Econometric Theory, **21**, 2005, 667–709. Printed in the United States of America.
DOI: 10.1017/S0266466605050371

GENERALIZED EMPIRICAL LIKELIHOOD ESTIMATORS AND TESTS UNDER PARTIAL, WEAK, AND STRONG IDENTIFICATION

Patrik Guggenberger
UCLA

Richard J. Smith
cemmap, UCL and IFS
and
University of Warwick

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, *Econometrica* 68, 1055–1096) when identification may be weak. The paper makes two main contributions. First, 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. Second, 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

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 adviser, 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, ISEG/Universidade Tecnica de Lisboa, Konstanz, Laval, Madison (Wisconsin), Mannheim, Maryland, NYU, Penn, Penn State, Pittsburgh, Princeton, Rice, Riverside, Rochester, San Diego, Texas A&M, UCLA, USC, and Yale. We thank all the seminar participants. Guggenberger and Smith received financial support through a Carl Arvid Anderson Prize Fellowship and a 2002 Leverhulme Major Research Fellowship, respectively. Address correspondence to Patrik Guggenberger, Bunche Hall 8385, Department of Economics, U.C.L.A., Box 951477, Los Angeles, CA 90095-1477, USA; e-mail: guggenbe@econ.ucla.edu.

heteroskedasticity. The tests have competitive power properties, especially for thick-tailed or asymmetric error distributions.

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 variable (IV) 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 of this paper is to ascertain the performance of generalized empirical likelihood (GEL) methods (Newey and Smith, 2004; Smith, 1997, 2001) for time series IV models specified by nonlinear moment restrictions when identification may be weak (as in Stock and Wright, 2000). In particular, the paper makes two principal contributions. First, the asymptotic distribution of the GEL estimator is derived for a weakly identified setup. Second, 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 (Stock and Wright, 2000), 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 because of its competitive bias properties. For example, Newey and Smith (2004) show that for many models the asymptotic bias of empirical likelihood (EL) does not grow with the number of moment restrictions, whereas 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 Stock and Wright (2000) 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 Newey and Smith (2004) 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.

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

The statistics proposed here are asymptotically pivotal in contrast to classical Wald and likelihood ratio statistics no matter what 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 the GMM-based K-statistic of Kleibergen (2001) 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_ρ, is a score-type statistic, being a quadratic form in the GEL criterion score vector. The third statistic, LM_ρ, 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 overidentified 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 independent and identically distributed (i.i.d.) linear IV 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 the work of Stock and Wright (2000), 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_θ(ii), which follows. 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 (Guggenberger, 2003, Ch. 1) to the particular time series setting with nonlinear moment restrictions considered here seems worthwhile, especially because there is essentially no cost (in terms of complications of the proofs) to making this extension. The proofs for consistency and for the asymp-

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

totic distribution of the GEL estimator build on Guggenberger (2003), which adapts those given in Newey and Smith (2004) for the i.i.d. strongly identified context.

Subsequent to the i.i.d. linear version of this paper, two related papers have appeared. First, Caner (2003) derives the asymptotic distribution of the exponential tilting (ET) estimator (see Imbens, Spady, and Johnson, 1998; 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. Second, 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 $GELR_\rho$ and a modified version of the *K*-statistic based on the Kitamura and Stutzer (1997) and Smith (2001) kernel smoothed GEL estimator that is efficient under strong identification; see also Guggenberger and Smith (2003).

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 $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalue of A in absolute value, respectively. By A' 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 $A(A'K^{-1}A)^{-1}A'K^{-1}$ 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, $\text{vec}(M)$ stands for the column vectorization of the $k \times p$ matrix M; i.e., if $M = (m_1,\ldots,m_p)$ then $\text{vec}(M) = (m_1',\ldots,m_p')'$. Finally, $\|M\|$ equals the square root of the largest eigenvalue of $M'M$.

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

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

identified. Intuitively, then, the moment conditions that 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^l-valued data and, for each $n \in N$, $g_n : G \times \Theta \to R^k$ a given function, where $G \subset R^l$ and $\Theta \subset R^p$ denotes the parameter space. The model has a true parameter θ_0 for which the moment condition

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

is satisfied. For $g_n(z_i,\theta)$ we will usually write $g_i(\theta)$.

Example 1 (i.i.d. linear IV regression)

Guggenberger (2003, Ch. 1) discusses in detail GEL estimation and testing for this model under weak identification. The structural form (SF) equation is given by

$$y = Y\theta_0 + u, \tag{2.2}$$

and the reduced form (RF) for Y by

$$Y = Z\Pi + V, \tag{2.3}$$

where $y,u \in R^n$, $Y,V \in R^{n\times p}$, $Z \in R^{n\times k}$, and $\Pi \in R^{k\times p}$. The matrix Y may contain both exogenous and endogenous variables, $Y = (X,W)$ say, where $X \in R^{n\times p_X}$ and $W \in R^{n\times p_W}$ 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 Π 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, $Z_i,\ldots(i = 1,\ldots,n)$ the ith row of the matrix Y, V, $Z,\ldots$ written as a column vector. Assuming the instruments and the structural error are uncorrelated, $Eu_iZ_i = 0$, it follows that $Eg_i(\theta_0) = 0$, where for each $i = 1,\ldots,n$, $g_i(\theta) := (y_i - Y_i'\theta)Z_i$. Note that in this example $g_i(\theta)$ depends on n if the RF coefficient matrix Π 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 Stock and Wright (2000) the moment conditions may result from conditional moment restrictions. Assume $E[h(Y_i,\theta_0)|F_i] = 0$, where $h : H \times \Theta \to R^{k_1}$, $H \subset R^{k_2}$, 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 $l = k_2 + k_3$.

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

2.2. Assumptions

This section is concerned with the asymptotic distribution of the GEL estimator for θ when some components of $\theta_0 = (\alpha_0', \beta_0')'$, α_0 say, $\alpha_0 \in A$, $A \subset R^{p_A}$, are only weakly identified. Intuitively, this means that the moment condition (2.1) is not very informative about α_0. For parameter vectors $\theta = (\alpha', \beta_0')'$, $Eg_n(z_i, \theta)$ may be very close to zero, not only for α close to α_0 but also when α is far from α_0. In that case, the restriction $Eg_n(z_i, \theta_0) = 0$ is not very helpful for making inference on α_0. Assumption ID, which follows, provides a theoretical asymptotic framework for this phenomenon, which is taken from Assumption C in Stock and Wright (2000, p. 1061). We refer the reader to Stock and Wright (2000, pp. 1060–1061), which provides substantial detailed motivation for this assumption and an explanation of why it models α_0 as weakly and β_0 as strongly identified.

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

$$\hat{g}(\theta) := n^{-1} \sum_{i=1}^{n} g_i(\theta), \qquad \hat{G}(\theta) := n^{-1} \sum_{i=1}^{n} G_i(\theta),$$

$$\Psi_n(\theta) := n^{1/2}(\hat{g}(\theta) - E\hat{g}(\theta)),$$

$$\hat{\Omega}(\theta) := n^{-1} \sum_{i=1}^{n} g_i(\theta) g_i(\theta)',$$

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]

$$\Omega(\theta) := \lim_{n \to \infty} E n^{-1} \sum_{i=1}^{n} g_i(\theta) g_i(\theta)',$$

$$\Delta(\theta_1, \theta_2) := \lim_{n \to \infty} E\Psi_n(\theta_1)\Psi_n(\theta_2)' \quad \text{and} \quad \Delta(\theta) := \Delta(\theta, \theta).$$

Let $\theta = (\alpha', \beta')'$, where $\alpha \in A$, $A \subset R^{p_A}$, $\beta \in B$, $B \subset R^{p_B}$, and $p_A + p_B = p$. Also let $\mathcal{N} \subset B$ denote an open neighborhood of β_0.

Assumption Θ. 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 Θ, $m_1(\theta_0) = 0$, and $m_2(\beta) = 0$ if and only if $\beta = \beta_0$.

(ii) $m_2 \in C^1(\mathcal{N})$.

(iii) Let $M_2(\beta) := (\partial m_2 / \partial \beta)(\beta) \in R^{k \times p_B}$. The expression $M_2(\beta_0)$ has full column rank p_B.

Next we detail the necessary moment assumptions.[2]

Assumption M.

(i) $\max_{1\le i\le n}\sup_{\theta\in\Theta}\|g_i(\theta)\| = o_p(n^{1/2})$.

(ii) $\Omega(\cdot)$ is in $C^0(A\times\{\beta_0\})$ and bounded on Θ, $\Omega(\theta)$ is nonsingular for all $\theta\in A\times\{\beta_0\}$, $\sup_{\theta\in\Theta}\|\hat{\Omega}(\theta)-\Omega(\theta)\| = o_p(1)$, $\sup_{\theta\in A\times\mathcal{N}} n^{-1}\sum_{i=1}^{n}\|g_i(\theta)g_i(\theta)'\| = O_p(1)$.

(iii) $\Psi_n \Rightarrow \Psi$, where $\Psi(\theta)$ is a Gaussian stochastic process on Θ 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 $\Pr(\sup_{\theta\in A\times\mathcal{N}}\|\Psi(\theta)\|\le M_\varepsilon) > 1-\varepsilon$.

Assumption M(i) adapts Assumption 1(d) of Newey and Smith (2004), $E\sup_{\beta\in B}\|g_i(\beta)\|^{\xi}<\infty$ for some $\xi>2$, from the i.i.d. setting with strong identification ($p_A=0$ and thus $\theta=\beta$ and $\Theta=B$) to the weakly identified setup considered here. A sufficient condition for M(i) in the time series context and under ID is given by

$$\sup_{i\ge 1} E\sup_{\theta\in\Theta}\|g_i(\theta)\|^{\xi}<\infty \quad \text{for some } \xi>2. \tag{2.4}$$

Indeed, a simple application of the Markov inequality shows that (2.4) implies $\max_{1\le i\le n}\sup_{\theta\in\Theta}\|g_i(\theta)\| = O_p(n^{1/\xi}) = o_p(n^{1/2})$. See the Appendix for a proof. Assumption M(ii), which adapts Assumption 1(e) of Newey and Smith to the weakly identified setup, ensures that $\hat{\Omega}(\theta)$ is nonsingular for $\theta\in A\times\mathcal{N}$. Assumption M(iii) is essentially the "high-level" Assumption B of Stock and Wright (2000, p. 1059) that states that Ψ_n obeys a functional central limit theorem. In Assumption B′, Stock and Wright 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 $\sup_{\theta\in\Theta}\|\hat{g}(\theta)-E\hat{g}(\theta)\|\to_p 0$. In the proof of Theorem 2 we require $\sup_{\theta\in A\times\mathcal{N}}\|\Psi(\theta)\|$ 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 Newey and Smith (2004) even if $\xi=2$; see Owen (1990, Lemma 3) for a proof. Hence, using Lemmas 7–9 given subsequently, their Assumption 1(d) can be weakened to $\xi\ge 2$ 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 (continued)

See Guggenberger (2003). For the linear IV model (2.2) Assumption ID can be expressed as the following assumption.

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

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

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

$$E\hat{g}(\theta) = Eg_i(\theta) = E(Z_i Z_i')(n^{-1/2} C_A, \Pi_B)(\theta_0 - \theta),$$

which implies that in the notation of ID(i), $m_{1n}(\theta) = m_1(\theta) = E(Z_i Z_i') C_A(\alpha_0 - \alpha)$ and $m_2(\beta) = E(Z_i Z_i')\Pi_B(\beta_0 - \beta)$. 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 Π 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) $EZ_i U_i' = 0$;
(iii) $E\|Z_i\|^4 < \infty$, $Q_{ZZ} := E(Z_i Z_i') > 0$, $Eu_i^2 Z_i Z_i'$, $Eu_i V_{ij} Z_i Z_i'$, and $EV_{ij} V_{ik} Z_i Z_i'$ exist and are finite for $j, k = 1, \ldots, p$, where V_{ij} denotes the jth 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* Θ *hold in the linear* IV *model (2.2). Then Assumptions 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). $E(U_i U_i' | Z_i) = \Sigma_U > 0$.

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 $\Omega(\theta) = Q_{ZZ} v_\alpha' \Sigma_{uV_A} v_\alpha$ for $\theta \in A \times \{\beta_0\}$, where $v_\alpha' := (1, (\alpha_0 - \alpha)')$ and Σ_{uV_A} is the $(1 + p_A) \times (1 + p_A)$ upper left submatrix of Σ_U. However, M′ is more general than HOM because it allows for conditional het-

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

eroskedasticity. For example, $u_i = \|Z_i\|\zeta_i$, where $\zeta_i \sim N(0,1)$ is independent of $Z_i \sim N(0, I_k)$, is compatible with M′.

2.3. The GEL Estimator

This section provides a formal definition of the GEL estimator of θ_0.

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

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

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[3]

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

where

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

Assumption ρ.

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

The definition of the GEL estimator $\hat{\theta}_\rho$ is adopted from Newey and Smith (2004). 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 et al. (1998) and Kitamura and Stutzer (1997). For a recent survey of GEL methods see Imbens (2002).[4]

2.4. First-Order Equivalence

This section obtains the asymptotic distribution of the GEL estimator $\hat{\theta}_\rho$ under Assumption ID. Theorem 2 shows that the weakly identified parameters of θ_0 are estimated inconsistently and their GEL estimator has a nonstandard limiting distribution whereas 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

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

Stock and Wright (2000, 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 ρ satisfies Assumption ρ.[5] This is in contrast to the asymptotic theory for k-class estimators under weak identification. As shown in Staiger and Stock (1997, Theorem 1), the nonstandard asymptotic distribution of the k-class estimator depends on κ defined by $n(k-1) \to_d \kappa$. Therefore, LIML and 2SLS are not first-order equivalent under weak identification.

