tilde{g}(\theta_{a\beta_0}) + o_p(1),$$

where we have used (A.16) for the last equation. $M_\alpha(vii)$ and $\text{ID}_\alpha$ imply $(\text{vec} \hat{G}_{A_1}/\partial \beta)(\tilde{\theta}) = \partial(n^{-1/2}m_1(\tilde{\theta}) + m_2((a_{02}, \beta_1)))/\partial \beta \partial \alpha_1 + o_p(1) \to_p 0$ (recall that $m_2$ does not depend on $\alpha_1$) and thus

$$n^{1/2} \text{vec} \hat{G}_{A_1}(\tilde{\theta}_a) = n^{1/2} \text{vec} \hat{G}_{A_1}(\theta_{a\beta_0}) + o_p(1). \quad (A.19)$$

Proceeding exactly as in the proof of Theorem 4, using (A.17), (A.19) and Assumptions $M_\alpha(vii)-(ix)$ it follows that

$$\text{vec}(D^*, n^{1/2}\tilde{g}(\theta_0)) = m + Mv + o_p(1), \quad (A.20)$$

where $M \in R^{(kp_A + kp_{A_2} + k) \times (kp_A + k)}$ and

$$M := \begin{pmatrix} -I_{kp_A} & \Delta_{A_1} \Delta^{-1} & 0 \\ 0 & 0 & I_{kp_A} \\ 0 & I_k & 0 \end{pmatrix}, \quad v := n^{-1/2} \sum_{i=1}^n \begin{pmatrix} \text{vec}G_{iA_1}(\theta_{a\beta_0}) \\ g_i(\theta_{a\beta_0}) \end{pmatrix}, \quad m := \text{vec}(0, -(\partial m_2/\partial \alpha_2), 0),$$

where the arguments $(a_{02}, \beta_1)$ in $M_{23}$ and $(\partial m_2/\partial \alpha_2)$ and $\theta_{a\beta_0}$ in $\Delta$ and $\Delta$ are omitted. By $M_\alpha(x)$, $v$ is asymptotically normal with full rank covariance matrix $V^\alpha(\theta_{a\beta_0})$ and thus the asymptotic covariance matrix of vec($D^*, n^{1/2}\tilde{g}(\theta_0)$) is given by $M V^\alpha(\theta_{a\beta_0}) M'$. For independence of $D^*$ and $n^{1/2}\tilde{g}(\theta_0)$ the upper right $k(p_{A_2} + p_{A_2}) \times k$-submatrix of $M V^\alpha(\theta_{a\beta_0}) M'$ must be 0. This is clear for the $kp_{A_2} \times k$-dimensional submatrix and we only have to show that the $kp_{A_1} \times k$ upper right submatrix

$$(-\Delta_{A_1} + \Delta_{A_1} \Delta^{-1}(I_k - M_{23}H)\Delta)(I_k - M_{23}H)' \quad (A.21)$$

is 0. Using (A.18), the matrix in (A.21) equals $-\Delta_{A_1} \Delta^{-1}P M_{23}(\Delta) M_{23}(\Delta)\Delta$ which is clearly 0. This proves the independence claim.

Now denote by $\overline{D}$ and $\overline{g}$ the limiting normal distributions of $D^*$ and $n^{1/2}\tilde{g}(\theta_0)$, implied by (A.20). Recall $M(a) = \Delta^{-1}M_{23}(\Delta)$, see (4.2). If the function $h : R^{k \times p_A} \to R^{p_{A_2} \times k}$ defined by $h(D) := (D'M(a)D)^{-1/2}D'$ for $D \in R^{k \times p_A}$ is continuous on a set $C \subset R^{k \times p_A}$ with $\text{Pr}(D \in C) = 1$, then by the continuous mapping theorem

$$(D'M(a)D)^{-1/2}D'M(a)D)^{-1/2}D' \Delta^{-1}g.$$

[34]
By (A.17) and (A.18) the latter variable is distributed as 
\[ W_\alpha(a) + \zeta_\alpha, \]
where
\[ W_\alpha(a) := (\mathcal{D}' M(a) \mathcal{D})^{-1/2} \mathcal{D}' M(a)m_1(\theta_{a\beta_0}). \] (A.22)

Therefore the theorem is proven once we have proved the continuity claim for \( h \). For this step of the proof we need the positive definite assumption for \( V^\alpha(\theta_{a\beta_0}) \) in \( M_\alpha(x) \).

