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

Patrik Guggenberger, Yale University

June 2002, this version October 2002
Preliminary and Incomplete

Abstract

The actual finite sample size of many popular structural coefficient tests depends on the strength of identification. For example, even though the likelihood ratio and Wald test statistics are asymptotically χ^2, employing a χ^2 critical value can lead to extreme size-distortions in finite samples in weakly identified situations.

In this paper, I therefore propose new types of test statistics whose actual sizes are independent of the strength of identification. This is made rigorous by showing that the asymptotic null distribution of these statistics is χ^2 under partial (Phillips (1989)), weak (Staiger and Stock (1997)), and strong identification.

The first statistic A_ρ is constructed from the criterion function of the Generalized Empirical Likelihood (GEL) estimator. The second statistic K_ρ^W generalizes the K statistic in Kleibergen (2001) from Generalized Method of Moments (GMM) to GEL. This statistic is given by a quadratic form of the first-order condition of the GEL estimator evaluated at the true parameter value. I show how K_ρ^W can be modified to test hypothesis involving a subvector of the structural parameter vector.

A Monte Carlo study reveals that the new tests have very competitive size and power properties under both weak and strong identification. Their main advantage lies in their robustness to certain features of the error distribution like asymmetry or thick tails. In over-identified problems, the statistic K_ρ^W is generally more powerful than A_ρ.

Finally, I derive the asymptotic distribution of the GEL estimator for the structural parameters in the linear model under weak identification. Similar to the findings of Stock and Wright (2000) for the GMM estimator, the resulting estimators have non-standard asymptotic distributions and are in general inconsistent.

Keywords:

JEL Classification Numbers:

1 Introduction

.....All the proofs are given in the Appendix.

By " $\rightarrow_d$ ", " $\rightarrow_p$ ", and " $\Rightarrow$ " I 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" I write "a.s.". For "with probability approaching 1" I write "wpa1". For a symmetric matrix A, "$A > 0$" means that A is positive definite. For a full rank matrix A, I denote by P_A the projection matrix on the column space of A, $A(A'A)^{-1}A'$, and define $M_A := I - P_A$. By "$\otimes$" I denote the Kronecker product. Finally $vec(M)$ stands for the column vectorization of the $k \times p$ matrix M, i.e. if $M = (m_1, ..., m_p)$ then $vec(M) = (m_1'...m_p')'$.

2 General Empirical Likelihood and Weak Identification

In this section, I derive the asymptotic distribution of the GEL estimator under weak asymptotics in the linear model. I then propose several test statistics for simple hypothesis involving the structural parameters, that are asymptotically similar under classical and local to zero asymptotics. I extend the statistics to tests for subvectors of the structural parameter vector and also derive their asymptotic distribution when some parameters are weakly and some are strongly identified.

2.1 The model and assumptions

I consider the following linear model where the structural equation is given by

$$y = Y\theta_0 + X\gamma_0 + u, \tag{2.1}$$

and the reduced form by

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

where $y, u \in R^n$, $Y, V \in R^{n\times p}$, $X \in R^{n\times q}$, $Z \in R^{n\times k}$, $\Pi \in R^{k\times p}$, $\gamma_0 \in R^q$, and $\theta_0 \in R^p$. The variables Z constitute a set of instruments for the endogenous variables Y. The variables X are assumed to be exogenous. Interest focuses on inference on the vector θ_0.

From now on, I assume wlog that $\gamma_0 = 0$. For, if $\gamma_0 \neq 0$, one can multiply equation (2.1) by the $n \times n$ matrix M_X, i.e. project all variables on the space orthogonal to the range space of X, and then work with the model

$$y^* = Y^*\theta_0 + u^*$$

instead, where $y^* := M_X y$, $Y^* := M_X Y$, and $u^* := M_X u$. I also assume $k \geq p$, the order condition for identification.

Under classical asymptotic theory, it is well known that in the above model the $2SLS$ estimator, $\widehat{\theta}_{2SLS} := (Y'P_Z Y)^{-1}Y'P_Z y$ is a consistent and asymptotically normal estimator for θ_0. However, there is strong theoretical and Monte Carlo evidence

that the asymptotic distribution can be a very poor approximation of the finite sample distribution, especially when the correlation between the instruments and the included endogenous variables is weak, see among many other references, Phillips (1989), Nelson and Startz (1990), and Staiger and Stock (1997). Therefore, alternative asymptotics have been proposed in the literature that better capture the finite sample behavior of the estimator when the correlation is weak. Phillips (1989) introduces the notion of partially identified models, in which only a subset of the structural parameters are identified. Staiger and Stock (1997) and Stock and Wright (2000) propose weakly identified models, where for each finite sample size, the model is formally identified, but where the correlation between the instruments and the endogenous variables fades away with n going to infinity. In these models, the $2SLS$ estimator is inconsistent and distributed asymptotically as a random mixture of normals. The asymptotic null distribution of Wald statistics, based on the $2SLS$ estimator and testing linear restrictions of the structural parameter vector, is in general not a χ^2 random variabel and depends on parameters that cannot be consistently estimated. In this paper, I propose alternative test statistics whose asymptotic null distribution is the same under strong and weak identification. The rigorous formulation of these notions is given in the next assumption.

Assumption 1: $\Pi = \Pi_n = n^{-\xi}C$, where C is a fixed $R^{k\times p}$ matrix. Either (i) or (ii) holds, where:

(i) "Classical or strong identification", $\xi = 0$ and C is of full column rank.

(ii)"Weak identification", $\xi = 1/2$.

Note that Assumption 1(ii) includes as a particular case, the completely unidentified model, in which $C = 0$. Below, I generalize Assumption 1 and allow simultaneously for weakly and strongly identified parameters. The generalized assumption then also includes the case of the partially identified model of Phillips (1989). For now, to simplify exposition, I assume that all parameters are either weakly or strongly identified.

Part (ii) of Assumption 1 is a now popular method of modeling weak correlation between the instruments Z and the included endogenous variables as n goes to infinity. Note that if Π_n was modeled as a fixed matrix independent of n, the mean of the F statistic testing $\Pi = 0$ would tend to infinity with n. Under Assumption 1 the mean is $O_p(1)$. (See Staiger and Stock (1997, p.560) for a more detailed motivation of the local to zero assumption.) ADD MORE MOTIVATION.

Define $U = (u, V)$. By $U_i, Z_i, ...$ for $i = 1, ..., n$ I denote the i^{th} row of the matrix $U, Z,...$ written as a column vector. By $U_{ij}, Z_{ij},...$ I denote the j^{th} components of the vectors $U_i, Z_i,....$

I make the following moment and distributional assumptions.

Assumption 2:

(i) $E(U_i|Z_i) = 0$ a.s.,

(ii) $E(U_iU_j'|Z_i, Z_j) = \Sigma_{UU}$ if $i = j$ and 0 otherwise a.s. (conditional homoskedasticity), where $\Sigma_{UU} > 0$,

(iii) $E||Z_i||^4 < \infty$, and $E(Z_iZ_i') = Q_{ZZ} > 0$,

(iv) Z_i is independent of Z_j and U_j for $i \neq j$,

(v) $\{U_i : i \geq 1\}$ are iid.

I decompose the matrix Σ_{UU} conformably with the dimensions of u and V. Write Σ_{UU} as $\begin{pmatrix} \sigma_{uu} & \Sigma'_{uV} \\ \Sigma_{uV} & \Sigma_{VV} \end{pmatrix}$, where σ_{uu}, Σ_{uV}, and Σ_{VV} are 1×1, $p \times 1$, and $p \times p$ matrices, respectively. If Y is endogenous in (2.1) then $\Sigma_{uV} \neq 0$. For given sample size n, define the random k-vector

$$g_{ni}(\theta) = (y_i - Y_i'\theta)Z_i.$$

I usually write $g_i(\theta)$ for $g_{ni}(\theta)$. By definition

$$g_i(\theta) = (u_i + (n^{-\xi}Z_i'C + V_i')(\theta_0 - \theta))Z_i. \tag{2.2}$$

Before defining the estimator of θ_0 I have to adequately restrict the parameter space.

Assumption 3: θ_0 is in the interior of the compact set $\Theta \subset R^p$.

The GEL estimator $\widehat{\theta}$ of θ_0 in (2.1) exploits the moment condition $Eg_i(\theta_0) = 0$. It is given by

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \sup_{\lambda \in \widehat{\Lambda}_n(\theta)} \widehat{P}(\theta, \lambda), ^1 \tag{2.3}$$

where

$$\widehat{P}(\theta, \lambda) := 2\sum_{i=1}^{n} \rho(\lambda' g_i(\theta))/n - 2\rho_0. \tag{2.4}$$

MENTION MINIMUM DISTANCE FORMULATION OF GEL. Here ρ is a real-valued function $Q \to R$, where Q is an open interval of the real line that contains 0, and $\widehat{\Lambda}_n(\theta) := \{\lambda \in R^k : \lambda' g_i(\theta) \in Q \text{ for } i = 1, ..., n\}$. If defined, let $\rho_j(\upsilon) := \partial^j \rho(\upsilon)/\partial \upsilon^j$ and $\rho_j := \rho_j(0)$ for nonnegative integers j.

Assumption 4:

(i) ρ is strictly concave on Q.

(ii) ρ is C^2 in some neighborhood of 0, and $\rho_1 = \rho_2 = -1$.