If defined, let

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

For $\theta = (\alpha',\beta')' \in \Theta$ and $b \in R^{p_B}$ 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* Θ, *ID, M, and* ρ *are satisfied. Then*

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

$$n^{1/2}(\hat{\beta} - \beta_0) = O_p(1).$$

(ii) The following more precise result holds. For any fixed $M > 0$ *let* $B_M := \{b \in R^{p_B} : \|b\| \le M\}$ *and define* $\theta_{\alpha b} := (\alpha', \beta_0' + n^{-1/2}b')'$. *Then, for* $(\alpha,b) \in A \times B_M$, $n\hat{P}(\theta_{\alpha b},\lambda(\theta_{\alpha b})) \Rightarrow P_{\alpha b} := P((\alpha',\beta_0')',b)$. *Assume there exists a random element* $(\alpha^*,b^*) \in A \times R^{p_B}$ *such that a.s.* $P_{\alpha^* b^*} < \inf_{(\alpha,b)\in(A\times R^{p_B})\backslash G} P_{\alpha b}$ *for every open set G that contains* (α^*,β^*). *Then*

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

Remark 1. Theorem 2(ii) is analogous to Theorem 1 in Stock and Wright (2000, p. 1062) for GMM. Note that from (A.5) in the Appendix $\lambda(\hat{\theta}) = -(\Omega((\hat{\alpha}',\beta_0')') + o_p(1))^{-1}\hat{g}(\hat{\theta}) = O_p(n^{-1/2})$ as $\hat{g}(\hat{\theta}) = O_p(n^{-1/2})$. Moreover, using the proof of Theorem 2 it can be shown that

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

Therefore, like $\hat{\beta}$, although $n^{1/2}$-consistent, $n^{1/2}\lambda(\hat{\theta})$ admits a nonstandard asymptotic distribution (see also Caner, 2003). If $p_A = 0$, where all parameters are strongly identified, $n^{1/2}\lambda(\hat{\theta}) \to_d N(0,\Omega^{-1}M_{M_2}(\Omega)\Delta M_{M_2}(\Omega)'\Omega^{-1})$, 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 (equation (A.7) in the Appendix) for $b^*(\alpha) := \arg\min_{b \in R^{p_B}} P_{\alpha b}$ for $\alpha \in A$. In particular, if $p_A = 0$, (A.7) shows that

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

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 preceding formula coincides with Theorem 3.2 of Newey and Smith (2004). 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 Newey and Smith, and the estimator $\hat{\beta}$ as defined previously 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 α_0.

Remark 2. Given the nonnormal asymptotic distribution of the GMM and GEL parameter estimates under weak identification (established in Theorem 1 in Stock and Wright, 2000, and our Theorem 2, 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 θ_0. The next section introduces alternative approaches that overcome these difficulties.

Example 1 (continued)

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 setup considered here there are at least two other statistics that share this property, namely, the Anderson and Rubin (1949) AR-statistic and the Kleibergen (2001, 2002a) K-statistic. The first statistic, $GELR_\rho(\theta)$, that we describe may be interpreted as a likelihood

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

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 first-order conditions (FOC) of $\hat{P}_\rho(\theta,\lambda)$ with respect to θ. 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[6]

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

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 setup and from EL to arbitrary ρ, the first statistic we consider is the renormalized GEL criterion function (2.7):

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

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

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

For $\theta \in \Theta$, define the $k \times p$ matrix

$$D_\rho(\theta) := \sum_{i=1}^n \rho_1(\lambda(\theta)' g_i(\theta)) G_i(\theta)/n. \tag{3.4}$$

Thus, (3.3) may be written as $\lambda(\hat{\theta})' D_\rho(\hat{\theta}) = 0'$. The test statistic is therefore given as a quadratic form in the score vector $\lambda(\theta)' D_\rho(\theta)$ evaluated at the hypothesized parameter vector θ

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

$$S_\rho(\theta) := n\lambda(\theta)' D_\rho(\theta)(D_\rho(\theta)'\tilde{\Omega}(\theta)^{-1}D_\rho(\theta))^{-1}D_\rho(\theta)'\lambda(\theta), \tag{3.5}$$

where ρ is any function satisfying Assumption ρ and $\tilde{\Omega}(\theta)$ is a consistent estimator of $\Delta(\theta)$. We also consider the following variant of $S_\rho(\theta)$:

$$\begin{aligned} LM_\rho(\theta) := {} & n\hat{g}(\theta)'\tilde{\Omega}(\theta)^{-1}D_\rho(\theta)(D_\rho(\theta)'\tilde{\Omega}(\theta)^{-1}D_\rho(\theta))^{-1} \\ & \times D_\rho(\theta)'\tilde{\Omega}(\theta)^{-1}\hat{g}(\theta) \end{aligned} \tag{3.6}$$

that substitutes $-\Delta(\theta)^{-1}\hat{g}(\theta)$ for $\lambda(\theta)$ in $S_\rho(\theta)$; see (A.8) in the Appendix, where it is shown that $n^{1/2}\lambda(\theta) = -\Delta(\theta)^{-1}n^{1/2}\hat{g}(\theta) + o_p(1)$. 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 preceding test statistics clearer, we adopted the notation $LM_\rho(\theta)$ and $S_\rho(\theta)$, respectively, in place of $K_\rho(\theta)$ and $K_\rho^L(\theta)$ 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 later, the sample average $\hat{\Omega}(\theta)$ may be used for $\tilde{\Omega}(\theta)$.[7] Note that when $\rho(v) = -(1+v)^2/2$, i.e., 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 $\lambda(\theta) = -\hat{\Omega}(\theta)^{-}\hat{g}(\theta)$ in the definition of $D_{CUE}(\theta)$, where $\hat{\Omega}(\theta)^{-}$ denotes any generalized inverse of $\hat{\Omega}(\theta)$.

As noted previously 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 the Kleibergen (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 $\hat{g}(\theta)'\tilde{\Omega}(\theta)^{-1}\hat{g}(\theta)$. It replaces $D_{CUE}(\theta)$ in $LM_{CUE}(\theta)$ by $\hat{G}(\theta) - \tilde{V}(\theta)(I_p \otimes (\tilde{\Omega}(\theta)^{-1}\hat{g}(\theta)))$, where $\tilde{V}(\theta)$ is an estimator for $\lim_{n\to\infty} E\{n^{-1}\sum_{i=1}^n\sum_{j=1}^n [G_i(\theta) - EG_i(\theta)][(I_p \otimes g_j(\theta)') - E(I_p \otimes g_j(\theta)')]\}$. The particular assumptions made on $\Delta(\theta)$ determine the choice of estimators $\tilde{\Omega}(\theta)$ and $\tilde{V}(\theta)$. If the sample average $\hat{\Omega}(\theta)$ 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, Newey and Smith (2004) 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 θ_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 θ_0. If, however, all parameters are weakly identified this argument is no longer valid. From Theorem 2, $\hat{\theta}$ is no longer consistent for θ_0. Therefore, although the FOC hold at $\hat{\theta}$, this does not imply automatically that they also approximately hold at the true value θ_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 θ_0. Consequently, the power properties of hypothesis tests for θ_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 subsequently and theoretically by Theorem 4.

We now 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) $\max_{1 \le i \le n} \|g_i(\theta)\| = o_p(n^{1/2})$;
(ii) $\Delta(\theta) > 0$, $\hat{\Omega}(\theta) \to_p \Delta(\theta)$, $n^{-1}\sum_{i=1}^n \|g_i(\theta)g_i(\theta)'\| = O_p(1)$;
(iii) $\Psi_n(\theta) \to_d \Psi(\theta)$, where $\Psi(\theta) \equiv N(0, \Delta(\theta))$.

Note that for $\theta = (\alpha', \beta_0')'$ M$_\theta$(iii) and ID imply that $\hat{g}(\theta) \to_p 0$. Thus, under M$_\theta$(iii) and ID the assumption $\hat{\Omega}(\theta) \to_p \Delta(\theta)$ in M$_\theta$(ii) is equivalent to the assumption $n^{-1}\sum_{i=1}^n (g_i(\theta) - \hat{g}(\theta))(g_i(\theta) - \hat{g}(\theta))' \to_p \Delta(\theta)$ for $\theta = (\alpha', \beta_0')'$, which is Assumption D$'$ in Stock and Wright (2000). 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 Stock and Wright, a possible application is the intertemporally separable consumption capital asset pricing model (CCAPM). Without assuming $\hat{\Omega}(\theta) \to_p \Delta(\theta)$, 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 setup considered here.

THEOREM 3. *Suppose ID, M$_\theta$(i)–(iii), and ρ hold for $\theta = (\alpha', \beta_0')'$. Then*

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

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

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

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 $\mathcal{M} \subset \Theta$ be an open neighborhood of θ.

Assumption M$_\theta$ (continued).

(iv) $\hat{g}(\cdot)$ is differentiable at $\bar{\theta}$ a.s. for each $\bar{\theta} \in \mathcal{M}$, $\hat{g}(\bar{\theta})$ is integrable (with respect to the probability measure) for all $\bar{\theta} \in \mathcal{M}$, $\sup_{\bar{\theta} \in \mathcal{M}} \|\hat{G}(\bar{\theta})\|$ is integrable, $m_{1n} \in C^1(\Theta)$, and $M_{1n}(\cdot) := (\partial m_{1n}/\partial\theta)(\cdot)$ converges uniformly on Θ to some function;

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

(v) $n^{-1}\sum_{i=1}^{n}(\text{vec}G_{iA}(\theta))g_i'(\theta)\to_p \Delta_A(\theta)$ ($\Delta_A(\theta)$ is defined below in (vii)), $\widetilde{\Omega}(\theta)\to_p \Delta(\theta)$, $\hat{G}_B(\theta):=n^{-1}\sum_{i=1}^{n}G_{iB}(\theta)\to_p E\hat{G}_B(\theta)$;

(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) $n^{-1/2}\sum_{i=1}^{n}((\text{vec}(G_{iA}(\theta)-EG_{iA}(\theta)))',(g_i(\theta)-Eg_i(\theta))')'\to_d N(0,V(\theta))$, where $V(\theta):=\lim_{n\to\infty}\text{var}(n^{-1/2}\sum_{i=1}^{n}((\text{vec}(G_{iA}(\theta))', g_i(\theta)')')\in R^{k(p_A+1)\times k(p_A+1)}$ is positive definite.

In M_θ(vii) write

$$V(\theta)=\begin{pmatrix}\Delta_{AA} & \Delta_A\\ \Delta_A' & \Delta\end{pmatrix}(\theta),\quad \text{where } \Delta_{AA}(\theta)\in R^{p_A k\times p_A k}.$$

Assumption M_θ(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 $M_{1n}(\theta)\to M_1(\theta):=(\partial m_1/\partial\theta)(\theta)$. Assumptions ID and M_θ thus imply that

$$E\hat{G}(\theta)=n^{-1/2}M_{1n}(\theta)+(0,M_2(\beta_0))\to(0,M_2(\beta_0)), \tag{3.7}$$

where by ID the limit matrix $(0,M_2(\beta_0))$ is of deficient rank p_B. Assumption M_θ(v) is comparable to M_θ(ii), where $\hat{\Omega}(\theta)\to_p\Delta(\theta)$ was assumed and extends M_θ(ii) to cross-product terms in vec $G_{iA}(\theta)$ and $g_i(\theta)$. Assumption M_θ(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_θ(vii), which is a stronger version of M_θ(iii) and states that a central limit theorem (CLT) holds simultaneously for the centered $g_i(\theta)$ and part of the derivative matrix, namely, vec $G_{iA}(\theta)$. Write $LM_\rho(\theta)=n\hat{g}'\widetilde{\Omega}^{-1}D(D'\widetilde{\Omega}^{-1}D)^{-1}D'\widetilde{\Omega}^{-1}\hat{g}$, where $D=D_\rho(\theta)$ and $\widetilde{\Omega}=\widetilde{\Omega}(\theta)$. As shown in the proof of Theorem 4, for $\theta=(\alpha',\beta_0')'$, Assumptions ID, ρ, M_θ(i)–(vi), and $\hat{G}_A(\theta):=n^{-1}\sum_{i=1}^{n}G_{iA}(\theta)\to_p E\hat{G}_A(\theta)$ imply that $D\to_p-(0,M_2(\beta_0))$. Therefore, the probability limit of $D'\widetilde{\Omega}^{-1}D$ is not invertible without renormalization. Define $D^*:=D\Lambda$ where the $p\times p$ diagonal matrix $\Lambda:=\text{diag}(n^{1/2},\ldots,n^{1/2},1,\ldots,1)$ with first p_A diagonal elements equal to $n^{1/2}$ and the remainder equal to unity. Hence,

$$LM_\rho(\theta)=n\hat{g}'\widetilde{\Omega}^{-1}D^*(D^{*\prime}\widetilde{\Omega}^{-1}D^*)^{-1}D^{*\prime}\widetilde{\Omega}^{-1}\hat{g}. \tag{3.8}$$

In the proof of Theorem 4 we show that under Assumptions ID, ρ, and M_θ(i)–(vi)

$$\text{vec}\,D^*=\text{vec}(0,-M_2(\beta_0))+\begin{pmatrix}-I_{kp_A} & \Delta_A(\theta)\Delta(\theta)^{-1}\\ 0 & 0\end{pmatrix}n^{-1/2}\sum_{i=1}^{n}\begin{pmatrix}\text{vec}G_{iA}(\theta)\\ g_i(\theta)\end{pmatrix}+o_p(1).$$

Assumption M_θ(vii), in particular the full rank assumption on $V(\theta)$, ensures that $D^{*\prime}\widetilde{\Omega}^{-1}D^*$ has full rank w.p.a.1. Assumption M_θ(vii) is closely related to

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

Assumption 1 of Kleibergen (2001). 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_θ(i)–(vii), and ρ hold for $\theta = (\alpha', \beta_0')'$. Then,*

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

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

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

Remark 1. The proof of Theorem 4 crucially hinges on the fact that $n^{1/2}\lambda(\theta_0)$ and $\operatorname{vec} D_\rho(\theta_0)$ (suitably normalized) from the FOC (3.3) are asymptotically jointly normally distributed and, moreover, are asymptotically independent. A similar result is critical also for the Kleibergen (2001) K-statistic, which generalizes the Brown and Newey (1998) analysis of efficient GMM moment estimation to the weakly identified setup. 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 θ_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 $\{GELR_\rho(\theta_0) \geq \chi_r^2(k)\}$, where $\chi_r^2(k)$ denotes the $(1 - r)$-critical value from the $\chi^2(k)$ distribution. A $(1 - r)$-confidence region for θ_0 is obtained by inverting the just-described test, i.e., $\{\theta \in \Theta : GELR_\rho(\theta) \leq \chi_r^2(k)\}$. 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 δ does not diverge to infinity 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 α_0 from the data.