It is enough to show that with probability 1, \( \text{rank}(M_{M_{2\beta}}(\Delta) \mathcal{D}) = p_A \). Because the span of the columns of \( M_{2\beta} \) equals the kernel of \( M_{M_{2\beta}}(\Delta) \) and \( \text{rank}(M_{2\beta}) = p_B \), the latter condition holds if \( \text{rank}(M_{2\beta}, \mathcal{D}) = p \). Denote by \( \mathcal{D}_{pA_2} \) the last \( p_{A_2} \) columns of \( \mathcal{D} \) which by (A.20) equal \( -(\partial m_2/\partial \alpha_2) \). By Assumption ID_\alpha, the matrix \( (\partial m_2/\partial (\alpha_2', \beta_0'))((\alpha_02, \beta_0)) \) has rank \( p_{A_2} + p_B \) and it remains to show that with probability one, this matrix is linearly independent of the first \( p_{A_1} \) columns of \( \mathcal{D} \), \( \mathcal{D}_{pA_1} \) say.

Using (A.20) and \( V^\alpha(\theta_{a\beta_0}) > 0 \), the covariance matrix of \( \text{vec}\mathcal{D}_{pA_1} \) is easily shown to have full column rank \( p_{A_1}k \). An argument analogous to the last step in the proof of Theorem 4 can then be applied to conclude the proof. \( \square \)
References


[36]


——— (2002b) Two independent pivotal statistics that test location and misspecification and add-up to the Anderson-Rubin statistic. Working paper.


Asterisks in each row denote the number closest to the 5% significance level. The size results are computed using R=10,000 simulation repetitions.
<table>
<thead>
<tr>
<th>$\Pi_1$</th>
<th>$n$</th>
<th>$k$</th>
<th>$\rho_{UV}$</th>
<th>$LM_{EL}$</th>
<th>$S_{EL}$</th>
<th>$LM_{CUE}$</th>
<th>$GELR_{CUE}$</th>
<th>$GELR_{ET}$</th>
<th>$K_{HET}$</th>
<th>$LR_M$</th>
<th>$2SLS_{HET}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>1</td>
<td>.0</td>
<td>3.9</td>
<td>15.9</td>
<td>3.9</td>
<td>8.1</td>
<td>4.9*</td>
<td>27.7</td>
<td>7.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>4.2</td>
<td>16.8</td>
<td>4.2</td>
<td>8.4</td>
<td>5.0*</td>
<td>26.1</td>
<td>8.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.99</td>
<td>3.9</td>
<td>16.3</td>
<td>3.9</td>
<td>7.8</td>
<td>4.8*</td>
<td>25.7</td>
<td>8.1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>.0</td>
<td>4.3</td>
<td>20.2</td>
<td>2.8</td>
<td>2.2</td>
<td>16.1</td>
<td>5.1*</td>
<td>13.1</td>
<td>6.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>4.0</td>
<td>19.9</td>
<td>2.6</td>
<td>15.9</td>
<td>4.6*</td>
<td>11.3</td>
<td>7.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.99</td>
<td>4.0</td>
<td>19.7</td>
<td>2.3</td>
<td>15.3</td>
<td>4.7*</td>
<td>10.8</td>
<td>14.4</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>.0</td>
<td>4.5*</td>
<td>29.9</td>
<td>1.8</td>
<td>1.2</td>
<td>31.7</td>
<td>6.2</td>
<td>8.8</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>4.2*</td>
<td>29.0</td>
<td>1.7</td>
<td>31.8</td>
<td>5.9</td>
<td>10.2</td>
<td>10.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.99</td>
<td>4.1</td>
<td>28.9</td>
<td>1.6</td>
<td>32.4</td>
<td>5.7*</td>
<td>9.4</td>
<td>28.9</td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>1</td>
<td>.0</td>
<td>4.8</td>
<td>7.4</td>
<td>4.8</td>
<td>4.8</td>
<td>5.9</td>
<td>5.0*</td>
<td>25.9</td>
<td>5.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>4.8</td>
<td>7.7</td>
<td>4.8</td>
<td>6.0</td>
<td>5.0*</td>
<td>27.6</td>
<td>5.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.99</td>
<td>5.0*</td>
<td>7.5</td>
<td>5.0*</td>
<td>5.0*</td>
<td>5.2</td>
<td>24.2</td>
<td>5.8</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>.0</td>
<td>5.1</td>
<td>8.4</td>
<td>4.3</td>
<td>4.0</td>
<td>7.8</td>
<td>5.0*</td>
<td>10.8</td>
<td>5.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>5.1*</td>
<td>8.0</td>
<td>4.4</td>
<td>7.7</td>
<td>4.9*</td>
<td>10.0</td>
<td>6.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.99</td>
<td>4.4*</td>
<td>7.3</td>
<td>3.7</td>
<td>4.3</td>
<td>4.3</td>
<td>8.8</td>
<td>7.2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>.0</td>
<td>5.0*</td>
<td>9.1</td>
<td>3.9</td>
<td>3.5</td>
<td>10.4</td>
<td>4.6</td>
<td>8.1</td>