This definition is adopted from Newey and Smith (2001) (NS from now on). I slightly modify their definition of $\widehat{P}(\theta, \lambda)$ by recentering and rescaling because it simplifies the presentation. The most popular GEL estimators are the Continuous-Updating Estimator (CUE), Empirical Likelihood (EL), and Exponential Tilting (ET), corresponding to $\rho(\upsilon) = -(1 + \upsilon)^2/2$, $\rho(\upsilon) = \ln(1 - \upsilon)$, and $\rho(\upsilon) = -\exp \upsilon$. The EL and ET estimators were introduced by Owen (1988, 1990) and Kitamura and Stutzer (1997), respectively.

[1] For Θ compact, ρ, and each g_i continuous it can be shown that an argmin $\widehat{\theta}$ really exists. In fact, $\sup_{\lambda \in \widehat{\Lambda}_n(\theta)} \widehat{P}(\theta, \lambda)$, 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 $\widehat{\theta}$ however is not implied. As a simple example, in the case $p = 2$, let the two components Y_{ij} $(j = 1, 2)$ of Y_i be independent Bernoulli random variables. Then, for each n, it happens with positive probability that $Y_{i1} = Y_{i2}$, for all $i = 1, ..., n$. In that case, if $\widehat{\theta} \in \Theta$ is an argmin vector of $\sup_{\lambda \in \widehat{\Lambda}_n(\theta)} \widehat{P}(\theta, \lambda)$, then each $\overline{\theta} \in \Theta$ with $\overline{\theta}_1 + \overline{\theta}_2 = \widehat{\theta}_1 + \widehat{\theta}_2$ is too. To uniquely define $\widehat{\theta}$, we could, for example, do the following. From the set of all vectors $\theta \in \Theta$ that minimize $\sup_{\lambda \in \widehat{\Lambda}_n(\theta)} \widehat{P}(\theta, \lambda)$, let $\widehat{\theta}$ be the vector that has smallest first component (if that does not pin down $\widehat{\theta}$ uniquely, discriminate the remaining vectors by the second component, and so on).

2.2 Asymptotics for GEL estimators under Weak Identification

In the following I want to derive the asymptotic distribution of $\widehat{\theta}$ under weak identification, i.e. under Assumption 1(ii). It is instructive to examine a simple case first, namely the case where ρ is quadratic. In that case, $Q = R$, and a second order Taylor expansion in λ of $\widehat{P}(\theta, \lambda)$ about 0 is exact. The former implies that for each $\theta \in \Theta$ we have $\widehat{\Lambda}_n(\theta) = R^k$ and thus the maximization in λ is unconstrained. The latter implies that for

$$\widehat{g}(\theta) := \sum_{i=1}^{n} g_i(\theta)/n,$$

and

$$\widehat{\Omega}(\theta_1, \theta_2) := \sum_{i=1}^{n} g_i(\theta_1) g_i(\theta_2)'/n, \ \widehat{\Omega}(\theta) := \widehat{\Omega}(\theta, \theta)$$

we have

$$\widehat{P}(\theta, \lambda) = -2\widehat{g}(\theta)'\lambda - \lambda'\widehat{\Omega}(\theta)\lambda. \tag{2.5}$$

By concavity of $\widehat{P}(\theta, \lambda)$ in λ, any solution $\lambda(\theta)$ to the FOC $0 = -\widehat{g}(\theta) - \widehat{\Omega}(\theta)\lambda$ maximizes $\widehat{P}(\theta, \lambda)$ with respect to λ for fixed θ. If by $\widehat{\Omega}(\theta)^-$ we denote the Moore-Penrose inverse of $\widehat{\Omega}(\theta)$, then $\lambda(\theta) := -\widehat{\Omega}(\theta)^- \widehat{g}(\theta)$ solves the FOC. The GEL objective function for quadratic ρ is thus given by

$$\widehat{P}(\theta, \lambda(\theta)) = \widehat{g}(\theta)'\widehat{\Omega}(\theta)^- \widehat{g}(\theta). \tag{2.6}$$

The previous argument was used in NS to show that for quadratic ρ the GEL estimator formally resembles the GMM Continuous-Updating estimator defined in Hansen, Heaton, and Yaron (1996)[2]. Both estimators minimize a quadratic form whose weighting matrix is continuously altered as θ changes. However, in the latter case the weighting matrix is the inverse of a consistent estimate of the covariance matrix of $\widehat{g}(\theta)$ while in the former it is usually not. In general, only in the case $\theta = \theta_0$, the matrix $\widehat{\Omega}(\theta)$ consistenly estimates the covariance matrix of $\widehat{g}(\theta)$. Even though the two estimators are in general numerically different they are both referred to in the literature as the CUE estimator. In this paper I distinguish the two estimators by writing CUE and CUE_{GMM} for the GEL and GMM Continuous-Updating estimator, respectively.

I now derive the asymptotic distribution of $\arg\min \widehat{P}(\theta, \lambda(\theta))$ under local to zero asymptotics. Define the $k \times k$ matrix

$$\Omega(\theta_1, \theta_2) := \lim_{n \to \infty} E g_i(\theta_1) g_i(\theta_2)' \text{ and write } \Omega(\theta) \text{ for } \Omega(\theta, \theta). \tag{2.7}$$

The next lemma establishes the probability limit of $\widehat{\Omega}(\theta)$ under weak asymptotics.

Lemma 1 *Under Assumptions 1(ii), 2, and 3 the following results hold.*

[2]The Continuous-Updating Estimator appears already in Pakes and Pollard (1989), see their Lemma (3.5) and Theorem (3.3). However most of the literature cites Hansen, Heaton, and Yaron (1996) when referring to the Continuous-Updating Estimator.

(i) We have $\Omega(\theta_1,\theta_2) = (1,(\theta_0-\theta_1)')\Sigma_{UU}(1,(\theta_0-\theta_2)')'Q_{ZZ}$ *and* $\sup_{\theta_1,\theta_2\in\Theta}||\widehat{\Omega}(\theta_1,\theta_2)-\Omega(\theta_1,\theta_2)|| \to_p 0$, *in particular,* $\sup_{\theta\in\Theta}||\widehat{\Omega}(\theta)-\Omega(\theta)|| \to_p 0$.

(ii) Let $\Psi(\theta)$ *be a* k*-dimensional Gaussian empirical process on* Θ *with mean zero and covariance function* $E\Psi(\theta_1)\Psi(\theta_2)' = \Omega(\theta_1,\theta_2)$. *Then,* $n^{1/2}\widehat{g}(\theta) \Rightarrow \Psi(\theta) + Q_{ZZ}C(\theta_0-\theta)$.

By positive definiteness of Σ_{UU} and Q_{ZZ} it follows that $\Omega(\theta)$ is positive definite for all $\theta\in\Theta$. Therefore, by the above lemma $\widehat{\Omega}(\theta)$ converges uniformly to a uniformly positive definite matrix and therefore wpa1 $\widehat{\Omega}(\theta)$ is invertible. It follows that $n\widehat{P}(\theta,\lambda(\theta)) \Rightarrow P(\theta)$, where

$$\begin{aligned} P(\theta,\overline{\theta}) &:= [\Psi(\theta)+Q_{ZZ}C(\theta_0-\theta)]'\Omega(\overline{\theta})^{-1}[\Psi(\theta)+Q_{ZZ}C(\theta_0-\theta)], \qquad (2.8)\\ P(\theta) &:= P(\theta,\theta). \end{aligned}$$

Assuming that the process $P(\theta)$ has a unique minimum, it follows from Lemma 3.2.1 in van der Vaart and Wellner (1996, p.286) that

$$\widehat{\theta}_{CUE} \to_d \arg\min_{\theta\in\Theta} P(\theta).$$

The analogous result, in the more general setup, where the linear model can also contain a number of strongly identified parameters, has been shown in Stock and Wright (2000) for CUE_{GMM}. They consider GMM estimation with weak identification and then specialize their results to the linear model for which they work out the asymptotic distribution of two stage least squares and CUE_{GMM}.

I now deal with the asymptotic distribution of arbitrary GEL estimators. For nonquadratic ρ the analysis becomes much more complicated. I closely follow the proof in NS (see their Lemmas A1 and A2). The main step of the proof is to show that the $\sup_{\lambda\in\widehat{\Lambda}_n(\theta)}\widehat{P}(\theta,\lambda)$ in (2.3) is actually a maximum wpa1. Then the following definition is justified (at least wpa1):

$$\lambda(\theta) := \arg\max_{\lambda\in\widehat{\Lambda}_n(\theta)} \widehat{P}(\theta,\lambda). \qquad (2.9)$$

It then follows that the FOC for a maximum at $\lambda(\theta)$ has to hold and a second order Taylor expansion of the FOC (this time with Lagrange remainder term) establishes the desired result as before.

Theorem 2 *Under Assumptions 1(ii), and 2-4 we have* $n\widehat{P}(\theta,\lambda(\theta)) \Rightarrow P(\theta)$, *and assuming that* $P(\theta)$ *has a unique minimum, the GEL estimator satisfies* $\widehat{\theta}_{GEL} \to_d \arg\min_{\theta\in\Theta} P(\theta)$.