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

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 overidentified 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 overidentification. The AR-statistic of Anderson and Rubin (1949) has a $\chi^2(k)$ limiting null distribution also. 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. Stock and Wright (2000, 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 $H_0\colon \theta = \theta_0$, which is shown to be asymptotically distributed as $\chi^2(p)$ under strong identification. However, because of 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 (continued)

Guggenberger (2003) derives the results given in Theorems 3 and 4 under Assumptions Θ, ID$'$, M$'$, and ρ allowing for alternatives $\alpha \in A$ and Pitman drift in the data generating process (DGP) for the strongly identified parameters to assess the asymptotic power properties of the tests; i.e., ID$'$ holds and for some fixed $b \in R^{p_B}$, $y = Y(\theta_0 + n^{-1/2}(0', b')') + u$. 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 preceding theorems because, as is easily shown, Assumptions ID$'$, M$'$, ρ, and $V(\theta) > 0$ imply M$_\theta$ for any consistent estimator $\tilde{\Omega}(\theta)$ of $\Omega(\theta)$. In particular, $V(\theta)$ has a simple representation. For $\theta = (\alpha', \beta_0')'$, $\Omega(\theta) = \Delta(\theta)$ and $\Delta_{AA}(\theta) = E(V_{iA}V_{iA}' \otimes Z_i Z_i')$, 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 focused 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, α_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 β by a GEL estimator $\hat{\beta}(\alpha)$. To make this idea more rigorous, define the GEL estimator $\hat{\beta}(\alpha)$ for β_0:

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

$$\hat{\beta}(\alpha) := \arg\min_{\beta \in B} \sup_{\lambda \in \hat{\Lambda}_n(\alpha',\beta')'} \hat{P}((\alpha',\beta')',\lambda). \tag{4.1}$$

We usually write $\hat{\beta}$ for $\hat{\beta}(\alpha)$ where there is no ambiguity. A requirement of the analysis that follows is that $\hat{\beta} \rightarrow_p \beta_0$ if $\alpha = \alpha_0$. Therefore, we assume that the nuisance parameters β_0 that are not involved in the hypothesis under test are strongly identified; see Theorem 2. On the other hand, the components of α_0 can be weakly or strongly identified, and in Assumption ID$_\alpha$, which follows, we assume the former holds for α_{01} and the latter for α_{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 α_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 θ_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 $\alpha_j \in A_j$, $A_j \subset R^{p_{A_j}}$, $(j = 1,2)$, $p_{A_1} + p_{A_2} = p_A$ and $\beta \in B$, $B \subset R^{p_B}$. Also let $\mathcal{N} \subset A_2 \times B$ be an open neighborhood of (α_{02},β_0).

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

Assumption ID$_\alpha$.

(i) $E\hat{g}(\theta) = n^{-1/2}m_{1n}(\theta) + m_2(\alpha_2,\beta)$, where $m_{1n}, m_1 : \Theta \rightarrow R^k$ and $m_2 : A_2 \times B \rightarrow R^k$ are continuous functions such that $m_{1n}(\theta) \rightarrow m_1(\theta)$ uniformly on Θ, $m_1(\theta_0) = 0$, and $m_2(\alpha_2,\beta) = 0$ if and only if $(\alpha_2,\beta) = (\alpha_{02},\beta_0)$.

(ii) $m_2 \in C^1(\mathcal{N})$.

(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 that α_{01} and (α_{02},β_0) are weakly and strongly identified, respectively. Assumptions A and ID$_\alpha$ adapt Assumptions Θ and ID in Section 2 for the subvector case.

Let

$$\hat{\theta}_\alpha := (\alpha',\hat{\beta}(\alpha)')' \quad \text{and} \quad \theta_{\alpha\beta} := (\alpha',\beta')'.$$

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

$$GELR_\rho^{sub}(\alpha) := GELR_\rho(\hat{\theta}_\alpha).$$

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

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, α_{01} here.

For $a_1 \in A_1$ let $a := (a_1', \alpha_{02}')'$ be a fixed vector whose strongly identified component α_{02} is the same as the corresponding component of the true parameter vector θ_0. Let $\mathcal{M} \subset B$ be an open neighborhood of β_0.

Assumption M_α.

(i) $\max_{1\leq i\leq n} \sup_{\beta\in B} \|g_i(\theta_{a\beta})\| = o_p(n^{1/2})$;

(ii) $\sup_{\beta\in B} \|\hat{\Omega}(\theta_{a\beta}) - \Gamma(\theta_{a\beta})\| \to_p 0$ for some matrix $\Gamma(\cdot)$ that is uniformly bounded on $\{\theta_{a\beta} : \beta \in B\}$, continuous at $\theta_{a\beta_0}$ and $\Gamma(\theta_{a\beta_0}) = \Delta(\theta_{a\beta_0}) > 0$ and $n^{-1}\sum_{i=1}^n \|g_i(\theta_{a\beta_0}) g_i(\theta_{a\beta_0})'\| = O_p(1)$;

(iii) $\Psi_n(\theta_{a\beta_0}) \to_d \Psi(\theta_{a\beta_0})$, where $\Psi(\theta_{a\beta_0}) \equiv N(0, \Delta(\theta_{a\beta_0}))$;

(iv) $\hat{G}_B(\cdot) := n^{-1}\sum_{i=1}^n (\partial g_i/\partial\beta)(\cdot)$ exists at $\theta_{a\beta}$ a.s. for each $\beta \in \mathcal{M}$, $\hat{g}(\theta_{a\beta})$ is integrable for all $\beta \in \mathcal{M}$, $\sup_{\beta\in\mathcal{M}} \|\hat{G}_B(\theta_{a\beta})\|$ 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 $\beta \in \mathcal{M}$;

(vi) $\sup_{\beta\in\mathcal{M}} n^{-1}\sum_{i=1}^n \|G_{iB}(\theta_{a\beta})\| = O_p(1)$.

Mutatis mutandis, M_α has the same interpretation as M_θ. For example M_α(ii) guarantees that $\lambda_{\max}(\hat{\Omega}(\hat{\theta}_a))$ is bounded and $\lambda_{\min}(\hat{\Omega}(\hat{\theta}_a))$ is bounded away from zero w.p.a.1, whereas M_α(iv) and ID_α imply that for $\beta \in \mathcal{M}$ we have $E\hat{G}_B(\theta_{a\beta}) = n^{-1/2}(\partial m_{1n}/\partial\beta)(\theta_{a\beta}) + (\partial m_2/\partial\beta)(\alpha_{02}, \beta) \to (\partial m_2/\partial\beta)(\alpha_{02}, \beta)$. By ID_α this last matrix has full column rank for $\beta = \beta_0$. If we assume that the $G_{iB}(\theta_{a\beta})$, $(i = 1, \ldots, n)$, viewed as functions of β, are continuous at β_0 a.s., then we can simplify M_α(vi) to $n^{-1}\sum_{i=1}^n \|G_{iB}(\theta_{a\beta_0})\| = O_p(1)$. A similar comment holds for the assumptions in the continuation of M_α that follows.

THEOREM 5. *Assume $1 \leq p_A < p$. Suppose Assumptions A, ID_α, M_α(i)–(vi), and ρ hold for some $a_1 \in A_1$ and $a = (a_1', \alpha_{02}')'$. Then,*

$$GELR_\rho^{sub}(a) \to_d \chi^2(k - p_B, \delta),$$

where the noncentrality parameter δ is given by

$$\delta := m_1(\theta_{a\beta_0})'\Delta(\theta_{a\beta_0})^{-1} M_{M_{2\beta}(\alpha_{02},\beta_0)}(\Delta(\theta_{a\beta_0})) m_1(\theta_{a\beta_0}),$$

where $M_{2\beta}(\cdot) := (\partial m_2/\partial\beta)(\cdot) \in R^{k\times p_B}$. In particular,

$$GELR_\rho^{sub}(\alpha_0) \to_d \chi^2(k - p_A).$$

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

Theorem 5 confirms that the subvector statistic $GELR_{\rho}^{sub}(\alpha_0)$, like the full vector statistic $GELR_{\rho}(\theta_0)$, is asymptotically pivotal. As before, this result can be used to construct hypothesis tests and confidence regions for α_0.

We now generalize the statistics S_{ρ} and LM_{ρ} to the subvector case. The asymptotic variance matrices of $n^{1/2}\hat{g}(\hat{\theta}_{\alpha})$ and $n^{1/2}\lambda(\hat{\theta}_{\alpha})$ differ from those of $n^{1/2}\hat{g}(\theta_{\alpha\beta_0})$ and $n^{1/2}\lambda(\theta_{\alpha\beta_0})$. Therefore different weighting matrices are required in the quadratic forms defining these subvector statistics. In the Appendix (see proofs of Theorems 5 and 6) it is shown that for $a = (a_1', \alpha_{02}')'$, $\lambda(\hat{\theta}_a) = \arg\max_{\lambda \in \hat{\Lambda}_n(\hat{\theta}_a)} \hat{P}(\hat{\theta}_a, \lambda)$ exists w.p.a.1 and that $n^{1/2}\lambda(\hat{\theta}_a)$ is asymptotically normal with covariance matrix $M(a)$, where for $\alpha = (\alpha_1', \alpha_2')' \in R^{p_A}$

$$M(\alpha) := \Delta(\theta_{\alpha\beta_0})^{-1} M_{M_{2\beta}(\alpha_2, \beta_0)}(\Delta(\theta_{\alpha\beta_0})). \tag{4.2}$$

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

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

which coincides with the definition of $D_{\rho}(\theta)$ in (3.4) when α is the full vector θ. Similarly to $S_{\rho}(\theta)$ in (3.5) the subvector test statistic $S_{\rho}^{sub}(\alpha)$ 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 $\widetilde{M}(\alpha)$ be an estimator of $M(\alpha)$ that is given by replacing the expressions $\Delta(\theta_{\alpha\beta_0})$ and $M_{2\beta}(\alpha_2, \beta_0)$ in $M(\alpha)$ by consistent estimators, $\widetilde{\Omega}$ and $\widetilde{M}_2$ say. By Assumptions M_{α}(ii) and M_{α}(iv)–(v) we may choose $\hat{\Omega}(\hat{\theta}_a)$ for $\widetilde{\Omega}$ and $\hat{G}_B(\hat{\theta}_a)$ for $\widetilde{M}_2$ when $\alpha = a = (a_1', \alpha_{02}')'$. Hence,

$$S_{\rho}^{sub}(\alpha) := n\lambda(\hat{\theta}_{\alpha})' D_{\rho}(\alpha)(D_{\rho}(\alpha)' \widetilde{M}(\alpha) D_{\rho}(\alpha))^{-1} D_{\rho}(\alpha)' \lambda(\hat{\theta}_{\alpha}).$$

The statistic $LM_{\rho}^{sub}(\alpha)$ is constructed like $S_{\rho}^{sub}(\alpha)$ but replaces $\lambda(\hat{\theta}_{\alpha})$ by $-\widetilde{\Omega}^{-1}\hat{g}(\hat{\theta}_{\alpha})$. Thus,

$$LM_{\rho}^{sub}(\alpha) := n\hat{g}(\hat{\theta}_{\alpha})' \widetilde{\Omega}^{-1} D_{\rho}(\alpha)(D_{\rho}(\alpha)' \widetilde{M}(\alpha) D_{\rho}(\alpha))^{-1} D_{\rho}(\alpha)' \widetilde{\Omega}^{-1} \hat{g}(\hat{\theta}_{\alpha}).$$

Let $a = (a_1', \alpha_{02}')'$, $\mathcal{M} \subset B$ be an open neighborhood of β_0, and $\hat{G}_{A_j}(\theta) := n^{-1} \sum_{i=1}^{n} (\partial g_i / \partial \alpha_j)(\theta)$, $(j = 1,2)$.

Assumption M_{α} (continued).

(vii) $\hat{G}_{A_1}(\theta_{a\beta})$ viewed as a function in β is continuously differentiable at β a.s. for each $\beta \in \mathcal{M}$, $(\partial \operatorname{vec} \hat{G}_{A_1} / \partial\beta)(\theta_{a\beta}) \to_p E(\partial \operatorname{vec} \hat{G}_{A_1} / \partial\beta)(\theta_{a\beta}) = (\partial E \operatorname{vec} \hat{G}_{A_1} / \partial\beta)(\theta_{a\beta})$, $\hat{G}_A(\theta_{a\beta}) \to_p E\hat{G}_A(\theta_{a\beta}) = (\partial E\hat{g}/\partial\alpha)(\theta_{a\beta})$, $(\partial \operatorname{vec}(\partial m_{1n}/\partial\alpha_1)/\partial\beta)(\theta_{a\beta}) \to (\partial \operatorname{vec}(\partial m_1/\partial\alpha_1)/\partial\beta)(\theta_{a\beta})$,

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

where convergence is uniform over $\beta \in \mathcal{M}$ in all cases, $\partial m_{1n}/\partial\alpha(\cdot)$ is continuous at $\theta_{a\beta}$ a.s. for each $\beta \in \mathcal{M}$, and $\partial m_{1n}/\partial\alpha(\theta_{a\beta})$ converges uniformly over $\beta \in \mathcal{M}$ to some function;

(viii) $n^{-1}\sum_{i=1}^{n}(\text{vec}\, G_{iA_1}(\theta_{a\beta}))g_i'(\theta_{a\beta}) \rightarrow_p \Phi(\theta_{a\beta})$ 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) of this assumption), $\tilde{\Omega}(\hat{\theta}_a) \rightarrow_p \Delta(\theta_{a\beta_0})$;

(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) $n^{-1/2}\sum_{i=1}^{n}((\text{vec}(G_{iA_1}(\theta_{a\beta_0}) - EG_{iA_1}(\theta_{a\beta_0})))', (g_i(\theta_{a\beta_0}) - Eg_i(\theta_{a\beta_0}))')' \rightarrow_d N(0, V^{\alpha}(\theta_{a\beta_0}))$, where $V^{\alpha}(\theta_{a\beta_0})$ is the appropriate submatrix of $V(\theta_{a\beta_0})$ defined in M_θ(vii); $V^{\alpha}(\theta_{a\beta_0})$ is positive definite.