<td>5.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>4.8*</td>
<td>9.1</td>
<td>3.5</td>
<td>10.0</td>
<td>4.4</td>
<td>8.0</td>
<td>6.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.99</td>
<td>4.9*</td>
<td>9.1</td>
<td>3.8</td>
<td>10.0</td>
<td>4.5</td>
<td>7.8</td>
<td>11.1</td>
<td></td>
</tr>
<tr>
<td>.1</td>
<td>50</td>
<td>1</td>
<td>.0</td>
<td>3.6</td>
<td>16.3</td>
<td>3.6</td>
<td>3.6</td>
<td>4.5*</td>
<td>26.3</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>3.7</td>
<td>16.4</td>
<td>3.7</td>
<td>3.7</td>
<td>4.7*</td>
<td>26.4</td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.99</td>
<td>3.9</td>
<td>16.5</td>
<td>3.9</td>
<td>3.9</td>
<td>4.9*</td>
<td>28.8</td>
<td>25.2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>.0</td>
<td>5.7*</td>
<td>23.1</td>
<td>3.6</td>
<td>2.1</td>
<td>16.0</td>
<td>6.4</td>
<td>18.1</td>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>5.7*</td>
<td>22.6</td>
<td>3.7</td>
<td>2.1</td>
<td>15.8</td>
<td>6.6</td>
<td>18.2</td>
<td>18.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.99</td>
<td>5.0*</td>
<td>23.0</td>
<td>2.8</td>
<td>2.1</td>
<td>15.7</td>
<td>5.2</td>
<td>22.3</td>
<td>93.6</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>.0</td>
<td>6.4*</td>
<td>33.3</td>
<td>3.2</td>
<td>1.1</td>
<td>31.8</td>
<td>8.7</td>
<td>16.7</td>
<td>2.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>6.2*</td>
<td>33.5</td>
<td>3.4</td>
<td>1.0</td>
<td>32.0</td>
<td>9.2</td>
<td>18.7</td>
<td>36.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.99</td>
<td>5.7*</td>
<td>33.7</td>
<td>2.2</td>
<td>1.1</td>
<td>31.5</td>
<td>7.0</td>
<td>19.8</td>
<td>99.8</td>
</tr>
<tr>
<td>250</td>
<td>1</td>
<td>.0</td>
<td>4.9*</td>
<td>7.5</td>
<td>4.9*</td>
<td>4.9*</td>
<td>6.1</td>
<td>5.2</td>
<td>25.0</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>5.1*</td>
<td>8.0</td>
<td>5.1*</td>
<td>5.1*</td>
<td>6.4</td>
<td>5.2</td>
<td>25.2</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.99</td>
<td>4.9*</td>
<td>7.5</td>
<td>4.9*</td>
<td>4.9*</td>
<td>6.1</td>
<td>5.1*</td>
<td>26.6</td>
<td>12.8</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>.0</td>
<td>4.7*</td>
<td>7.8</td>
<td>4.0</td>
<td>3.8</td>
<td>6.9</td>
<td>4.5</td>
<td>15.3</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>4.9*</td>
<td>8.1</td>
<td>4.1</td>
<td>4.2</td>
<td>8.1</td>
<td>4.6</td>
<td>15.5</td>
<td>15.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.99</td>
<td>5.1*</td>
<td>8.1</td>
<td>4.5</td>
<td>4.3</td>
<td>7.8</td>
<td>4.9*</td>
<td>15.1</td>
<td>75.6</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>.0</td>
<td>5.2</td>
<td>9.3</td>
<td>4.0</td>
<td>3.8</td>
<td>9.8</td>
<td>4.9*</td>
<td>12.4</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>5.7</td>
<td>10.4</td>
<td>4.2</td>
<td>4.0</td>
<td>10.2</td>
<td>5.2*</td>
<td>11.9</td>
<td>34.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.99</td>
<td>5.2*</td>
<td>9.9</td>
<td>3.6</td>
<td>3.8</td>
<td>10.6</td>
<td>4.5</td>
<td>12.6</td>
<td>98.4</td>
</tr>
</tbody>
</table>

Asterisks in each row denote the number closest to the 5% significance level. The size results are computed using $R=10,000$ simulation repetitions.
(i) Power curves, strong instrument, normal errors

- $\text{LM}_\text{EL}$
- $K$
- $\text{LR}_\text{M}$
- $2\text{SLS}_{\text{HET}}$

true theta
(ii) Power curves, strong instrument, t[2] errors

The graph shows the power curves for different methods under strong instrument conditions. The x-axis represents the true theta, ranging from -0.4 to 0.4. The y-axis represents the power, ranging from 0 to 100. The curves are labeled as follows:

- $L_\text{MEL}$
- $K$
- $L_\text{RM}$
- $2\text{SLS}_{\text{HET}}$

The $L_\text{MEL}$ curve is the green solid line, $K$ is the cyan dotted line, $L_\text{RM}$ is the red dashed line, and $2\text{SLS}_{\text{HET}}$ is the magenta dotted line.
iii) Power curves, strong instrument, chi-square errors

- $L M_{EL}$
- $K$
- $L R_M$
- $2 S L S_{H E T}$

true theta