INTERPRET RESULT The theorem shows that with weak instruments GEL estimation is in general inconsistent. Also, the theorem implies that no matter which concave function ρ we use to define the GEL estimator, the asymptotic distribution of the structural coefficient estimates in a linear model with weak instruments is the same. In particular, there is no first-order difference between the three most

commonly used GEL estimators: CUE, ET, and EL. COMPARE RESULTS TO STAIGER AND STOCK RESULTS FOR OLS, 2SLS, LIML For a "strongly identified" model with a finite number of moment restrictions NS find also that different GEL estimators are first order equivalent. However, under strong identification, the GEL estimators are consistent and asymptotically normal.

Remark 1 *Assumption 2 implies that* $E[\sup_{\theta\in\Theta} ||g_i(\theta)||^2] < \infty$[3]. *In their paper on GEL with strong instruments, in their Assumption 1(d), NS assume that* $E[\sup_{\theta\in\Theta} ||g_i(\theta)||^s] < \infty$ *for some* $s > 2$. *They could weaken this assumption to* $s = 2$ *and still prove consistency and asymptotic normality of the GEL estimator (their Theorems 3.1 and 3.2) by modifying their proof along the lines of my proof of Theorem 2. It then follows that consistency and asymptotic normality of GEL in NS under Assumption 1(i) can be established under the same assumptions as for two-step efficient GMM, in Hansen (1982).*

2.3 Test statistics

Because the limiting distribution in Theorem 2 is nonstandard and involves quantities that cannot be consistently estimated, it can not be exploited in a straightforward manner to construct confidence intervals for θ_0. However, similar to Stock and Wright (2000, p.1063), confidence intervals for θ_0 and hypothesis tests of $\theta = \theta_0$ can be constructed directly from the objective function. This is a consequence of the following corollary to Theorem 2. Define

$$A_\rho(\theta) := n\widehat{P}(\theta, \lambda(\theta)). \tag{2.10}$$

For CUE, equation (2.6) shows that when $\theta \neq \theta_0$, we get $A_\rho(\theta) \to_d \chi^2(k)$ for $\xi > 1/2$, and $A_\rho(\theta)$ diverges to infinity at a rate of $n^{1-2\xi}$ for $\xi < 1/2$. I thus derive the asymptotic distribution of $A_\rho(\theta_0)$ for $\xi = 0$ under a local rather than a fixed alternative.

Corollary 3 *Suppose Assumptions 2 and 4, and fix* $\theta \in \Theta$.

(*i*) *Under Assumption 1(i), we have for any fixed* $d \in R^p$

$$A_\rho(\theta_0 + n^{-1/2}d) \to_d \chi^2(k, \delta),$$

where the noncentrality parameter δ *is given by*

$$\delta = \sigma_{uu}^{-1} ||Q_{ZZ}^{1/2} C d||^2.$$

(*ii*) *Under Assumption 1(ii) we have*

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

[3]To prove this note that $E(u_i^2 Z_i Z_i') = E(E(u_i^2|Z_i) Z_i Z_i') = \sigma_{uu} Q_{ZZ}$ and thus $E||u_i Z_i||^2 < \infty$. An analogous argument shows that $E||V_{ij} Z_i||^2 < \infty$ $(j = 1, ..., p)$. Also, $E|Z_{ij_1} Z_{ij_2}|^2 < \infty$ $(j_1, j_2 = 1, ..., k)$ and thus by compactness of Θ and by Minkowski's inequality the claim follows.

where the noncentrality parameter δ is given by

$$\delta = ||\Omega(\theta)^{-1/2} Q_{ZZ} C(\theta_0 - \theta)||^2.$$

In particular, independent of the value $\xi \in [0, \infty)$, we have $A_\rho(\theta_0) \to_d \chi^2(k)$.

MENTION TESTS OF IMBENS AND SPADY (2002).

The corollary provides a straightforward method for constructing confidence sets and perform hypothesis tests involving θ_0 that are asymptotically valid. For example, to test the hypothesis $\theta = \theta_0$ at significance level r, reject the hypothesis iff $A_\rho(\theta_0)$ exceeds $\chi^2_r(k)$, the $(1-r)$ $\chi^2(k)$ critical value. $(1-r)$ confidence regions for θ_0, obtained by inverting the just described test, are given by $\{\theta \in \Theta |\ A_\rho(\theta) \leq \chi^2_r(k)\}$. As Corollary 3 shows the power of the hypothesis test depends on $\delta = \delta((\theta_0 - \theta), C, Q_{ZZ}, \Omega(\theta))$. In general, one would expect the power to increase with $||\theta_0 - \theta||$ increasing (everything else remaining constant). Also, increasing $||C||$, i.e. working with stronger instruments, should increase the power of the test.

Corollary 3(ii) shows that under weak instruments the above hypothesis test for $\theta = \theta_0$ is inconsistent. The noncentrality parameter δ of the asymptotic χ^2 distribution under the alternative does not converge to infinity for increasing sample size, and therefore the rejection rate under the alternative does not converge to 1. However, since the asymptotic distribution under the null is $\chi^2(k)$ independent of ξ, the test has correct asymptotic size with weak and strong identification.

When EL methods were introduced in the late eighties by Owen, they were first used to construct confidence regions for means of *iid* random variables. Corollary 3(i) is a direct generalization to GEL of the well known EL result that -2 times the logarithm of the empirical likelihood ratio converges in distribution to a chi squared random variable (see Owen (1988) p.237, 238 or Owen (1990)).

For the CUE_{GMM} with weak instruments, Stock and Wright (2000, Theorem 2) derive the asymptotic distribution of the analogue to $A_\rho(\theta)$ at the true value θ_0.

A drawback of the type of test statistic derived from the result in Corollary 3 is that its limiting distribution has a degrees of freedom parameter equal to the number of instruments. In general this has a negative impact on the power properties of hypothesis tests in overidentified situations. Kleibergen (2001, 2002) introduced a statistic, called K statistic, for hypothesis tests in a GMM framework whose limiting distribution is chi-squared with degrees of freedom equal to the number of parameters to be estimated. The test statistic is given by a quadratic form of the derivative of the GMM objective function evaluated at the true value θ_0.

Applying Kleibergen's (2001) idea to GEL I construct a quadratic form from the GEL FOC condition for θ evaluated at the true value θ_0. If the minimum of the objective function $\widehat{P}(\theta, \lambda(\theta))$ is obtained in the interior of the parameter space Θ, the following FOC has to hold

$$\lambda(\theta)' \sum_{i=1}^{n} \rho_1(\lambda(\theta)' g_i(\theta)) G_i(\theta)/n = 0, \tag{2.11}$$

where the $k \times p$ matrix $G_i(\theta)$ is given by $\partial g_i(\theta)/\partial\theta$ and where $\lambda(\theta)$ is defined in (2.9) above (for a proof see NS, Section 2.2 or equation (4.11) below). For $\theta \in \Theta$ I define the $k \times p$ matrix

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

The expression on the left hand side of equation (2.11) can thus be written as $\lambda(\theta)' D_\rho(\theta)$.

Under Assumption 1(ii), for CUE we have seen above that $\lambda(\theta) = -\widehat{\Omega}(\theta)^- \widehat{g}(\theta)$ wpa1 and thus by Lemma 1(i) and (ii) $n^{1/2}\lambda(\theta_0) \to_d N(0, \Omega(\theta_0)^{-1})$. In the Appendix I show that the last statement holds for all GEL estimators. If $D_\rho(\theta_0)$ and $\lambda(\theta_0)$ were asymptotically independent we could premultiply the (appropriately normalized) statistic $D_\rho(\theta_0)' n^{1/2}\lambda(\theta_0)$ by the factor $(D_\rho(\theta)'\Omega(\theta)^{-1} D_\rho(\theta))^{-1/2}$ to get a limiting $N(0, I_p)$ distribution. From that statistic we could then construct a quadratic form with limiting $\chi^2(p)$ distribution. The Appendix provides a rigorous treatment of the above steps. The resulting statistic can be written compactly as

$$K_\rho^W(\theta) := n\widehat{g}(\theta)'\Omega(\theta)^{-1/2} P_{\Omega(\theta)^{-1/2} D_\rho(\theta)} \Omega(\theta)^{-1/2} \widehat{g}(\theta), \tag{2.12}$$

where ρ is any function satisfying Assumption 4. I also consider the following variant of $K_\rho^W(\theta)$ that does not substitute $\lambda(\theta)$ by $-\Omega(\theta)^{-1}\widehat{g}(\theta)$

$$K_\rho^L(\theta) := n\lambda(\theta)'\Omega(\theta)^{1/2} P_{\Omega(\theta)^{-1/2} D_\rho(\theta)} \Omega(\theta)^{1/2} \lambda(\theta). \tag{2.13}$$

I use the superscripts W and L for the two test statistics because they have an interpretation as Wald-type and LM-type (Lagrange-Multiplier-type) statistics, respectively.

The intuition for the test statistics is based on the classical case of strong identification, i.e. the case considered in Assumption 1(i). In that case, we know from NS that $\widehat{\theta}$ is $n^{1/2}$-consistent. Therefore, if the FOC (2.11) hold at $\widehat{\theta}$, then, at least asymptotically, they also hold at the true value θ_0. The statistic $K_\rho^W(\theta)$ can then be interpreted as a quadratic form whose criterion is expected to be small at the true value θ_0.