In M_α(x) write

$$V^{\alpha}(\theta) = \begin{pmatrix} \Delta_{A_1A_1} & \Delta_{A_1} \\ \Delta_{A_1}' & \Delta \end{pmatrix}(\theta), \quad \text{where } \Delta_{A_1A_1}(\theta) \in R^{p_{A_1}k \times p_{A_1}k}.$$

Assumption M_α(x) is the key assumption and plays a role similar to M_θ(vii). Assumption M_α(vii) extends M_α(iv) by explicitly assuming that integration and differentiation can be exchanged in the expectation of $\hat{G}_{A_1}(\theta_{a\beta})$, whereas M_α(iv) gave primitive conditions that imply that exchange holds for $\hat{g}(\theta_{a\beta})$. Assumptions M_α(v), M_α(vii), and ID_α imply that $(\partial\,\text{vec}\,\hat{G}_{A_1}/\partial\beta)(\hat{\theta}_a) \rightarrow_p 0$, which is an important result used in the proof of the next theorem; in a linear model this result is trivially true because $\partial\,\text{vec}\,\hat{G}_{A_1}/\partial\beta \equiv 0$. Assumptions M_α(vii)–(x) are analogous to M_θ(iv)–(vii) with A_1 and A_2 now playing the roles of A and B, respectively.

THEOREM 6. *Assume $1 \le p_A < p$. Suppose Assumptions A, ID_α, M_α(i)–(x), and ρ hold for $a = (a_1', \alpha_{02}')'$ for $a_1 \in A_1$. Then,*

$$S_\rho^{sub}(a), LM_\rho^{sub}(a) \rightarrow_d (W_\alpha(a) + \zeta_\alpha)'(W_\alpha(a) + \zeta_\alpha),$$

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 ζ_α and W_α are independent. We have $W_\alpha(\alpha_0) \equiv 0$, and therefore

$$S_\rho^{sub}(\alpha_0), LM_\rho^{sub}(\alpha_0) \rightarrow_d \chi^2(p_A).$$

Remark 1. The subvector statistics are asymptotically pivotal when elements of α_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 α_0. Compared to the GMM-subvector statistic of Kleibergen (2001)the statistics $S_\rho^{sub}(a)$ and $LM_\rho^{sub}(a)$ are appealing because of their compact formulation.

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

Remark 2. Even though it is unclear how the asymptotic distribution of these test statistics might be derived without assuming strong identification of β_0, it is obvious that neither $S_\rho^{sub}(\alpha_0)$ nor $LM_\rho^{sub}(\alpha_0)$ would converge to a $\chi^2(p_A)$ random variable. In general the quantities $n^{1/2}\lambda(\hat{\theta}_{\alpha_0})$ in $S_\rho^{sub}(\alpha_0)$ and $n^{1/2}\hat{g}(\hat{\theta}_{\alpha_0})$ in $LM_\rho^{sub}(\alpha_0)$ are no longer asymptotically normal because of their dependence on the GEL estimator $\hat{\beta}(\alpha_0)$, which as a direct consequence of Theorem 2 has a nonstandard limiting distribution if β_0 is not strongly identified. Moreover, the subvector version of the K-statistic of Kleibergen (2001) also experiences the same problem in these circumstances as the (GMM) CUE of β_0 has a nonnormal limiting distribution under weak identification (see Stock and Wright, 2000). Somewhat surprisingly, however, Monte Carlo simulations by the authors (not reported here) for the subvector statistic $LM_\rho^{sub}(\alpha_0)$ indicate that its size properties are not much affected by the strength or weakness of identification of β_0. Startz, Zivot, and Nelson (2004) report similar findings from Monte Carlo simulations for the subvector test statistic of Kleibergen (2001).

Example 1 (continued)

Guggenberger (2003) derives the corresponding results. Note that Assumptions Θ, ID$'$, M$'$, and ρ 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_1} components. Let $Y = (X, W)$ denote the partition of the included variables of the structural equation into exogenous and endogenous variables. Partition $\theta_0 = (\theta_{X0}', \theta_{W0}')'$ and $\theta = (\theta_X', \theta_W')'$ conformably. Valid inference is possible on any subvector of θ_{W0} if the appropriate assumptions given previously 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., $M_Xy = M_XW\theta_{W0} + M_Xu$. If Assumption M$_\alpha$ holds for $\theta_{W0} = (\alpha_{W0}, \beta_{W0})$ and $g_i(\theta_W) := M_{X,i}'(y - W\theta_W)Z_i$, where $M_{X,i}$ denotes the ith row of M_X written as a column vector, valid inference may be undertaken on α_{W0}.

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 DGP is given by model (2.2) considered in Example 1 and is similar to that in Kleibergen (2002a, p. 1791), namely,

$$y = Y\theta_0 + u, \qquad \textbf{(5.1)}$$

$$Y = Z\Pi + V.$$

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

There are 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, i.e., $k = 1$, $\Pi = \Pi_1$, whereas, in the overidentified 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 $\Sigma \in R^{2\times 2}$ with diagonal elements unity and off-diagonal elements ρ_{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 $B_i|u_i + 2| - (1 - B_i)|u_i + 2|$ where B_i is Bernoulli (0.5,0.5) 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 $GELR_\rho(\theta)$ in (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 $S_\rho(\theta)$ in (3.5) and $LM_\rho(\theta)$ in (3.6) with $\tilde{\Omega}(\theta)$ replaced by $\hat{\Omega}(\theta)$. As noted previously, for CUE, $S_\rho(\theta)$ and $LM_\rho(\theta)$ 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

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

Kleibergen (2001, 2002a), one assuming homoskedastic errors K, the other robust to conditional heteroskedasticity K_{HET}, the conditional likelihood ratio test LR_M of Moreira (2003), and two versions of the two-stage least squares (2SLS) Wald statistic $2SLS$ (see, e.g., Wooldridge, 2002, pp. 98, 100), one assuming homoskedastic errors ($2SLS_{HOM}$) and the other robust to conditional heteroskedasticity ($2SLS_{HET}$).[9] Under $H_0\colon \theta_0 = 0$, $AR(\theta_0) \to_d \chi^2(k)$ and $K(\theta_0) \to_d \chi^2(p)$. 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 $H_0\colon \theta = \theta_0$ and strong identification.

5.3. Size Comparison

Empirical sizes are calculated using 5% asymptotic critical values for all of the preceding statistics for DGPs (5.1) corresponding to all 54 possible combinations of sample size $n = 50, 100, 250$, number of instruments $k = 1, 5, 10$, SF and RF error correlation $\rho_{uV} = 0.0, 0.5, 0.99$, and RF coefficient $\Pi_1 = 0.1, 1.0$ for Designs (I)–(IV) and (I_{HET})–(IV_{HET}).[10]

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 Tables 1 and 2, which follow, 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 ρ_{uV} allows the degree of endogeneity of Y to be varied. Whereas for $\rho_{uV} = 0$, Y is exogenous, Y is strongly endogenous for $\rho_{uV} = 0.99$. We include the just-identified case, $k = 1$, and two overidentified-cases, $k = 5$ and 10.

We now describe the results for Designs (I) and (I_{HET}) given in Tables 1 and 2, respectively, which exclude those for $GELR_{EL}$, S_{ET}, 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).

First, we consider the separate effects of Π_1, n, ρ_{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 Π_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. Although 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

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

Table 1. Size results for Design (I) at 5% significance level

Π_1	n	k	ρ_{uV}	LM_{EL}	S_{EL}	LM_{CUE}	$GELR_{CUE}$	$GELR_{ET}$	K	LR_M	$2SLS_{HOM}$
1	50	1	0.0	4.6	9.0	4.6	4.6	6.3	5.1*	5.9	4.3
			0.5	4.8*	9.0	4.8*	4.8*	6.7	5.7	5.4	5.2*
			0.99	4.9*	9.1	4.9*	4.9*	6.9	6.1	5.7	6.2
		5	0.0	4.3	15.5	2.6	2.8	13.5	5.6*	6.3	3.8
			0.5	4.1	15.0	2.6	2.5	13.3	5.4	5.2*	5.8
			0.99	4.2	15.0	2.6	2.7	12.9	5.6	5.4*	12.3
		10	0.0	4.4	26.8	1.6	1.3	28.9	5.9	5.4*	2.8
			0.5	4.2*	26.1	1.7	1.5	28.8	6.3	6.6	9.2
			0.99	4.0	25.8	1.6	1.3	29.5	5.7*	6.3	27.1
	250	1	0.0	5.3	5.6	5.3	5.3	5.8	5.3	5.3	5.1*
			0.5	5.1*	5.4	5.1*	5.1*	5.4	5.3	5.9	5.2
			0.99	5.0*	5.4	5.0*	5.0*	5.3	5.1	4.3	5.2
		5	0.0	5.0*	6.5	4.4	4.4	6.4	5.2	5.6	4.8
			0.5	5.0*	6.3	4.4	4.6	6.6	5.3	5.2	5.6
			0.99	4.5	5.9	4.0	4.5	6.8	4.8*	4.4	6.7
		10	0.0	4.8*	7.6	3.7	4.1	8.9	5.3	5.4	4.7
			0.5	4.7	7.6	3.5	4.0	8.8	5.1	5.0*	6.1
			0.99	4.8	7.5	3.5	3.8	8.9	5.1*	5.2	10.9
0.1	50	1	0.0	4.4	8.9	4.4	4.4	6.4	5.3*	5.3*	0.0
			0.5	4.4*	8.8	4.4*	4.4*	6.6	5.7	5.7	2.2
			0.99	4.8*	8.9	4.8*	4.8*	6.5	5.7	6.5	24.9
		5	0.0	5.6*	17.4	3.8	2.7	13.6	6.6	7.1	0.6
			0.5	5.3*	16.9	3.5	2.7	12.9	6.6	7.1	16.3
			0.99	4.1*	15.9	2.7	2.7	13.2	6.0	6.7	96.5
		10	0.0	6.2*	30.3	3.1	1.3	29.3	8.6	9.2	1.0
			0.5	6.1*	30.5	3.2	1.3	28.9	8.2	10.3	34.3
			0.99	5.0*	27.4	2.0	1.3	28.9	6.7	6.5	99.9
	250	1	0.0	4.8	5.3	4.8	4.8	5.2	5.1*	4.5	0.1
			0.5	5.1*	5.5	5.1*	5.1*	5.4	5.3	4.5	3.3
			0.99	5.2*	5.6	5.2*	5.2*	5.6	5.3	5.4	13.2
		5	0.0	4.7	6.2	4.1	4.0	6.0	4.8*	5.2*	0.5
			0.5	4.9*	6.2	4.2	4.5	6.9	4.8	5.5	15.8
			0.99	5.1*	6.4	4.6	4.8	6.9	5.5	6.2	80.1
		10	0.0	5.3*	7.9	4.0	3.9	8.5	5.3*	5.7	1.5
			0.5	5.3	8.4	4.0	4.3	8.8	5.3	5.1*	35.2
			0.99	4.9*	7.8	3.7	4.0	9.0	5.4	6.1	98.8

Note: 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 2. Size results for Design (I_{HET}) at 5% significance level

Π_1	n	k	ρ_{uV}	LM_{EL}	S_{EL}	LM_{CUE}	$GELR_{CUE}$	$GELR_{ET}$	K_{HET}	LR_M	$2SLS_{HET}$
1	50	1	0.0	3.9	15.9	3.9	3.9	8.1	4.9*	27.7	7.8
			0.5	4.2	16.8	4.2	4.2	8.4	5.0*	26.1	8.2
			0.99	3.9	16.3	3.9	3.9	7.8	4.8*	25.7	8.1
		5	0.0	4.3	20.2	2.8	2.2	16.1	5.1*	13.1	6.0
			0.5	4.0	19.9	2.6	2.0	15.9	4.6*	11.3	7.9
			0.99	4.0	19.7	2.3	2.0	15.3	4.7*	10.8	14.4
		10	0.0	4.5*	29.9	1.8	1.2	31.7	6.2	8.8	4.4
			0.5	4.2*	29.0	1.7	1.2	31.8	5.9	10.2	10.7
			0.99	4.1	28.9	1.6	1.2	32.4	5.7*	9.4	28.9
	250	1	0.0	4.8	7.4	4.8	4.8	5.9	5.0*	25.9	5.7
			0.5	4.8	7.7	4.8	4.8	6.0	5.0*	27.6	5.9
			0.99	5.0*	7.5	5.0*	5.0*	5.9	5.2	24.2	5.8
		5	0.0	5.1	8.4	4.3	4.0	7.8	5.0*	10.8	5.6
			0.5	5.1*	8.0	4.4	4.1	7.7	4.9*	10.0	6.0
			0.99	4.4*	7.3	3.7	4.3	8.0	4.3	8.8	7.2
		10	0.0	5.0*	9.1	3.9	3.5	10.4	4.6	8.1	5.2
			0.5	4.8*	9.1	3.5	3.6	10.0	4.4	8.0	6.5
			0.99	4.9*	9.1	3.8	3.6	10.0	4.5	7.8	11.1
0.1	50	1	0.0	3.6	16.3	3.6	3.6	7.7	4.5*	26.3	0.4
			0.5	3.7	16.4	3.7	3.7	8.1	4.7*	26.4	3.1
			0.99	3.9	16.5	3.9	3.9	8.1	4.9*	28.8	25.2
		5	0.0	5.7*	23.1	3.6	2.1	16.0	6.4	18.1	1.4
			0.5	5.7*	22.6	3.7	2.1	15.8	6.6	18.2	18.4
			0.99	5.0*	23.0	2.8	2.1	15.7	5.2	22.3	93.6
		10	0.0	6.4*	33.3	3.2	1.1	31.8	8.7	16.7	2.2
			0.5	6.2*	33.5	3.4	1.0	32.0	9.2	18.7	36.1
			0.99	5.7*	33.7	2.2	1.1	31.5	7.0	19.8	99.8
	250	1	0.0	4.9*	7.5	4.9*	4.9*	6.1	5.2	25.0	0.3
			0.5	5.1*	8.0	5.1*	5.1*	6.4	5.2	25.2	3.2
			0.99	4.9*	7.5	4.9*	4.9*	6.1	5.1*	26.6	12.8
		5	0.0	4.7*	7.8	4.0	3.8	6.9	4.5	15.3	0.8
			0.5	4.9*	8.1	4.1	4.2	8.1	4.6	15.5	15.9
			0.99	5.1*	8.1	4.5	4.3	7.8	4.9*	15.1	75.6
		10	0.0	5.2	9.3	4.0	3.8	9.8	4.9*	12.4	1.8
			0.5	5.7	10.4	4.2	4.0	10.2	5.2*	11.9	34.3
			0.99	5.2*	9.9	3.6	3.8	10.6	4.5	12.6	98.4

Note: Asterisks in each row denote the number closest to the 5% significance level. The size results are computed using $R = 10{,}000$ simulation repetitions.

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

$2SLS$ statistics, $GELR_{ET}$ and S_{EL}, severely overreject in overidentified and strongly endogenous cases when $n = 50$. Even though they still overreject 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 ρ_{uV}. The slight dependence of the size results in Table 1 on ρ_{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 ρ_{uV} affects their size properties. Moreover, there is little dependence of the results on ρ_{uV}. However, for $2SLS$, increasing ρ_{uV} leads to severe overrejection when combined with overidentification, especially so in the weak instrument case.

Increasing the number of instruments k usually leads to overrejection 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 overidentified situations. In all other cases, $2SLS$ has competitive size properties. The statistics $GELR_{ET}$ and S_{EL} severely overreject in overidentified 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 underreject, especially in overidentified 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 overreject. 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.

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 overreject, 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 overidentified 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

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

closest to the 5% significance level across all the statistics. The test statistic 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 preceding 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. Whereas with normal errors and weak identification $2SLS$ severely overrejects, with these error distributions it severely underrejects.

To summarize this size study, LM_{EL}, LM_{CUE}, $GELR_{CUE}$, and K_{HET} have reliable size properties across all designs that appear independent of both the strength or weakness of identification and possible conditional heteroskedasticity. The test statistic $2SLS$ 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 preceding statistics and DGPs (5.1) corresponding to all 16 possible combinations of sample size $n = 100$, 250, number of instruments $k = 5$, 10, SF and RF error correlation $\rho_{uV} = 0.5$, 0.99, and RF 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 preceding size comparison. We do so because size correction of LR_M is not straightforward as a result of the conditional construction of LR_M and, as shown before, 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 θ_0. The null hypothesis under test is again $H_0 : \theta_0 = 0$. For weak identification ($\Pi_1 = 0.1$), θ_0 takes values in the interval $[-4.0,4.0]$ whereas, 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 that follow, 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}.

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

We do not report power results for the statistics S_{EL} and S_{ET} because, as seen earlier, their size properties appear to be quite poor for the sample sizes considered here. When $k = 1$, AR and K are numerically identical. In overidentified 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_ρ for $k = 1$ but leads to a less powerful test for $k > 1$. Also EL and ET versions of $GELR_\rho$ have rather unreliable size properties for the sample sizes considered here. Therefore we do not report detailed results for $GELR_\rho$.

We first focus on the separate effects of Π_1, n, ρ_{uV}, and k on power.

With strong identification all statistics have a U-shaped power curve. With the exception of $2SLS_{HET}$, the lowest point of the power curve is usually achieved at $\theta_0 = 0$. In Designs (I) and (II), $2SLS_{HET}$ is usually biased, taking on its lowest value at a negative θ_0 value in the interval $[-0.2,0.0]$. When θ_0 is weakly identified, the power curves of LM_{EL}, K, and LR_M are generally very flat across all θ_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 $\rho_{uV} = 0.99$, although being flat at about 5% for positive θ_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 $LM_{EL}(0) = GELR_{CUE}(0) = n\hat{g}(0)\hat{\Omega}(0)^{-1}\hat{g}(0)$. We have $\hat{\Omega}(0) \rightarrow_p E(u_i + Y_i\theta_0)^2$, which in Design (I) with $\Pi_1 = 0.1$ equals $1 + 2\theta_0\rho_{uV} + (1.01)\theta_0^2$. If $\rho_{uV} = 0.99$ this expression is minimized at around $\theta_0 = -0.98$ where it equals approximately 0.03. Therefore, this peak is caused by $\hat{\Omega}(0)^{-1}$ taking on large values for θ_0 in the neighborhood of -1.

For negative θ_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.

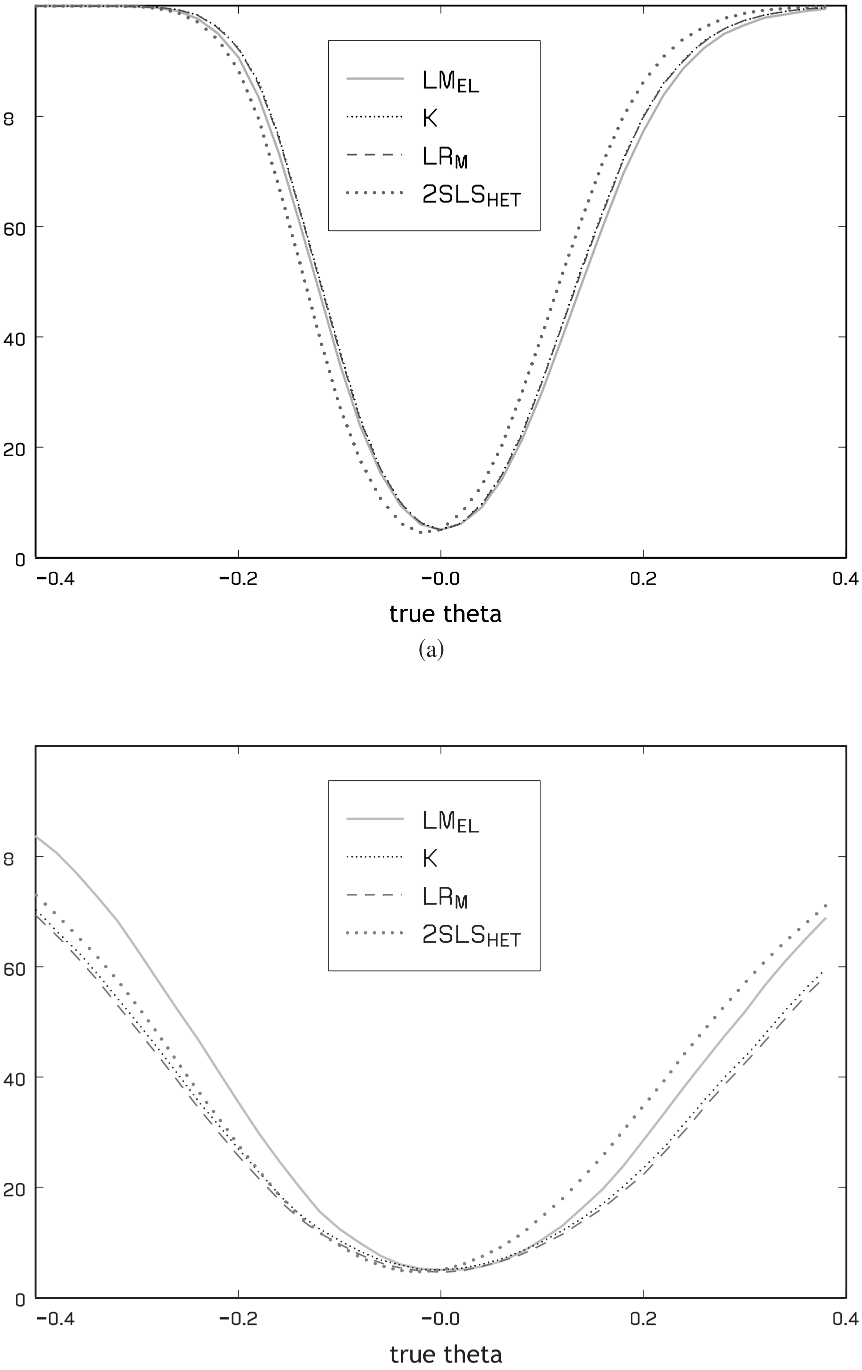
As to be expected the tests are more powerful when n is increased from 100 to 250. This holds uniformly across all statistics and designs with a more pronounced power increase in the strongly identified cases.

There does not seem to be a systematic effect due to ρ_{uV} as it varies with the specific design. For reasons explained previously, the shape of the power curves can change dramatically in Design (I) when ρ_{uV} 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 1a–c display the power curves of the four statistics for Designs (I)–(III) in the case

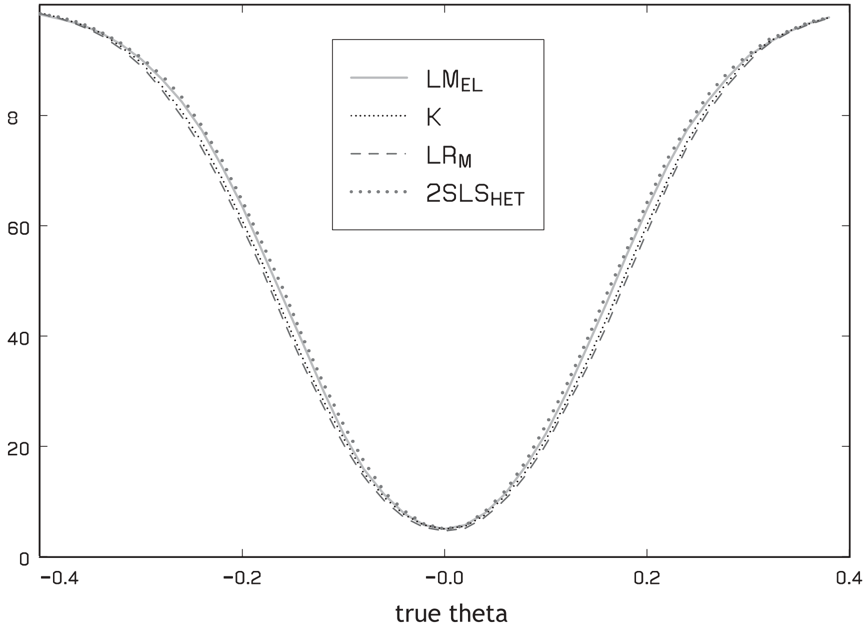
https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press



(a)

(b)

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press



(c)

FIGURE 1. Power curves, strong instrument. (a) Normal errors, (b) $t(2)$ errors, (c) χ^2 errors.

$\Pi_1 = 1.0$, $n = 250$, $\rho_{uV} = 0.5$, 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 θ_0 values but less so for positive θ_0. However, the Wald test $2SLS_{HET}$ is by far the most powerful test in all three designs. Except for some small negative θ_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 1b and 1c). However, LR_M and K uniformly dominate LM_{EL} in Design (I) (see Figure 1a). 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. Although 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.

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

With strong identification, the Wald test is the most powerful test for positive θ_0 values. For negative θ_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 whereas 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 previously the size properties of LR_M deteriorate substantially in the presence of heteroskedasticity.

NOTES

[1] Note that $\Delta(\theta)$ is $\Omega(\theta)$ in Stock and Wright (2000). We choose our notation for $\Omega(\theta)$ for consistency with Newey and Smith (2004).

[2] Weak convergence here is defined with respect to the sup-norm on function spaces and euclidean norm on R^k.

[3] For compact Θ, continuous ρ, and g_i $(i = 1,\dots,n)$, 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 θ, 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 $i = 1,\dots,n$ is positive. If $Y_{i1} = Y_{i2}$ for every $i = 1,\dots,n$ and $\hat{\theta} \in \Theta$ is an argmin of $\sup_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta,\lambda)$, then each $\bar{\theta} \in \Theta$ with $\bar{\theta}_1 + \bar{\theta}_2 = \hat{\theta}_1 + \hat{\theta}_2$ is also. 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 the 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.)

[4] A choice of $\hat{\Omega}(\theta)^{-1}$ as the weighting matrix $W_T(\bar{\theta}_T(\theta))$ in Stock and Wright (2000, equation (2.2), p. 1058), i.e., $(\sum_{i=1}^{n} g_i(\theta)g_i(\theta)'/n)^{-1}$, results in the CUE which is the GEL estimator based on $\rho(v) = -(1+v)^2/2$; see Newey and Smith (2004, Theorem 2.1). Hansen, Heaton, and Yaron (1996) and Pakes and Pollard (1989) define the (GMM) CUE using the centered weighting matrix $(\sum_{i=1}^{n}(g_i(\theta) - \hat{g}(\theta))(g_i(\theta) - \hat{g}(\theta))'/n)^{-1}$. However, as shown in Newey and Smith (2004, footnote 2), both versions of the CUE are numerically identical.

[5] The proof of Theorem 2 uses a second-order Taylor expansion of $\hat{P}_\rho(\theta,\lambda)$ in λ about 0 in which the only impact of ρ asymptotically is through ρ_1 and ρ_2, which are both -1.

[6] Newey and Smith (2004) 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.