Under weak identification, i.e. the case in Assumption 1(ii), the argument has to be modified. As proved above, $\widehat{\theta}$ is no longer consistent for θ_0. Therefore, the fact that the FOC hold at $\widehat{\theta}$ does not imply automatically that they have to hold at the true value θ_0, not even approximately or asymptotically. However, as shown in Lemma 10 below, under weak identification the FOC $n^{-1}\sum_{i=1}^n \rho_1(\lambda' g_i(\theta)) g_i(\theta) = 0$ not only holds at $\widehat{\theta}$ wpa1 but holds uniformly over $\theta \in \Theta$ wpa1. Therefore, the FOC is not a condition that asymptotically pins down the true value θ_0, but a condition that holds asymptotically for all $\theta \in \Theta$. Under weak identification, we therefore should not expect that hypothesis tests for θ_0 based on the statistics $K_\rho^L(\theta)$ or $K_\rho^W(\theta)$ have good power properties. This is corroborated by the Monte Carlo simulations below and by the next Corollary. However, the tests are asymptotically similar under both cases of Assumption 1. Everything stated for $K_\rho^L(\theta)$ and $K_\rho^W(\theta)$ in this paragraph also applies to Kleibergen's K statistic, see Kleibergen (2001, 2002).

The next result provides the asymptotics for $K_\rho^W(\theta)$ and $K_\rho^L(\theta)$ for fixed arbitrary $\theta \in \Theta$ under weak asymptotics. For $\xi < 1/2$, when $\theta \neq \theta_0$, the factor $n^{1/2}\widehat{g}(\theta)$ converges to infinity at a rate of $n^{(1/2)-\xi}$. Under strong asymptotics I thus only give the asymptotic distribution for θ_0.

Note that in my linear model we have $G_i(\theta) = G_i = -n^{-\xi}Z_iZ_i'C - Z_iV_i' = -Z_iY_i'$.

Corollary 4 *Suppose Assumptions 2 and 4 and fix $\theta \in \Theta$.*

(i) Under Assumption 1(i) we have $K_\rho^W(\theta_0) \to_d \chi^2(p)$, and $K_\rho^L(\theta_0) \to_d \chi^2(p)$.

(ii) Under Assumption 1(ii) the asymptotic distribution of the test statistics $K_\rho^W(\theta)$ and $K_\rho^L(\theta)$ is identical and given by

$$(W(\theta) + \zeta)'(W(\theta) + \zeta),$$

where $\zeta \sim N(0, I_p)$,where the nonstandard distribution of the random p-vector $W(\theta)$ is defined in (4.8), and where ζ and $W(\theta)$ are independent. Because $W(\theta_0) \equiv 0$, for $\theta = \theta_0$, we have $K_\rho^W(\theta_0) \to_d \chi^2(p)$, and $K_\rho^L(\theta_0) \to_d \chi^2(p)$.

COMMENT ON POWER PROPERTIES UNDER WEAK IDENTIFICATION.

To use the above corollary for hypothesis tests or for the construction of confidence intervals we have to replace the unknown matrix $\Omega(\theta)^{-1}$ by a consistent estimate. For example, we can use the sample average

$$\widehat{\Omega}(\theta)^{-1} = (\sum_{i=1}^n g_i(\theta)g_i(\theta)'/n)^{-1}. \tag{2.14}$$

Recall that the CUE FOC for $\lambda(\theta)$ is given by $\lambda(\theta) = -\widehat{\Omega}(\theta)^{-1}\widehat{g}(\theta)$ from which it follows that if we estimate $\Omega(\theta)^{-1}$ by $\widehat{\Omega}(\theta)^{-1}$ then for CUE the LM-type and Wald-type statistics are numerically equivalent. For other GEL estimators however the two statistics do generally not coincide. Kleibergen's K statistic does not coincide with K_ρ^W for quadratic ρ. The K statistic uses the FOC for CUE_{GMM} defined in Hansen, Heaton, and Yaron (1996) while K_ρ^W uses the FOC for GEL for quadratic ρ. I have mentioned earlier that the two estimators do in general not coincide.

2.4 Some extensions

2.4.1 Tests for subvectors of the parameter vector

In general, an applied researcher is interested in inference on a subvector of the parameter vector rather than inference on the whole parameter vector. For example, in determining the impact of education on wage, the set of regressors may include besides others: education, work experience, a dummy variable for marriage, and the number of children. However, interest focuses only on the parameter for education. Likewise, when examining the impact of jobtraining on productivity the sole purpose of inference is the parameter of the dummy variable for jobtraining and not the parameters of the other regressors like wage, quality of work environment, etc..

Therefore, I now generalize the procedure that led to the test statistic in Corollary 4 to a setup that focuses on inference for subvectors. In doing so, I apply the method

in Kleibergen (2001) to the GEL case. While so far, the asymptotic null distribution of all tests considered was independent of the specific case in Assumption 1, for the subvector test the asymptotic null distribution depends on whether the model is weakly or strongly identified.

I start off with some notation. Let $\theta_0 = (\alpha_0', \beta_0')'$, $\theta = (\alpha', \beta')'$, where α, α_0, and β, β_0 are p_1 and p_2 dimensional vectors, respectively, where $p_1 + p_2 = p$ and $p_2 < p$. Assume we are interested in inference on β_0. Define $h_{i\beta^*}(\alpha) := g_i(\alpha, \beta^*)$ for a fixed hypothesized p_2-vector β^*. Then $Eh_{i\beta^*}(\alpha_0) = 0$ if $\beta^* = \beta_0$. I write $h_i(\alpha)$ for $h_{i\beta^*}(\alpha)$. For $\lambda \in \widehat{\Lambda}_n(\alpha) := \{\lambda \in R^k : \lambda' h_i(\alpha) \in Q \text{ for } i = 1, ..., n\}$, let $\widehat{P}(\alpha, \lambda) := 2\sum_{i=1}^n \rho(\lambda' h_i(\alpha))/n - 2\rho_0$ and define the GEL estimator $\widehat{\alpha}$ for α_0 by $\widehat{\alpha} = \widehat{\alpha}(\beta^*) := \arg\min_{\alpha \in \{\alpha \in R^{p_1} : (\alpha', \beta^{*\prime})' \in \Theta\}} \sup_{\lambda \in \widehat{\Lambda}_n(\alpha)} \widehat{P}(\alpha, \lambda)$.

Suppose Assumption 1(i), i.e. the classical case of strong identification, and $\beta^* = \beta_0$. From NS, Lemma A2 and Theorem 3.2, it follows that $\lambda(\widehat{\alpha}) := \arg\max_{\lambda \in \widehat{\Lambda}_n(\widehat{\alpha})} \widehat{P}(\widehat{\alpha}, \lambda)$ exists wpa1, and that $n^{1/2}\lambda(\widehat{\alpha}) \to_d N(0, P)$[4], where

$$P := (\sigma_{uu} Q_{ZZ})^{-1} - (\sigma_{uu} Q_{ZZ})^{-1} EH_i [EH_i' (\sigma_{uu} Q_{ZZ})^{-1} EH_i]^{-1} EH_i' (\sigma_{uu} Q_{ZZ})^{-1},$$

where $H_i := (\partial h_i(\alpha)/\partial \alpha) = -Z_i Z_i' \Pi_{p_1} - Z_i V_{ip_1}' = -Z_i Y_{ip_1}'$, Π_{p_1} is the $k \times p_1$ matrix consisting of the first p_1 columns of Π and likewise, V_{ip_1} and Y_{ip_1} are the subvectors that consist of the first p_1 components of V_i and Y_i, respectively. It is assumed that Π_{p_1} has full rank p_1. The GEL FOC for $\widehat{\alpha}$ is given by

$$0 = \lambda(\widehat{\alpha})' \sum_{i=1}^n \rho_1(\lambda(\widehat{\alpha})' h_i(\widehat{\alpha})) H_i/n = \lambda(\widehat{\alpha})' D_{\rho\beta_0}(\widehat{\alpha}),$$

where I have defined the $k \times p_1$ matrix $D_{\rho\beta_0}(\alpha) := \sum_{i=1}^n \rho_1(\lambda(\alpha)' h_i(\alpha)) H_i/n$. The FOC motivates the following test statistics

$$\begin{aligned} K^L_{\rho\beta_0}(\widehat{\alpha}) :&= n\lambda(\widehat{\alpha})' W_{\rho\beta_0}(\widehat{\alpha}) \lambda(\widehat{\alpha}), \\ K^W_{\rho\beta_0}(\widehat{\alpha}) :&= n\widehat{h}(\widehat{\alpha})' (\sigma_{uu} Q_{ZZ})^{-1} W_{\rho\beta_0}(\widehat{\alpha}) (\sigma_{uu} Q_{ZZ})^{-1} \widehat{h}(\widehat{\alpha}), \\ \text{where } W_{\rho\beta_0}(\widehat{\alpha}) :&= D_{\rho\beta_0}(\widehat{\alpha}) (D_{\rho\beta_0}(\widehat{\alpha})' P D_{\rho\beta_0}(\widehat{\alpha}))^{-1} D_{\rho\beta_0}(\widehat{\alpha})', \end{aligned}$$

and where $\widehat{h}(\alpha) := \sum_{i=1}^n h_i(\alpha)/n$.