[7] Alternatively, instead of using uniform weights in the definition of $\hat{\Omega}(\theta)$ one could use empirical probabilities that are associated with each GEL estimator; see Section 2 of Newey and Smith (2004). However, preliminary Monte Carlo simulations (not reported here) showed no clear improvement in the performance of the test statistics.

[8] To calculate $GELR_\rho(\theta)$, $S_\rho(\theta)$, and $LM_\rho(\theta)$ for EL and ET, the globally concave maximization problem $\max_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta,\lambda)$ must be solved numerically. To do so we implement a variant of the Newton–Raphson algorithm. We initialize the algorithm by setting λ equal to the zero vec-

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

tor. At each iteration the algorithm tries several shrinking step sizes in the search direction and accepts the first one that increases the function value compared to the previous value for λ. This procedure enforces an "uphill climbing" feature of the algorithm.

9. The statistics are defined as follows:

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

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

$$K(\theta) := (y - Y\theta)' P_{\tilde{Y}(\theta)} (y - Y\theta)/s_{uu}(\theta),$$

where $\tilde{Y}(\theta) := Z\tilde{\Pi}(\theta)$, $\tilde{\Pi}(\theta) = (Z'Z)^{-1} Z'[Y - (y - Y\theta) s_{uV}(\theta)/s_{uu}(\theta)]$, and $s_{uV}(\theta) := (y - Y\theta)' M_Z Y/(n - k)$. The statistic $K(\theta)$ (Kleibergen, 2002a), is not robust to conditional heteroskedasticity. However, a version of the K-statistic in Kleibergen (2001, equation (22)) that uses a heteroskedasticity consistent estimator for the covariance matrix of $g_i(\theta)$ overcomes this drawback. For model (5.1), the statistic is given by

$$K_{HET}(\theta) := n\hat{g}(\theta)' \bar{\Omega}(\theta)^{-1} D(\theta) (D(\theta)' \bar{\Omega}(\theta)^{-1} D(\theta))^{-1} D(\theta)' \bar{\Omega}(\theta)^{-1} \hat{g}(\theta),$$

where $\bar{\Omega}(\theta) := \hat{\Omega}(\theta) - \hat{g}(\theta)\hat{g}(\theta)'$, $D(\theta) := \sum_{i=1}^{n} G_i - n\hat{V}(\theta)\bar{\Omega}(\theta)^{-1}\hat{g}(\theta)$, $\hat{G} := \sum_{i=1}^{n} G_i/n$, and $\hat{V}(\theta) := \sum_{i=1}^{n} (G_i - \hat{G})(g_i(\theta) - \hat{g}(\theta))'/n$. The statistic $K_{HET}(\theta)$ is identical in structure to $LM_{CUE}(\theta)$ except the centered components $g_i(\theta) - \hat{g}(\theta)$ and $G_i - \hat{G}$ are used in place of $g_i(\theta)$ and G_i, respectively. Note that $G_i := G_i(\theta)$ does not depend on θ in a linear model. For the LR_M statistic, see Moreira (2003, Sect. 3). Finally, the Wald statistics are given by

$$2SLS_{HOM} := \hat{\theta}' W^{-1} \hat{\theta}, \qquad 2SLS_{HET} := \hat{\theta}' W_{HET}^{-1} \hat{\theta},$$

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

10. 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 $\rho_{uV} = 0.99$. 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).

To investigate the sensitivity of the results in Kleibergen (2002a) to the choice of Z, we iterated Kleibergen's (2002a) procedure 100 times; i.e., each time we simulated a matrix Z of instruments that we then kept fixed across $R = 1{,}000$ samples of the DGP (5.1). We found strong dependence of the numerical results of the Monte Carlo experiment on Z. For example, in the case $\Pi_1 = 1$, $k = 1$, the power of the K-statistic to reject the hypothesis $\theta_0 = 0$ when $\theta_0 = 0.4$ varied from about 60% to 95% in the 100 experiments. For the specific Z that Kleibergen (2002a) generates, he reports power of about 93% (see his Figure 1, p. 1793).

REFERENCES

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

Andrews, D.W.K. (1994) Empirical process methods in econometrics. In R. Engle & D. McFadden (eds.), *Handbook of Econometrics*, vol. 4, 2247–2294. North-Holland.

Brown, B.W. & W.K. Newey (1998) Efficient semiparametric estimation of expectations. *Econometrica* 66, 453–464.

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

Caner, M. (2003) Exponential Tilting with Weak Instruments: Estimation and Testing. Working paper, University of Pittsburgh.

Dufour, J. (1997) Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica* 65, 1365–1387.

Guggenberger, P. (2003) Econometric essays on generalized empirical likelihood, long-memory time series, and volatility. Ph.D. thesis, Yale University.

Guggenberger, P. & R.J. Smith (2003) Generalized Empirical Likelihood Tests in Time Series Models with Potential Identification Failure. Working paper, UCLA and University of Warwick.

Guggenberger, P. & M. Wolf (2004) Subsampling Tests of Parameter Hypotheses and Overidentifying Restrictions with Possible Failure of Identification. Working paper, UCLA.

Hansen, L.P. (1982) Large sample properties of generalized method of moment estimators. *Econometrica* 50, 1029–1054.

Hansen, L.P., J. Heaton, & A. Yaron (1996) Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics* 14, 262–280.

Imbens, G. (1997) One-step estimators for over-identified generalized method of moments models. *Review of Economic Studies* 64, 359–383.

Imbens, G. (2002) Generalized method of moments and empirical likelihood. *Journal of Business & Economic Statistics* 20, 493–506.

Imbens, G., R.H. Spady, & P. Johnson (1998) Information theoretic approaches to inference in moment condition models. *Econometrica* 66, 333–357.

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

Kitamura, Y. & M. Stutzer (1997) An information-theoretic alternative to generalized method of moments estimation. *Econometrica* 65, 861–874.

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

Kleibergen, F. (2002a) Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica* 70, 1781–1805.

Kleibergen, F. (2002b) Two Independent Pivotal Statistics That Test Location and Misspecification and Add-Up to the Anderson-Rubin Statistic. Working paper, Brown University.

Moreira, M.J. (2003) A conditional likelihood ratio test for structural models. *Econometrica* 71, 1027–1048.

Nelson, C.R. & R. Startz (1990) Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica* 58, 967–976.

Newey, W.K. (1985) Generalized method of moments specification testing. *Journal of Econometrics* 29, 229–256.

Newey, W.K. & R.J. Smith (2004) Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* 72, 219–255.

Newey, W.K. & K.D. West (1987) Hypothesis testing with efficient method of moments estimation. *International Economic Review* 28, 777–787.

Otsu, T. (2003) Generalized Empirical Likelihood Inference under Weak Identification. Working paper, University of Wisconsin.

Owen, A. (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75, 237–249.

Owen, A. (1990) Empirical likelihood ratio confidence regions. *Annals of Statistics* 18, 90–120.

Pakes, A. & D. Pollard (1989) Simulation and the asymptotics of optimization estimators. *Econometrica* 57, 1027–1057.

Phillips, P.C.B. (1984) The exact distribution of LIML: I. *International Economic Review* 25, 249–261.

Phillips, P.C.B. (1989) Partially identified econometric models. *Econometric Theory* 5, 181–240.

Qin, J. & J. Lawless (1994) Empirical likelihood and general estimating equations. *Annals of Statistics* 22, 300–325.

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

Smith, R.J. (1997) Alternative semi-parametric likelihood approaches to generalized method of moments estimation. *Economic Journal* 107, 503–519.

Smith, R.J. (2001) GEL Criteria for Moment Condition Models. Working paper, University of Bristol. Revised version CWP 19/04, cemmap, IFS and UCL. http://cemmap.ifs.org.uk/wps/cwp0419.pdf.

Staiger, D. & J.H. Stock (1997) Instrumental variables regression with weak instruments. *Econometrica* 65, 557–586.

Startz, R., E. Zivot, & C.R. Nelson (2004) Improved inference in weakly identified instrumental variables regression. Forthcoming in *Frontiers of Analysis and Applied Research: Essays in Honor of Peter C.B. Phillips*. Cambridge University Press.

Stock, J.H. & J.H. Wright (2000) GMM with weak identification. *Econometrica* 68, 1055–1096.

Stock, J.H., J.H. Wright, & M. Yogo (2002) A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics* 20, 518–529.

van der Vaart, A.W. & J.A. Wellner (1996) *Weak Convergence and Empirical Processes*. Springer.

Wooldridge, J. (2002) *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

APPENDIX: Proofs

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

$$\Pr\left\{\left(\max_{1 \leq i \leq n} f_i\right) n^{-1/\xi} > C^{1/\xi}\right\} \leq \sum_{i=1}^{n} \Pr\{f_i^{\xi} > nC\} \leq \sum_{i=1}^{n} \frac{1}{nC} E(f_i^{\xi}) \leq K/C < \varepsilon,$$

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 $(\max_{1 \leq i \leq n} f_i) n^{-1/\xi} = O_p(1)$ and thus $(\max_{1 \leq i \leq n} f_i) = o_p(n^{1/2})$ by $\xi > 2$. Thus (2.4) implies M(i). ■

Proof of Lemma 1. ID holds trivially. By (2.2) and (2.3), $g_i(\theta) = (y_i - Y_i'\theta)Z_i = Z_i(Z_i'\Pi + V_i')(\theta_0 - \theta) + Z_i u_i$. 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 $EW_i^2 < \infty$, $\max_{1 \leq i \leq n} |W_i| = o_p(n^{1/2})$; see Owen (1990, Lemma 3) for a proof. By the definition of $g_i(\theta)$ and the triangle inequality,

$$\max_{1 \leq i \leq n} \sup_{\theta \in \Theta} \|g_i(\theta)\| \leq \max_{1 \leq i \leq n} \sup_{\theta \in \Theta} (\|Z_i Z_i' \Pi(\theta_0 - \theta)\| + \|Z_i V_i'(\theta_0 - \theta)\| + \|Z_i u_i\|).$$

By Assumption M$'$(iii), we can apply the just-mentioned result to each of the three summands in the preceding equation, which proves the result.

Next M(ii) is shown. By the i.i.d. assumption, $\Omega(\theta) = \lim_{n \to \infty} Eg_i(\theta) g_i(\theta)'$, and continuity and boundedness in M(ii) follow immediately from M$'$(iii) and compactness of Θ. The same is true for the $O_p(1)$ statement in M(ii). Finally, uniform convergence follows from the weak law of large numbers and compactness of Θ.

Next M(iii) is proved. 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:

$$\sup_{\|\theta_1-\theta_2\|<\delta} \|\Psi_n(\cdot,\theta_1) - \Psi_n(\cdot,\theta_2)\|$$

$$= \sup_{\|\theta_1-\theta_2\|<\delta} \|(\beta_2-\beta_1)'n^{-1/2}\sum_{i=1}^{n} \Pi_B'(Z_iZ_i' - Q_{ZZ}) + (\theta_2-\theta_1)'n^{-1/2}\sum_{i=1}^{n} V_iZ_i'\|,$$

where the last expression is bounded by $\delta O_p(1)$ by the CLT. Furthermore, Θ 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 the Guggenberger (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 Newey and Smith (2004) for the proofs of our theorems. These modifications are necessary because unlike Newey and Smith 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 $c_n := n^{-1/2}\max_{1\le i\le n}\sup_{\theta\in\Theta_n}\|g_i(\theta)\|$. Let $\Lambda_n := \{\lambda \in R^k : \|\lambda\| \le n^{-1/2}c_n^{-1/2}\}$ if $c_n > 0$ and $\Lambda_n = R^k$ otherwise. Write "u.w.p.a.1" for "uniformly over $\theta \in \Theta_n$ w.p.a.1."

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

Then $\sup_{\theta\in\Theta_n,\lambda\in\Lambda_n,1\le i\le n}|\lambda' g_i(\theta)| \to_p 0$ *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

$$\sup_{\theta\in\Theta_n,\lambda\in\Lambda_n,1\le i\le n}|\lambda' g_i(\theta)| \le n^{-1/2}c_n^{-1/2}\max_{1\le i\le n}\sup_{\theta\in\Theta_n}\|g_i(\theta)\|$$

$$= n^{-1/2}c_n^{-1/2}n^{1/2}c_n = c_n^{1/2} = o_p(1),$$

which also immediately implies the second part. ■

LEMMA 8. *Suppose* $\max_{1\le i\le n}\sup_{\theta\in\Theta_n}\|g_i(\theta)\| = o_p(n^{1/2})$, $\lambda_{\min}(\hat\Omega(\theta)) \ge \varepsilon$ *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* ρ *holds.*

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. Without loss of generality $c_n \neq 0$, and thus Λ_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 ρ, $\hat P(\theta,\lambda)$ (as a function in λ for fixed θ) is C^2 on some open neighborhood of Λ_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 λ_θ^* on the line segment joining 0 and λ_θ such that for some positive constants C_1 and C_2

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

$$\le -2\lambda_\theta'\hat{g}(\theta) - C_1\lambda_\theta'\hat{\Omega}(\theta)\lambda_\theta \le 2\|\lambda_\theta\|\,\|\hat{g}(\theta)\| - C_2\|\lambda_\theta\|^2 \qquad \textbf{(A.1)}$$

u.w.p.a.1, where the second inequality follows as $\max_{1\le i\le n}\rho_2(\lambda_\theta^{*\prime}g_i(\theta)) < -\frac{1}{2}$ u.w.p.a.1 from Lemma 7, continuity of $\rho_2(\cdot)$ at zero, and $\rho_2 = -1$. The last inequality follows from $\lambda_{\min}(\hat{\Omega}(\theta)) \ge \varepsilon > 0$ u.w.p.a.1. Now, (A.1) implies that $(C_2/2)\|\lambda_\theta\| \le \|\hat{g}(\theta)\|$ 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 exist $M_\epsilon < \infty$ and $n_\epsilon \in \mathbb{N}$ such that $\Pr(\|n^{1/2}\lambda_\theta\| \le M_\epsilon) > 1 - \epsilon/2$ uniformly over $\theta \in \Theta_n$ and $\Pr(c_n^{-1/2} > M_\epsilon) > 1 - \epsilon/2$ for all $n \ge n_\epsilon$. Then $\Pr(\lambda_\theta \in \text{int}(\Lambda_n)) = \Pr(\|n^{1/2}\lambda_\theta\| < c_n^{-1/2}) \ge \Pr((\|n^{1/2}\lambda_\theta\| \le M_\epsilon) \wedge (c_n^{-1/2} > M_\epsilon)) > 1 - \epsilon$ for $n \ge n_\epsilon$ 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 λ for fixed θ) 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 before $\lambda_\theta = O_p(n^{-1/2})$ uniformly over $\theta \in \Theta_n$. Thus the second and by (A.1) the third parts of the lemma follow. ■

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

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

$$\hat{\theta}_{d_1} := (d_1', \hat{\theta}_2(d_1)')' \in R^p, \qquad \theta_{d_1} := (d_1', \theta_{02}')' \in R^p.$$

By u.w.p.a.1 we denote "uniformly over $d_1 \in \Theta_1$ w.p.a.1."