Theorem 5 *Suppose Assumptions 1(i), and 2-4 hold, and that Π_{p_1}, the $k \times p_1$ matrix consisting of the first p_1 columns of Π, has full rank p_1. Then*

$$K^L_{\rho\beta_0}(\widehat{\alpha}) \to_d \chi^2_{p_1}, \text{ and } K^W_{\rho\beta_0}(\widehat{\alpha}) \to_d \chi^2_{p_1}.$$

In order to use the result in the theorem for hypothesis testing or the construction of confidence intervals for β_0, we have to replace the unknown quantities P and $\sigma_{uu} Q_{ZZ}$ in the statistics $K^L_{\rho\beta_0}(\widehat{\alpha})$ and $K^W_{\rho\beta_0}(\widehat{\alpha})$ by consistent estimators. If $\beta = \beta_0$,

[4] As mentioned above in Remark 1, these results of NS can be established under my slightly weaker moment assumptions.

the matrix $\sigma_{uu}Q_{ZZ}$ is consistently estimated by $\sum_{i=1}^{n} h_i(\widehat{\alpha})h_i(\widehat{\alpha})'/n$, since under Assumption 1(i) $\widehat{\alpha} \to_p \alpha_0$, see NS, Theorem 3.2. Using this estimate and replacing EH_i by its sample average, we obtain a consistent estimate of P. The two test statistics $K^L_{\rho\beta_0}(\widehat{\alpha})$ and $K^W_{\rho\beta_0}(\widehat{\alpha})$ are again numerically identical in the case $\rho(\upsilon) = -(1+\upsilon)^2/2$ if P and $\sigma_{uu}Q_{ZZ}$ are replaced by these estimators.

Even though it appears difficult to derive the asymptotic distribution under Assumption 1(ii), there is strong evidence that the statistics $K^L_{\rho\beta_0}(\widehat{\alpha})$ and $K^W_{\rho\beta_0}(\widehat{\alpha})$ no longer converge to a $\chi^2_{p_1}$ random variable. The reason is that in general the quantities $n^{1/2}\lambda(\widehat{\alpha})$ in $K^L_{\rho\beta_0}(\widehat{\alpha})$ and $n^{1/2}\widehat{h}(\widehat{\alpha})$ in $K^W_{\rho\beta_0}(\widehat{\alpha})$ no longer converge to a normal distribution asymptotically, because of their dependence on $\widehat{\alpha}$, which, in direct consequence of Theorem 2, has a nonstandard asymptotic distribution.

If one applies the same ideas to Kleibergen's (2002) K statistic to come up with a subvector test, one runs into the same problem. In fact, for the CUE_{GMM}, $\widehat{\alpha}$ also has a nonnormal limiting distribution, as shown in Stock and Wright (2000).

It therefore still remains to find a statistic for a subvector test that is similar under both weak and strong identification.

2.4.2 Weak and strong identification

I now generalize Assumption 1 to a scenario where some parameters are weakly and some strongly identified and then derive asymptotic results for the GEL estimator in this setup.

Assumption 1': $\Pi_n = (\Pi_A, \Pi_B)$, $C = (C_A, C_B)$, $\Pi_A = n^{-1/2}C_A$, $\Pi_B = C_B$. The matrices Π_A and C_A are $R^{k\times p_1}$, and Π_B and C_B are $R^{k\times p_2}$, where $p_1 + p_2 = p$, and p_1 and $p_2 \geq 1$. Let C_B have full column rank.

Conformably with Π_n, I write $Y = (Y_A, Y_B)$, $\theta = (\alpha', \beta')'$, $\widehat{\theta} = (\widehat{\alpha}', \widehat{\beta}')'$ and $\theta_0 = (\alpha_0', \beta_0')'$. Assumption 1' specializes the GMM weak identification assumption of Stock and Wright (2000) to the linear model (see their Assumption C, p. 1061, and the application to the linear model p.1070). It defines the parameter vector α_0 as weakly identified and β_0 as strongly identified. Assumption 1' contains as a particular case the partially identified model of Phillips (1989). Choosing p_1 and setting $C_A = 0$, we obtain a model in which C has any desired (less than full) rank.

To simplify notation, from now on I write (α, β) for $(\alpha', \beta')'$. Define $A := \{\alpha \in R^{p_1} | \exists \beta \in R^{p_2} \text{ s.t. } (\alpha, \beta) \in \Theta\}$, the projection of Θ on the first p_1 components, and similarly $B := \{\beta \in R^{p_2} | \exists \alpha \in R^{p_1} \text{ s.t. } (\alpha, \beta) \in \Theta\}$. Because Θ is compact, the same holds for A and B.

An analogous statement to Lemma 1 is given by the next result. Define the $k \times k$ matrix

$$\Omega_\beta(\theta_1, \theta_2) := [\Omega(\theta_1, \theta_2) + E((Z_i'C_B(\beta_0 - \beta_1))(Z_i'C_B(\beta_0 - \beta_2))Z_iZ_i')].$$

Write $\Omega_\beta(\theta)$ for $\Omega_\beta(\theta, \theta)$.

Lemma 6 *Under Assumptions 1', 2, and 3 the following results hold.*

(*i*) $\sup_{\theta_1,\theta_2\in\Theta} ||\widehat{\Omega}(\theta_1,\theta_2)-\Omega_\beta(\theta_1,\theta_2)|| \to_p 0$, *in particular,* $\sup_{\theta\in\Theta} ||\widehat{\Omega}(\theta)-\Omega_\beta(\theta)|| \to_p 0$.

(*ii*) $\sup_{\theta\in\Theta} ||\widehat{g}(\theta) - Q_{ZZ}C_B(\beta_0-\beta)|| \to_p 0$.

The next theorem establishes the asymptotic behavior of $\widehat{\theta}$. Let $K \subset R^{p_2}$ be an arbitrary compact neighborhood of 0.

Theorem 7 *Under Assumptions 1', and 2-4, we have that*

$$(i)\ \widehat{\alpha} \text{ is (in general) inconsistent, and } n^{1/2}(\widehat{\beta}-\beta_0) = O_p(1).$$

Furthermore, for $(\alpha, b) \in A \times K$,

$$(ii)\ n\widehat{P}((\alpha,\beta_0+n^{-1/2}b),\lambda((\alpha,\beta_0+n^{-1/2}b))) \Rightarrow P((\alpha,\beta_0+b),(\alpha,\beta_0)).$$

Therefore, assuming a unique $\arg\min_{(\alpha,b)\in A\times K} P((\alpha,\beta_0+b),(\alpha,\beta_0))$ *exists, we have*

$$(\widehat{\alpha}, n^{1/2}(\widehat{\beta}-\beta_0)) \to_d (\alpha^*,\beta^*) := \arg\min_{(\alpha,b)\in A\times K} P((\alpha,\beta_0+b),(\alpha,\beta_0)),$$

The theorem shows that $\widehat{\beta}$ is $n^{1/2}$-consistent with $n^{1/2}(\widehat{\beta}-\beta_0)$ being nonstandard in general. The reason why the asymptotic distribution is nonnormal is a consequence of the inconsistent estimation of $\widehat{\alpha}$. An equivalent result has been obtained in Stock and Wright (2000), see their Theorem 1, for GMM estimators with weak identification. Their result contains as one particular case the CUE_{GMM}.

When all parameters are strongly identified, i.e. $p_1 = 0$, the above theorem does not apply. The resulting distribution in that case is given in NS. The estimator is consistent and asymptotically normal. The distribution of the estimator when $p_2 = 0$ was given above in Theorem 2. Therefore, together with the result in Theorem 7, all possible combinations of $p_1$and p_2 are covered.

As before I get the following corollary.

Corollary 8 *Let Assumptions 1', and 2-4 hold. Then for fixed* $(\alpha, b) \in A \times K$ *we have*

$$n\widehat{P}((\alpha,\beta_0+n^{-1/2}b),\lambda((\alpha,\beta_0+n^{-1/2}b))) \to_d \eta\chi^2(k,\delta),$$

where the noncentrality parameter δ *and the parameter* η *are given by*

$$\delta := ||\Omega(\alpha,\beta_0)^{-1/2}Q_{ZZ}C(\alpha_0-\alpha,-b)||^2/\eta,$$
$$\eta := ||\Sigma_{UU}^{-1/2}(1,\alpha_0-\alpha,-b)||^2/||\Sigma_{UU}^{-1/2}(1,\alpha_0-\alpha,0)||^2$$

In particular, at the true value we have $n\widehat{P}(\theta_0,\lambda(\theta_0)) \to_d \chi^2(k)$.

DERIVE ASYMPTOTICS FOR K TYPE STATISTICS UNDER ASSUMPTION 1'.

3 Monte Carlo Experiment

To assess the finite sample performance of the hypothesis tests constructed from the asymptotic results in Corollaries 3 and 4 I conduct a small Monte Carlo study. The experimental setup is taken from Kleibergen (2002). The data generating process (DGP) is given by model (2.1)

$$\begin{aligned} y &= Y\theta_0 + u, \\ Y &= Z\Pi + V, \end{aligned} \tag{3.15}$$

with $p = 1$, $n = 100$, and $Z \sim N(0, I_k \otimes I_n)$. I choose two values for k, namely $k = 1$ and $k = 5$, the just-identified case and a over-identified case, respectively. In the over-identified case I let $\Pi = (\Pi_1, 0, 0, 0, 0)$, i.e. I add on a number of irrelevant instruments. In both cases I look at two different values of the real number Π_1, namely $\Pi_1 = 0.1$ which I call the weak instrument case, and $\Pi_1 = 1$ which I call the strong instrument case.

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

Error distributions

I experiment with several distributions for (u, V) to investigate the robustness of the test statistics to several possible features of the error distribution.

In the first Design (I) I let $(u, V)' \sim N(0, \Sigma \otimes I_n)$, where $\Sigma \in R^{2\times 2}$ with diagonal elements 1 and off-diagonal elements 0.99. The specification of the covariance matrix Σ implies that y and Y are strongly endogenous. This is the case considered in Kleibergen (2002).

In the second Design (II) I examine the robustness of the performance of the test statistics towards thick tails in the error distribution of the structural equation. I modify Design (I) by using $u_i/(w_i/r)^{1/2}$ instead of u_i, where w_i is a chi-squared random variable with r degrees of freedom independent of u_i and V_i, i.e. this time the error in the structural equation has a t-distribution with r degrees of freedom. Design II(i) and II(ii) take $r = 2, 3$, respectively. Design II(i) does not satisfy the assumptions on second moments made in Assumption 2. Because in empirical applications it is usually only assumed but unknown whether or not second moments of the error distribution exist it is important to know about the performance of a test statistic if the errors do not have second moments.

Design (III) modifies Design (I) by exchanging u_i by $u_i^2 - 1$ i.e. this time the error in the structural equation has a recentered chi-squared distribution with one degree of freedom. This case examines robustness towards an asymmetric structural error distribution.

In Design (IV) I take a bimodal distribution for u_i. Let B_i have a Bernoulli (.5,.5) distribution that is independent of all other random variables. Replace u_i from Design (I) by $B_i|u_i + 2| - (1 - B_i)|u_i + 2|$. The resulting error distribution has peaks at -2 and +2.

Test statistics

The following test statistics are compared in the Monte Carlo study.

I include the Anderson-Rubin test statistic (AR), see Anderson-Rubin (1949) or Kleibergen (2002)

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

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

I also calculate three statistics $A_\rho(\theta)$ from (2.10), for $\rho(\upsilon) = -(1+\upsilon)^2/2$ (CUE), $\rho(\upsilon) = \ln(1-\upsilon)$ (EL), and $\rho(\upsilon) = -\exp\upsilon$ (ET).

All of the above statistics are asymptotically distributed as $\chi^2(k)$ under the null. By contrast, the following statistics are asymptotically distributed as $\chi^2(p)$ under the null.

I include the K statistic, recently proposed by Kleibergen (2002), and given by

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

where $\widetilde{Y}(\theta) := Z\widetilde{\Pi}(\theta)$, $\widetilde{\Pi}(\theta) = (Z'Z)^{-1}Z'[Y - (y - Y\theta)s_{uV}(\theta)/s_{uu}(\theta)]$, and $s_{uV}(\theta) := (y - Y\theta)' M_Z Y/(n-k)$.

I calculate three statistics for each of $K_\rho^W(\theta)$ and $K_\rho^L(\theta)$ defined in (2.12) and (2.13) with the same choices for ρ as for $A_\rho(\theta)$ above and where $\Omega(\theta)$ is replaced by the consistent estimator $\widehat{\Omega}(\theta)$, see (2.14).

Moreira (2002) introduces a testing procedure for simple hypothesis tests for the parameters of the structural model that has exact size in finite samples if errors are normal with known covariance matrix Λ for the reduced form errors. The tests are shown to be asymptotically similar for more general nonnormal errors and estimated covariance matrix. I include Moreira's conditional likelihood ratio test in my simulation (see Section 3 in Moreira (2002) for motivation). For the model (3.15) with only one endogenous variable it can be described as follows. Let

$$\begin{aligned}
LR_M &: = \frac{1}{2}[\overline{S}'\overline{S} - \overline{T}'\overline{T} + \{(\overline{S}'\overline{S} + \overline{T}'\overline{T})^2 - 4(\overline{S}'\overline{S}\overline{T}'\overline{T} - (\overline{S}'\overline{T})^2)\}^{1/2}], \quad (3.16)\\
\text{where } \overline{S} &: = (Z'Z)^{-1/2} S (b_0' \widehat{\Lambda} b_0)^{-1/2},\ \overline{T} := (Z'Z)^{-1/2} T (a_0' \widehat{\Lambda}^{-1} a_0)^{-1/2},\\
\text{where } S &: = Z'(y - Y\theta_0),\ T := Z'(y, Y)\widehat{\Lambda}^{-1} a_0,\\
\text{where } a_0 &: = (\theta_0, 1)',\ b_0 := (1, -\theta_0)',\ \text{and } \widehat{\Lambda} := (y, Y)' M_Z (y, Y)/(n-k).
\end{aligned}$$

In the normal model with known Λ and under the null, Moreira (2002) shows that $\overline{S}'\overline{S} \sim \chi^2(k)$. The test now works as follows. Given $\overline{T}'\overline{T}$, simulate R independent samples from a $\chi^2(k)$ distribution for $\overline{S}'\overline{S}$ and calculate the resulting R values for LR_M from (3.16): $LR_{M,1}, ..., LR_{MR}$. For a given size α, let the real number $c(\alpha)$ be the $(1-\alpha)$ quantile of the sample $\{LR_{M,1}, ..., LR_{MR}\}$. Reject the null, iff $LR_M > c(\alpha)$.

Finally, I include the Wald statistic for two-stage least squares (denoted $2SLS$), see for example Wooldridge (2002, p. 98)

$$2SLS := \widehat{\theta}' W^{-1} \widehat{\theta},$$

where $\widehat{\theta} := (Y' P_Z Y)^{-1} Y' P_Z y$, and W is a covariance matrix estimate of $\widehat{\theta}$. I use $W := \widehat{\sigma}^2 (Y' P_Z Y)^{-1}$ where $\widehat{\sigma}^2 := (n-k)^{-1} \sum_{i=1}^n \widehat{u}_i^2$ and $\widehat{u}_i := y_i - Y_i' \widehat{\theta}$.

Recall that for CUE, $K_\rho^W(\theta)$ and $K_\rho^L(\theta)$ are numerically identical. Also, in the just-identified case $k = p = 1$, the AR and K statistics coincide, see Kleibergen (2002).

To calculate $A_\rho(\theta)$, $K_\rho^W(\theta)$, and $K_\rho^L(\theta)$ for EL and ET, I have to solve the globally concave maximization problem $\max_{\lambda \in \widehat{\Lambda}_n(\theta)} \widehat{P}(\theta, \lambda)$ numerically. To do that I implement a variant of the Newton-Raphson algorithm. I start the algorithm by setting λ equal to the zero vector. In 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" of the algorithm.

It should be expected that in the over-identified case the statistics that converge to a $\chi^2(p)$ under the null are more powerful than the statistics that converge to a $\chi^2(k)$ under the null.

Size comparison

Tables I (1)-(4) contain the observed sizes at the 5% asymptotic critical values of all the statistics and all the error distributions described above. The sizes are computed using 10,000 samples from the DGP in (3.15)[5]. I also use 10,000 realizations from a $\chi^2(k)$ to estimate the cut-off value in Moreira's statistic. Table I(1) summarizes the results for the case $\Pi_1 = 1$, $k = 5$, Table I(2) for $\Pi_1 = 1$, $k = 1$, Table I(3) for $\Pi_1 = .1$, $k = 5$, and Table I(4) for $\Pi_1 = .1$, $k = 1$.

The performance of $2SLS$ depends crucially on the strength of the instrument and on the number of over-identifying restrictions. The actual sizes are close to the nominal 5% size only in the just-identified case with strong instruments. If $k = 5$ the test is somewhat off with strong instruments (the actual sizes range over the interval [4.3-9.3] across the different error distributions), and performs disastrously with weak instruments (the actual sizes range over the interval [0.8-92.2] !). When instruments are weak and $k = 1$ the actual sizes of 2SLS are also far from the nominal size, ranging from 0.2 to 18.3%. In short, the weak instrument case corresponds to a setup where 2SLS is not a reliable statistic for inference in small samples. Even in the strong instrument case, 2SLS is not recommended in the over-identified model.

The statistics K_{EL}^L,and K_{ET}^L consistently strongly over-reject in all of the experiments (actual sizes across all experiments range from 5.5-24.1% and 6.5-16.0%, respectively). The performance of A_{EL}, and A_{ET} is acceptable in the just-identified case but bad in the over-identified case. None of the statistics in this paragraph can

[5]Kleibergen (2002) 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 (3.15). I simulate a new matrix Z with each sample of the DGP (3.15). As a consequence, my results do not coincide with the results that Kleibergen (2002) reports.

To investigate the sensitivity of the results in Kleibergen (2002) to the choice of Z, I iterated Kleibergen's (2002) type of procedure 100 times, i.e. each time I simulated a matrix Z of instruments that I then kept fixed across $R = 1000$ samples of the DGP (3.15). I 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 $\theta = 0$ when $\theta_0 = .4$, varied from about 60% to 95% in the 100 experiments. For the specific Z that Kleibergen (2002) generated, the reported power is about 93%, see his Figure 1.

therefore be recommended for empirical applications that resemble the setup of that Monte Carlo study.