LEMMA 9. *Suppose* $\max_{1\le i\le n}\sup_{\theta\in\Theta_1\times\Theta_2}\|g_i(\theta)\| = o_p(n^{1/2})$, $\lambda_{\max}(\hat{\Omega}(\hat{\theta}_{d_1})) \le \kappa$ *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* ρ *holds.*

Then $\hat{g}(\hat{\theta}_{d_1}) = O_p(n^{-1/2})$ *uniformly over* $d_1 \in \Theta_1$.

Proof. Without loss of generality $\hat{g}(\hat{\theta}_{d_1}) \ne 0$ can be assumed. Define $\underline{\lambda} := -n^{-1/2}\hat{g}(\hat{\theta}_{d_1})/\|\hat{g}(\hat{\theta}_{d_1})\|$. Note that $\underline{\lambda} \in \Lambda_n$ and thus $\underline{\lambda} \in \hat{\Lambda}_n(\theta)$ 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 $\tilde{\lambda}$ on the line segment joining 0 and $\underline{\lambda}$ such that for some positive constants C_1 and C_2

$$\hat{P}(\hat{\theta}_{d_1},\underline{\lambda}) = -2\underline{\lambda}'\hat{g}(\hat{\theta}_{d_1}) + \underline{\lambda}'\left[\sum_{i=1}^{n}\rho_2(\tilde{\lambda}'g_i(\hat{\theta}_{d_1}))g_i(\hat{\theta}_{d_1})g_i(\hat{\theta}_{d_1})'/n\right]\underline{\lambda}$$

$$\ge 2n^{-1/2}\|\hat{g}(\hat{\theta}_{d_1})\| - C_1\underline{\lambda}'\left[\sum_{i=1}^{n}g_i(\hat{\theta}_{d_1})g_i(\hat{\theta}_{d_1})'/n\right]\underline{\lambda}$$

$$\ge 2n^{-1/2}\|\hat{g}(\hat{\theta}_{d_1})\| - C_2n^{-1} \qquad \textbf{(A.2)}$$

u.w.p.a.1, where the first inequality follows from Lemma 7, which implies that $\min_{1\le i\le n}$ $\rho_2(\tilde{\lambda}' g_i(\hat{\theta}_{d_1})) \ge -1.5$ u.w.p.a.1. The second inequality follows by $\lambda_{\max}(\hat{\Omega}(\hat{\theta}_{d_1})) \le \kappa < \infty$ u.w.p.a.1. The definition of $\hat{\theta}_{d_1}$ implies

$$\hat{P}(\hat{\theta}_{d_1}, \underline{\lambda}) \le \sup_{\lambda\in\hat{\Lambda}_n(\hat{\theta}_{d_1})} \hat{P}(\hat{\theta}_{d_1}, \lambda) \le \sup_{\lambda\in\hat{\Lambda}_n(\theta_{d_1})} \hat{P}(\theta_{d_1}, \lambda) = O_p(n^{-1}) \tag{A.3}$$

uniformly over $d_1 \in \Theta_1$. Combining equations (A.2) and (A.3) implies $n^{-1/2}\|\hat{g}(\hat{\theta}_{d_1})\| = O_p(n^{-1})$ uniformly over $d_1 \in \Theta_1$. ■

Proof of Theorem 2. (i) We first show consistency of $\hat{\beta}$. By Assumption ID and M(iii) $\sup_{\theta\in\Theta}\|\hat{g}(\theta) - (n^{-1/2}m_{1n}(\theta) + m_2(\beta))\| \to_p 0$, where $m_2(\beta) = 0$ if and only if $\beta = \beta_0$. Therefore, $\hat{g}(\hat{\theta}) = o_p(1)$ 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 $\lambda_{\max}(\hat{\Omega}(\hat{\theta})) \le \kappa$ w.p.a.1 for some $\kappa < \infty$, and thus Lemma 9 (applied to the case $p_1 = 0$, $\Theta_2 = \Theta$) implies $\hat{g}(\hat{\theta}) = O_p(n^{-1/2})$.

Next we establish $n^{1/2}$-consistency of $\hat{\beta}$. By consistency of $\hat{\beta}$ and Assumption M(ii) $\lambda_{\min}(\hat{\Omega}(\hat{\theta})) \ge \varepsilon$ w.p.a.1 for some $\varepsilon > 0$, and thus Lemma 8 for the case $\Theta_n = \{\hat{\theta}\}$ 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 λ around 0, there exists a mean value $\tilde{\lambda}$ between 0 and $\hat{\lambda}$ (that may be different for each row) such that

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

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

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

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

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

The same argument as for $\hat{\Omega}_{\tilde{\lambda}\hat{\theta}}$ proves $\hat{\Omega}_{\lambda^*\hat{\theta}} \to_p \Omega((\hat{\alpha}', \beta_0')')$. We therefore have $\hat{P}(\hat{\theta}, \hat{\lambda}) = \hat{g}(\hat{\theta})'(\Omega((\hat{\alpha}', \beta_0')')^{-1} + o_p(1))\hat{g}(\hat{\theta})$. By the definition of $\hat{\theta}$,

$$\begin{aligned} & n\hat{P}(\hat{\theta}, \hat{\lambda}) - n\hat{P}(\theta_0, \lambda(\theta_0)) \\ &\quad = n^{1/2}\hat{g}(\hat{\theta})'(\Omega((\hat{\alpha}', \beta_0')')^{-1} + o_p(1))n^{1/2}\hat{g}(\hat{\theta}) - n^{1/2}\hat{g}(\theta_0)'(\Omega(\theta_0)^{-1} \\ &\qquad + o_p(1))n^{1/2}\hat{g}(\theta_0) \le 0. \end{aligned}$$

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

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 Stock and Wright (2000, p. 1091, line six from the top) can now be applied to prove $n^{1/2}$-consistency of $\hat{\beta}$, where the symmetric matrix $\Omega((\hat{\alpha}',\beta_0')')^{-1} + o_p(1)$ plays the role of $W_T(\bar{\theta}_T(\hat{\theta}))$ in Stock and Wright. Note that in equation (A.4) in Stock and Wright, Assumption M(iii) of bounded sample paths w.p.a.1 is used. Finally, note that $\lambda_{\min}(\Omega((\hat{\alpha}',\beta_0')')^{-1} + o_p(1))$ is bounded away from zero w.p.a.1.

(ii) By Assumption M(iii)

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

and by ID we have for some mean-vector $\bar{\beta}$ between β_0 and $\beta_0 + n^{-1/2}b$ (that may differ across rows)

$$n^{1/2}E\hat{g}(\theta_{\alpha b}) = m_{1n}(\theta_{\alpha b}) + n^{1/2}m_2(\beta_0 + n^{-1/2}b) = m_{1n}(\theta_{\alpha b}) + M_2(\bar{\beta})b.$$

Because the latter expression is bounded, it follows that $\hat{g}(\theta_{\alpha b}) = O_p(n^{-1/2})$ u.w.p.a.1, where u.w.p.a.1 stands for "uniformly over $(\alpha,b) \in A \times B_M$ w.p.a.1." Therefore, by Lemma 8, $\lambda(\theta_{\alpha b})$ such that $\hat{P}(\theta_{\alpha b},\lambda(\theta_{\alpha b})) = \sup_{\lambda\in\hat{\Lambda}_n(\theta_{\alpha b})} \hat{P}(\theta_{\alpha b},\lambda)$ exists u.w.p.a.1 and $\lambda(\theta_{\alpha b}) = O_p(n^{-1/2})$ uniformly over $(\alpha,b) \in A \times B_M$. This implies that the FOC $n^{-1}\sum_{i=1}^{n} \rho_1(\lambda' g_i(\theta))g_i(\theta) = 0$ have to hold at $\lambda = \lambda(\theta_{\alpha b})$ and $\theta = \theta_{\alpha b}$ u.w.p.a.1. Expanding the FOC and using the same steps and notation as in part (i), it follows that $\lambda(\theta_{\alpha b}) = -(\hat{\Omega}_{\tilde{\lambda}\theta_{\alpha b}})^{-1}\hat{g}(\theta_{\alpha b})$, and upon inserting this into a second-order Taylor expansion of $\hat{P}(\theta,\lambda)$ we have

$$\hat{P}(\theta_{\alpha b},\lambda(\theta_{\alpha b})) = 2\hat{g}(\theta_{\alpha b})'\hat{\Omega}_{\tilde{\lambda}\theta_{\alpha b}}^{-1}\hat{g}(\theta_{\alpha b}) - \hat{g}(\theta_{\alpha b})'\hat{\Omega}_{\tilde{\lambda}\theta_{\alpha b}}^{-1}\hat{\Omega}_{\lambda^*\theta_{\alpha b}}\hat{\Omega}_{\tilde{\lambda}\theta_{\alpha b}}^{-1}\hat{g}(\theta_{\alpha b})$$

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

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

For given $\alpha \in A$, one can calculate $\arg\min_{b\in R^{p_B}} P_{\alpha b}$ by solving the FOC for b. Writing Ω 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 $\alpha^* = \arg\min_{\alpha\in A} P_{\alpha b^*(\alpha)}$. ■

Proof of Theorem 3. Applying Lemma 8 to the case $\Theta_n = \{\theta\}$, it follows that $\lambda(\theta) \in \hat{\Lambda}_n(\theta)$ exists such that $\hat{P}(\theta,\lambda(\theta)) = \sup_{\lambda\in\hat{\Lambda}_n(\theta)} \hat{P}(\theta,\lambda)$. Using the same steps and notation as in the proof of Theorem 2 leads to

$$\hat{P}(\theta,\lambda(\theta)) = 2\hat{g}(\theta)'\hat{\Omega}_{\tilde{\lambda}\theta}^{-1}\hat{g}(\theta) - \hat{g}(\theta)'\hat{\Omega}_{\tilde{\lambda}\theta}^{-1}\hat{\Omega}_{\lambda^*\theta}\hat{\Omega}_{\tilde{\lambda}\theta}^{-1}\hat{g}(\theta)$$

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

w.p.a.1, where by M_θ(ii) both $\hat\Omega_{\tilde\lambda\theta}$ and $\hat\Omega_{\lambda^*\theta}$ converge in probability to $\Delta(\theta)$. By M_θ(iii),

$$n^{1/2}\hat g(\theta) \to_d N(m_1(\theta), \Delta(\theta)),$$

from which the result follows. ■

Proof of Theorem 4. Using M_θ(i)–(iii) and an argument similar to the 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 equation (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 θ; 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 ξ_i between 0 and $\lambda' g_i$, and thus by (A.8) and the definition of Λ we have

$$\begin{aligned} D^* &= -n^{-1}\sum_{i=1}^{n}(n^{1/2}G_{iA}, G_{iB}) - n^{-3/2}\sum_{i=1}^{n}[\rho_2(\xi_i)(n^{1/2}G_{iA}, G_{iB})g_i'\Delta^{-1}n^{1/2}\hat g] + o_p(1) \\ &= -\left(n^{-1/2}\sum_{i=1}^{n}G_{iA} - n^{-1}\sum_{i=1}^{n}G_{iA}\,g_i'\Delta^{-1}n^{1/2}\hat g, M_2(\beta_0)\right) + o_p(1), \end{aligned}$$

where for the last equality we use (3.7) and Assumptions M_θ(v)–(vi). By Assumption M_θ(v) it thus follows that

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

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

$$M := \begin{pmatrix} -I_{kp_A} & \Delta_A\Delta^{-1} \\ 0 & 0 \\ 0 & I_k \end{pmatrix}, \qquad v := n^{-1/2}\sum_{i=1}^{n}\begin{pmatrix} \text{vec}G_{iA} \\ g_i \end{pmatrix};$$

M and v have dimensions $(kp_A + kp_B + k) \times (kp_A + k)$ and $(kp_A + k) \times 1$, respectively. By Assumption ID, M_θ(vii), and (3.7) $v \to_d N(w_2, V(\theta))$, where $w_2 := ((\text{vec}\,M_{1A})', m_1')'$ and M_{1A} are the first p_A columns of M_1. Therefore

$$\text{vec}(D^*, n^{1/2}\hat g) \to_d N\left(w_1 + Mw_2, \begin{pmatrix} \Psi & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Delta \end{pmatrix}\right), \tag{A.9}$$

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

We now derive the asymptotic distribution of $LM_\rho(\theta)$. Denote by $\bar D$ and $\bar g$ the limiting normal distributions of D^* and $n^{1/2}\hat g$, respectively (see equation (A.9)). Subsequently we show that the function $h: R^{k\times p} \to R^{p\times k}$ defined by $h(D) := (D'\Delta^{-1}D)^{-1/2}D'$ for