The remaining statistics are relatively reliable across all the different scenarios. The statistics A_{CUE}, K^W_{CUE}, and K^W_{ET} consistently under-reject (with actual size ranging from 2.6-5.0%, 3.0-5.0%, and 3.6-5.0%. respectively, across all scenarios). The same is true for K^W_{EL} except for three scenarios for the case $\Pi_1 = .1$, $k = 5$ where the actual sizes are 5.1, 5.4, and 6.1%. In contrast, the statistics LR_M, AR, and K usually slightly over-reject across all designs (actual sizes range from 5.4-6.9%, 5.2-6.5%, and 4.8-6.0%, respectively).

Based on the size results of my Monte Carlo study I think that K^W_{EL} and K enjoy a slight advantage over the remaining statistics. The actual size of K^W_{EL} is closest to the nominal size across all statistics in 9 of the 20 cases considered. The K statistic wins 8 of the 20 comparisons.

Power comparison

For each of the 20 setups reported in Tables I (1)-(4) I calculate power curves at a 5% significance level for all the above statistics, but report detailed results only for those that have reliable size properties for weak and strong identification in the above size experiment, i.e. the statistics K^W_{CUE}, K^W_{EL}, K^W_{ET}, A_{CUE}, LR_M, AR, and K. Except for LR_M, I report size-corrected power curves, using cut-off values calculated in the size comparison above. Due to the conditional construction of LR_M, size-correction for this statistic is not completely straightforward, and I therefore calculate a power curve for LR_M that has not been size-corrected.

I use 1,000 samples from the DGP in (3.15) for various values of the true value θ_0 and test the hypothesis that $\theta = 0$. I use 1,000 realizations from a $\chi^2(k)$ random variable to estimate the cut-off value for the LR_M statistic. For the results that I actually report in the Figures below, I use 10,000 samples.

With strong identification all statistics have a U-shaped power curve. With the exception of $2SLS$, the lowest point of the power curve is usually achieved at the true value θ_0. In the overidentified case, the test $2SLS$ is usually biased, taking on its lowest value at a negative θ value.

With weak identification the power curves do not have a clear pattern. In most cases, for postive true values of θ_0, the power curves are very flat, hardly exceeding the significance level of the test. For negative true values of θ_0, the power curves are sharply peaked at one particular value and then flatten out (with normal and bimodal errors) or slowly grow with increasing $|\theta_0|$ (with asymmetric and thick tail errors).

Across all scenarios the statistics K^W_{CUE}, K^W_{EL}, and K^W_{ET} have very similar performance and therefore I only report results for K^W_{EL}. In the case $k = 1$, AR and K are numerically identical. In the case $k = 5$, as also found in Kleibergen (2002), K performs considerably better than AR when identification is strong, and performs somewhat better when identification is weak. I therefore do not report results for AR but refer to Kleibergen (2002) for the comparision of K and AR.

I now discuss the performance of the remaining statistics K^W_{EL}, A_{CUE}, LR_M, and K.

Figures I(1)-(4) and Figure II (1)-(4) display power curves of the four statistics with strong identification for $k = 1$ and $k = 5$, respectively, for the error distributions of Design I, II(i), III, and IV. The outcome in Design II(ii) qualitatively resembles II(i) and I therefore do not separately report it to save space.

In the just-identified case, K^W_{EL} and A_{CUE} are virtually identical throughout all scenarios. Except for the normal and bimodal case, they slightly outperform the other two statistics. In the normal case, the performance of all four statistics is virtually identical, in the bimodal case there is a small advantage to the LR_M statistic.

When moving to the over-identified case, K^W_{EL} and A_{CUE} are no longer identical. From the theoretical results, it was to be expected that K^W_{EL} has better power properties than A_{CUE} in over-identified situations because the former is distributed asymptotically as $\chi^2(p)$ while the latter is distributed as $\chi^2(k)$ under the null. In all scenarios of the over-identified case, K^W_{EL} and K have better power properties than A_{CUE} and LR_M. The ranking of K^W_{EL} and K and the ranking of A_{CUE} and LR_M is not so clear cut. For errors with thick tails and asymmetric errors K^W_{EL} is usually more powerful than K, whereas K dominates K^W_{EL} for normal and bimodal errors. The comparison of A_{CUE} and LR_M is further complicated by the fact that LR_M over-rejects somewhat. It is probably fair to say that A_{CUE} dominates LR_M for asymmetric errors and errors with thick tails while for the remaining cases LR_M takes the lead.

Moving from the just-identified to the over-identified case results in considerable power loss for A_{CUE} and LR_M. Quite surprisingly, this conclusion is not true for K^W_{EL} and K. On the contrary, for some cases, the K^W_{EL} statistic is more powerful in the over-identified case, compare, for example, Figures I(2) and II(2), the case of $t(2)$ errors.

With weak identification the power curves of all four statistics have the shape described above. Power of all the tests is very low (except for a peak point in the normal and bimodal case) and their performance is not significantly different from each other.

Overall the statistics that perform best are K^W_{EL} and K. The former is more robust to errors with thick tails or asymmetric errors, the latter performs better with normal or bimodal errors. Both statistics remain powerful in over-identified scenarios.

Acknowledgements

My thanks go to Richard Smith whose many suggestions helped to substantially improve the content of this paper and with whom I am working on a generalization of this project to the time series context. This paper has also benefited from very helpful comments by my advisors Donald Andrews and Peter Phillips. I would also like to thank Frank Kleibergen for helpful correspondence and Vadim Marner for help with the simulation section.

4 Appendix of Proofs

Proof of Lemma 1. (i) First note that the probability limit of $\widehat{\Omega}(\theta_1, \theta_2)$ equals that of $n^{-1}\sum_{i=1}^n [u_i + V_i'(\theta_0 - \theta_1)][u_i + V_i'(\theta_0 - \theta_2)]Z_i Z_i'$. This holds because the terms in $\widehat{\Omega}(\theta_1, \theta_2)$ that are crossproducts involving $n^{-1/2}Z_i'C$ terms are negligible. More precisely, we have uniformly over $\theta_1, \theta_2 \in \Theta$,

$$n^{-1}\sum_{i=1}^n u_i n^{-1/2} Z_i' C(\theta_0 - \theta_2) Z_i Z_i' = O_p(n^{-1/2}),$$
$$n^{-1}\sum_{i=1}^n n^{-1} Z_i' C(\theta_0 - \theta_1) Z_i' C(\theta_0 - \theta_2) Z_i Z_i' = O_p(n^{-1}),$$
$$n^{-1}\sum_{i=1}^n n^{-1/2} Z_i' C(\theta_0 - \theta_1) V_i'(\theta_0 - \theta_2) Z_i Z_i' = O_p(n^{-1/2}),$$

by Assumption 2, compactness of Θ, and the weak law of large numbers (WLLN). The result now follows by the WLLN, compactness of Θ, and by taking conditional expectations.

(ii) Because $\sup_{\theta\in\Theta} ||n^{-1}\sum_{i=1}^n Z_i' C(\theta_0 - \theta) Z_i - Q_{ZZ} C(\theta_0 - \theta)|| \to_p 0$, we only have to deal with the empirical process $\nu_n(\cdot, \theta) := n^{-1/2}\sum_{i=1}^n (u_i + V_i'(\theta_0 - \theta))Z_i$. Fidi convergence follows by the CLT and stochastic equicontinuity follows by the fact that $(\theta_0 - \theta)$ enters $\nu_n(\cdot, \theta)$ linearly:

$$\sup_{||\theta_1 - \theta_2|| < \delta} ||\nu_n(\cdot, \theta_1) - \nu_n(\cdot, \theta_2)|| = \sup_{||\theta_1 - \theta_2|| < \delta} ||(\theta_2 - \theta_1)' n^{-1/2} \sum_{i=1}^n V_i Z_i|| = \delta O_p(1).$$

By assumption, Θ is compact and thus the proposition on p.2251 in Andrews (1994) can be applied which yields the desired result. □

The next two Lemmas are a modified version of Lemmas A1 and A2 in NS. The modifications are necessary because we work with local to zero asymptotics and because my moment assumptions are slightly weaker, see Remark 1. The lemmas are needed for the proof of Theorem 2.

Lemma 9 *Let* $b_{ni} := \sup_{\theta\in\Theta} ||g_{ni}(\theta)||$, $c_n := n^{-1/2}\max_{1\le i\le n} b_{ni}$, *and* $\Lambda_n := \{\lambda : ||\lambda|| \le n^{-1/2} c_n^{-1/2}\}$. *Then*

(*i*) $\sup_{\theta\in\Theta, \lambda\in\Lambda_n, 1\le i\le n} |\lambda' g_i(\theta)|$ *converges to* 0 *a.s.*.

(*ii*) *Wpa1,* $\Lambda_n \subset \widehat{\Lambda}_n(\theta)$, *uniformly over all* $\theta \in \Theta$.