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

$D \in R^{k\times p}$ is continuous on a set $C \subset R^{k\times p}$ with $\Pr(\bar{D} \in C) = 1$. By the continuous mapping theorem and $\mathrm{M}_\theta(\mathrm{v})$ we have

$$(D^{*\prime}\tilde{\Omega}^{-1}D^*)^{-1/2}D^{*\prime}\tilde{\Omega}^{-1}n^{1/2}\hat{g} \to_d (\bar{D}'\Delta^{-1}\bar{D})^{-1/2}\bar{D}'\Delta^{-1}\bar{g}. \qquad \textbf{(A.10)}$$

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

$$W = W(\alpha) := (\bar{D}'\Delta^{-1}\bar{D})^{-1/2}\bar{D}'\Delta^{-1}m_1(\theta), \qquad \textbf{(A.11)}$$

$\zeta \sim N(0, I_p)$, and W and ζ 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 $\bar{D}$ has full column rank a.s. From (A.9) it follows that the last p_B columns of $\bar{D}$ equal $-M_2(\beta_0)$, which has full column rank by assumption. Define $O := \{o \in R^{kp_A} : \exists \tilde{o} \in R^{k\times p_A}$, s.t. $o = \mathrm{vec}(\tilde{o})$ and the $k \times p$-matrix $(\tilde{o}, -M_2(\beta_0))$ 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 $\bar{D}$, $\bar{D}_{p_A}$ say, it has been shown that $\mathrm{vec}\bar{D}_{p_A}$ is normally distributed with full rank covariance matrix Ψ. This implies that for any measurable set $O^* \subset R^{kp_A}$ with Lebesgue measure 0, it holds that $\Pr(\mathrm{vec}(\bar{D}_{p_A}) \in O^*) = 0$, in particular, for $O^* = O$. This proves the continuity claim for h. ■

Proof of Theorem 5. By Assumptions $\mathrm{M}_\alpha(\mathrm{v})$ and ID_α $\hat{g}(\hat{\theta}_a) = m_2((\alpha_{02}, \hat{\beta})) + o_p(1)$, and by Lemmas 8 and 9 (applied to $\Theta_n = \{\theta_{a\beta_0}\}$ and $\Theta_1 = \{a\}$, $\Theta_2 = B$, respectively) we have $\hat{g}(\hat{\theta}_a) = O_p(n^{-1/2})$. Assumption ID_α then implies consistency of $\hat{\beta}$. Applying Lemma 8 to the case $\Theta_n = \{\hat{\theta}_a\}$ implies that the FOC for λ must hold in the definition of $\hat{\theta}_a$ (see equation (A.4)). Then repeating the analysis that leads to (A.6) in the proof of Theorem 2, we have by $\mathrm{M}_\alpha(\mathrm{ii})$

$$GELR_\rho^{sub}(a) = n^{1/2}\hat{g}(\hat{\theta}_a)'\Delta(\theta_{a\beta_0})^{-1}n^{1/2}\hat{g}(\hat{\theta}_a) + o_p(1). \qquad \textbf{(A.12)}$$

The next goal is to derive the asymptotic distribution of $n^{1/2}\hat{g}(\hat{\theta}_a)$. Our analysis follows Newey and Smith (2004); see their proof of Theorem 3.2. Differentiating the FOC (A.4) with respect to λ yields the matrix $n^{-1}\sum_{i=1}^n \rho_2(\hat{\lambda}'g_i(\hat{\theta}_a))\; g_i(\hat{\theta}_a)g_i(\hat{\theta}_a)'$, which by $\mathrm{M}_\alpha(\mathrm{ii})$ 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

$$n^{-1}\sum_{i=1}^n \rho_1(\hat{\lambda}'g_i(\hat{\theta}_a))(\partial g_i/\partial\beta)'(\hat{\theta}_a)\hat{\lambda} = 0 \qquad \textbf{(A.13)}$$

w.p.a.1. Also, a mean-value expansion of (A.4) in (β, λ) about $(\beta_0, 0)$ yields (where $g_i(\theta)$ inside ρ_1 is kept constant at $g_i(\hat{\theta}_a)$)

$$\begin{aligned}&-\hat{g}(\theta_{a\beta_0}) + n^{-1}\sum_{i=1}^n [\rho_1(\bar{\lambda}'g_i(\hat{\theta}_a))G_{iB}(\theta_{a\bar{\beta}})(\hat{\beta} - \beta_0)\\ &\quad + \rho_2(\bar{\lambda}'g_i(\hat{\theta}_a))g_i(\theta_{a\bar{\beta}})g_i(\hat{\theta}_a)'\hat{\lambda}] = 0, \qquad \textbf{(A.14)}\end{aligned}$$

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press

where $(\bar{\beta}',\bar{\lambda}')$ are mean values on the line segment that joins $(\beta_0',0')$ and $(\hat{\beta}',\hat{\lambda}')$ that may be different for each row. Combining the p_B rows of (A.13) with the k rows of (A.14) we get

$$\begin{pmatrix} 0 \\ -\hat{g}(\theta_{a\beta_0}) \end{pmatrix} + M \begin{pmatrix} \hat{\beta}-\beta_0 \\ \hat{\lambda} \end{pmatrix} = 0, \qquad \textbf{(A.15)}$$

where the $(p_B+k)\times(p_B+k)$ matrix M has been implicitly defined. By M_α(ii) and M_α(iv)–(vi) the matrix M converges in probability to $\overline{M}$, where (writing $M_{2\beta}$ for $M_{2\beta}((\alpha_{02},\beta_0))$)

$$\overline{M} := -\begin{pmatrix} 0 & M_{2\beta}' \\ M_{2\beta} & \Delta(\theta_{a\beta_0}) \end{pmatrix} \quad \text{and} \quad \overline{M}^{-1} = -\begin{pmatrix} -\Sigma & H \\ H' & P \end{pmatrix}$$

and (omitting the argument $\theta_{a\beta_0}$)

$$\Sigma := (M_{2\beta}'\Delta^{-1}M_{2\beta})^{-1}, \qquad H := \Sigma M_{2\beta}'\Delta^{-1}, \qquad \text{and} \quad P := \Delta^{-1} - \Delta^{-1}M_{2\beta}\Sigma M_{2\beta}'\Delta^{-1}.$$

It follows that M is nonsingular w.p.a.1. Equation (A.15) implies that w.p.a.1

$$n^{1/2}((\hat{\beta}-\beta_0)',\hat{\lambda}')' = M^{-1}(0',n^{1/2}\hat{g}(\theta_{a\beta_0})')'. \qquad \textbf{(A.16)}$$

An expansion of $\hat{g}(\hat{\theta}_a)$ in β around β_0 and the preceding lead to

$$\hat{g}(\hat{\theta}_a) = \hat{g}(\theta_{a\beta_0}) + \hat{G}_B(\bar{\theta})(\hat{\beta}-\beta_0) = (I_k - M_{2\beta}H)\hat{g}(\theta_{a\beta_0}) + o_p(n^{-1/2}) \qquad \textbf{(A.17)}$$

for some appropriate mean value $\bar{\theta}$. Note that

$$I_k - M_{2\beta}H = M_{M_{2\beta}}(\Delta(\theta_{a\beta_0})), \qquad \textbf{(A.18)}$$

which has rank $k-p_B$. From (A.12), $GELR_\rho^{sub}(a) \to_d \xi'\Delta(\theta_{a\beta_0})^{-1}M_{M_{2\beta}}(\Delta(\theta_{a\beta_0}))\xi$, 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, $n^{1/2}\lambda(\hat{\theta}_a) = -\Delta(\theta_{a\beta_0})^{-1}n^{1/2}\hat{g}(\hat{\theta}_a) + o_p(1)$. Hence, the result for $LM_\rho^{sub}(a)$ implies the result for $S_\rho^{sub}(a)$.

As in the proof of Theorem 4 renormalize $D^* := D_\rho(a)\Lambda$, where the diagonal $p_A \times p_A$ matrix $\Lambda := \mathrm{diag}(n^{1/2},\dots,n^{1/2},1,\dots,1)$ has first p_{A_1} diagonal elements equal to $n^{1/2}$ and the remaining p_{A_2} elements equal to unity. We now show that D^* and $n^{1/2}\hat{g}(\hat{\theta}_a)$ are asymptotically independent. By a mean-value expansion about $\theta_{a\beta_0}$ and Assumption M_α(vii) we have for some mean value $\tilde{\theta} = (a',\tilde{\beta}')'$ (that may be different for each row)

$$\begin{aligned} n^{1/2}\,\mathrm{vec}\,\hat{G}_{A_1}(\hat{\theta}_a) &= n^{1/2}\,\mathrm{vec}\,\hat{G}_{A_1}(\theta_{a\beta_0}) + (\partial\,\mathrm{vec}\,\hat{G}_{A_1}/\partial\beta)(\tilde{\theta})n^{1/2}(\hat{\beta}-\beta_0) \\ &= n^{1/2}\,\mathrm{vec}\,\hat{G}_{A_1}(\theta_{a\beta_0}) - (\partial\,\mathrm{vec}\,\hat{G}_{A_1}/\partial\beta)(\tilde{\theta})Hn^{1/2}\hat{g}(\theta_{a\beta_0}) + o_p(1), \end{aligned}$$

where we have used (A.16) for the last equation. Assumptions M_α(vii) and ID_α imply $(\partial\,\mathrm{vec}\,\hat{G}_{A_1}/\partial\beta)(\tilde{\theta}) = \partial(n^{-1/2}m_1(\tilde{\theta}) + m_2((\alpha_{02},\tilde{\beta})))/\partial\beta\partial\alpha_1 + o_p(1) \to_p 0$ (recall that m_2 does not depend on α_1) and thus

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

Proceeding exactly as in the proof of Theorem 4, using (A.17), (A.19), and Assumptions M_α(vii)–(ix), it follows that

$$\operatorname{vec}(D^*, n^{1/2}\hat{g}(\hat{\theta}_a)) = m + Mv + o_p(1), \qquad \textbf{(A.20)}$$

where $M \in R^{(kp_{A_1}+kp_{A_2}+k)\times(kp_{A_1}+k)}$ and

$$M := \begin{pmatrix} -I_{kp_{A_1}} & \Delta_{A_1}\Delta^{-1} \\ 0 & 0 \\ 0 & I_k \end{pmatrix}\begin{pmatrix} I_{kp_{A_1}} & 0 \\ 0 & I_k - M_{2\beta}H \end{pmatrix},$$

$$v := n^{-1/2}\sum_{i=1}^{n}\begin{pmatrix} \operatorname{vec}G_{iA_1}(\theta_{a\beta_0}) \\ g_i(\theta_{a\beta_0}) \end{pmatrix}, \qquad m := \operatorname{vec}(0,-(\partial m_2/\partial\alpha_2),0),$$

where the arguments (α_{02},β_0) in $M_{2\beta}$ and $(\partial m_2/\partial\alpha_2)$ and $\theta_{a\beta_0}$ in Δ_{A_1} and Δ are omitted. By M_α(x), v is asymptotically normal with full rank covariance matrix $V^\alpha(\theta_{a\beta_0})$, and thus the asymptotic covariance matrix of $\operatorname{vec}(D^*, n^{1/2}\hat{g}(\hat{\theta}_a))$ is given by $MV^\alpha(\theta_{a\beta_0})M'$. For independence of D^* and $n^{1/2}\hat{g}(\hat{\theta}_a)$ the upper right $k(p_{A_1}+p_{A_2})\times k$-submatrix of $MV^\alpha(\theta_{a\beta_0})M'$ 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

$$(-\Delta_{A_1} + \Delta_{A_1}\Delta^{-1}(I_k - M_{2\beta}H)\Delta)(I_k - M_{2\beta}H)' \qquad \textbf{(A.21)}$$

is 0. Using (A.18), the matrix in (A.21) equals $-\Delta_{A_1}\Delta^{-1}P_{M_{2\beta}}(\Delta)M_{M_{2\beta}}(\Delta)\Delta$, which is clearly 0. This proves the independence claim.

Now denote by $\bar{D}$ and $\bar{g}$ the limiting normal distributions of D^* and $n^{1/2}\hat{g}(\hat{\theta}_a)$, implied by (A.20). Recall $M(a) = \Delta^{-1}M_{M_{2\beta}}(\Delta)$ (see equation (4.2)). If the function $h: R^{k\times p_A} \to R^{p_A\times k}$ 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 $\Pr(\bar{D} \in C) = 1$, then by the continuous mapping theorem

$$(D^{*\prime}M(a)D^*)^{-1/2}D^{*\prime}\Delta^{-1}n^{1/2}\hat{g}(\hat{\theta}_a) \to_d (\bar{D}'M(a)\bar{D})^{-1/2}\bar{D}'\Delta^{-1}\bar{g}.$$

By (A.17) and (A.18) the latter variable is distributed as $W_\alpha(a) + \zeta_\alpha$, where

$$W_\alpha(a) := (\bar{D}'M(a)\bar{D})^{-1/2}\bar{D}'M(a)m_1(\theta_{a\beta_0}). \qquad \textbf{(A.22)}$$

Therefore the theorem is proved 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_α(x). It is enough to show that with probability 1, $\operatorname{rank}(M_{M_{2\beta}}(\Delta)\bar{D}) = p_A$. Because the span of the columns of $M_{2\beta}$ equals the kernel of $M_{M_{2\beta}}(\Delta)$ and $\operatorname{rank}(M_{2\beta}) = p_B$, the latter condition holds if $\operatorname{rank}(M_{2\beta},\bar{D}) = p$. Denote by $\bar{D}_{p_{A_2}}$ the last p_{A_2} columns of $\bar{D}$, which by (A.20) equal $-(\partial m_2/\partial\alpha_2)$. By Assumption ID_α, the matrix $(\partial m_2/\partial(\alpha_2',\beta')')((\alpha_{02},\beta_0))$ 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 $\bar{D}$, $\bar{D}_{p_{A_1}}$ say. Using (A.20) and $V^\alpha(\theta_{a\beta_0}) > 0$, the covariance matrix of $\operatorname{vec}\bar{D}_{p_{A_1}}$ 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. ■

https://doi.org/10.1017/S0266466605050371 Published online by Cambridge University Press