Proof. An application of the Borel-Cantelli Lemma shows that for real-valued *iid* random variables W_i such that $EW_i^2 < \infty$ we have $\max_{1\le i\le n} |W_i| = o(n^{1/2})$, see Owen (1990, Lemma 3) for a proof. Because

$$\max_{1\le i\le n} b_{ni} \le \max_{1\le i\le n} \sup_{\theta\in\Theta} (||u_i Z_i|| + n^{-1/2}||Z_i' C(\theta_0 - \theta) Z_i|| + ||V_i(\theta_0 - \theta) Z_i'||), \quad (4.1)$$

applying this result to each of the three summands in (4.1) and using Assumption 2 implies that $\max_{1\leq i\leq n} b_{ni} = o(n^{1/2})$ and $c_n = o(1)$. Therefore part (i) follows from

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

which also immediately implies (ii). □

Lemma 10 *$\lambda(\theta) := \arg\max_{\lambda\in\widehat{\Lambda}_n(\theta)} \widehat{P}(\theta,\lambda)$ exists uwpa1, and $\lambda(\theta) = O_p(n^{-1/2})$ uniformly over $\theta\in\Theta$.*

Proof. Define $\lambda_\theta := \arg\max_{\lambda\in\Lambda_n} \widehat{P}(\theta,\lambda)$. This definition makes sense because a continuous function takes on its maximum on a compact set and by Lemma 9 wpa1 $\widehat{P}(\theta,\lambda)$ (as a function in λ for fixed θ) is C^2 on some open neighborhood of Λ_n. I now show that actually $\widehat{P}(\theta,\lambda_\theta) = \sup_{\lambda\in\widehat{\Lambda}_n(\theta)} \widehat{P}(\theta,\lambda)$ which then proves the first part of the lemma. By a second order Taylor expansion around $\lambda = 0$, there is λ_θ^* on the line segment $\overline{0\lambda_\theta}$, such that for some positive constants C_1 and C_2

$$\begin{aligned} 0 &= \widehat{P}(\theta,0) \leq \widehat{P}(\theta,\lambda_\theta) = -2\lambda_\theta'\widehat{g}(\theta) + \lambda_\theta'[\sum_{i=1}^n \rho_2(\lambda_\theta^{*\prime} g_i(\theta))g_i(\theta)g_i(\theta)'/n]\lambda_\theta \\ &\leq -2\lambda_\theta'\widehat{g}(\theta) - C_1\lambda_\theta'\widehat{\Omega}(\theta)\lambda_\theta \leq 2||\lambda_\theta||\ ||\widehat{g}(\theta)|| - C_2||\lambda_\theta||^2 \end{aligned} \tag{4.2}$$

where the second to last inequality follows from the fact that by Lemma 9, continuity of $\rho_2(\cdot)$, and $\rho_2 = -1$ we have uwpa1 $\max_{1\leq i\leq n}\rho_2(\lambda_\theta^{*\prime} g_i(\theta)) < -1/2$. The last inequality follows from Lemma 1(i), and positive definiteness of Σ_{UU} and Q_{ZZ} which imply that $\widehat{\Omega}(\theta)$ converges uniformly to a uniformly positive definite matrix and therefore uwpa1 its smallest eigenvalue is bounded away from zero. (4.2) implies that $(C_2/2)||\lambda_\theta|| \leq ||\widehat{g}(\theta)||$ uwpa1, the latter being $O_p(n^{-1/2})$ uniformly over $\theta\in\Theta$. This follows by Assumption 1(ii), the CLT and compactness of Θ. The second part of the lemma is therefore proven. Furthermore, it follows that $\lambda_\theta \in int(\Lambda_n)$ uwpa1. To prove this, let $\varepsilon > 0$. There exists $M_\varepsilon < \infty$ and $n_\varepsilon \in \mathbb{N}$ s.t. $\Pr(||n^{1/2}\lambda_\theta|| \leq M_\varepsilon) > 1-\varepsilon$ for all $n \geq n_\varepsilon$ uniformly over $\theta\in\Theta$. Choose $n(\varepsilon) > n_\varepsilon$ so big that $c_n^{-1/2} > M_\varepsilon$ for all $n \geq n(\varepsilon)$. Then $\Pr(\lambda_\theta \in int(\Lambda_n)) = \Pr(||n^{1/2}\lambda_\theta|| < c_n^{-1/2}) \geq \Pr(||n^{1/2}\lambda_\theta|| \leq M_\varepsilon) > 1-\varepsilon$ for $n \geq n(\varepsilon)$.

Hence the FOC for an interior maximum $\partial\widehat{P}(\theta,\lambda_\theta)/\partial\lambda = 0$ holds. By Lemma 9, uwpa1 $\lambda_\theta \in \widehat{\Lambda}_n(\theta)$ and thus by concavity of $\widehat{P}(\theta,\lambda)$ (as a function in λ for fixed θ) and convexity of $\widehat{\Lambda}_n(\theta)$ the first part of the lemma follows. □

Proof of Theorem 2. In this proof I write "uwpa1" for "uniformly over $\theta\in\Theta$ wpa1". By Lemma 10 we know that uwpa1 there exists a $\lambda(\theta)$ that satisfies the arg sup requirement in (2.3) and that $\lambda(\theta) = O_p(n^{-1/2})$ holds uniformly over $\theta\in\Theta$. Thus, the FOC with respect to λ, $n^{-1}\sum_{i=1}^n \rho_1(\lambda' g_i(\theta))g_i(\theta) = 0$ has to hold at $\lambda(\theta)$ uwpa1. Expanding the FOC in λ around 0, we get for some mean value $\widetilde{\lambda}$ on the line segment $\overline{0\lambda(\theta)}$

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

where the matrix $\widehat{\Omega}_{\widetilde{\lambda}\theta}$ has been implicitly defined. Because $\lambda(\theta) = O_p(n^{-1/2})$, Lemma 9 and Lemma 1(i) imply that $\widehat{\Omega}_{\widetilde{\lambda}\theta}$ is invertible uwpa1. Therefore $\lambda(\theta) = -\widehat{\Omega}^{-1}_{\widetilde{\lambda}\theta}\widehat{g}(\theta)$ uwpa1. Inserting this into a second order Taylor expansion for $\widehat{P}(\theta,\lambda)$ (with mean value λ^*, like in (4.2) above) we find that

$$\widehat{P}(\theta,\lambda(\theta)) = 2\widehat{g}(\theta)'\widehat{\Omega}^{-1}_{\widetilde{\lambda}\theta}\widehat{g}(\theta) - \widehat{g}(\theta)'\widehat{\Omega}^{-1}_{\widetilde{\lambda}\theta}\widehat{\Omega}_{\lambda^*\theta}\widehat{\Omega}^{-1}_{\widetilde{\lambda}\theta}\widehat{g}(\theta). \tag{4.3}$$

By Lemma 1(i) and Lemma 9, both matrices $\widehat{\Omega}_{\widetilde{\lambda}\theta}$, and $\widehat{\Omega}_{\lambda^*\theta}$ converge in probability to $\Omega(\theta)$ uniformly over $\theta \in \Theta$. Lemma 1(ii) implies that $n\widehat{P}(\theta,\lambda(\theta))$ converges weakly to $P(\theta)$. The theorem then follows from Lemma 3.2.1 in van der Vaart and Wellner (1996, p.286). □

Proof of Corollary 3. (i) From Lemma A2 in NS, we know that for $\theta_n := \theta_0 + n^{-1/2}d$, $\lambda(\theta_n)$ exists wpa1. Therefore, wpa1, the FOC $n^{-1}\sum_{i=1}^n \rho_1(\lambda(\theta_n)'g_i(\theta_n))g_i(\theta_n) = 0$ has to hold. Similar to the derivation of (4.3), expanding the FOC up to second order, solving for $\lambda(\theta_n)$, and plugging $\lambda(\theta_n)$ into a second order Taylor expansion for $n\widehat{P}(\theta_n,\lambda(\theta_n))$, shows that the latter is distributed asymptotically as $n\widehat{g}(\theta_n)'[\lim_{n\to\infty} E(g_i(\theta_n)g_i'(\theta_n))^{-1}]\widehat{g}(\theta_n)$. We have $n^{1/2}\widehat{g}(\theta_n) \to_d N(-Q_{ZZ}Cd, \sigma_{uu}Q_{ZZ})$ and $E(g_i(\theta_n)g_i'(\theta_n))^{-1} \to (\sigma_{uu}Q_{ZZ})^{-1}$, which immediately implies the claim.

(ii) The proof follows from the proof of Corollary 8. □

Proof of Corollary 4. (i) Note first that Lemma 9 is still valid for b_{ni} replaced by $b_i := \sup_{\theta\in\Theta}||g_i(\theta)||$, and $g_i(\theta)$ defined in (2.2) for $\xi = 0$. Noting that $g_i(\theta_0) = u_iZ_i$ and thus $\widehat{g}(\theta_0) = O(n^{-1/2})$, an argument as in (4.2) replacing θ by θ_0 and λ_θ by $\lambda(\theta_0)$ leads to $\lambda(\theta_0) = O(n^{-1/2})$. Proceeding as in Lemma 10 this implies that $\lambda(\theta_0) = \arg\sup_{\lambda\in\widehat{\Lambda}_n(\theta)} \widehat{P}(\theta_0,\lambda)$ exists wpa1. Therefore the FOC $n^{-1}\sum_{i=1}^n \rho_1(\lambda(\theta_0)'g_i(\theta_0))g_i(\theta_0) = 0$ holds wpa1. As earlier, a first-order Taylor expansion of $\rho_1(\lambda(\theta_0)'g_i(\theta_0))$ then leads to $\lambda(\theta_0) = -\Omega(\theta_0)^{-1}\widehat{g}(\theta_0) + o_p(1)$ and thus $n^{1/2}\lambda(\theta_0) \to_d N(0,\Omega(\theta_0)^{-1})$. Finally note that $D_\rho(\theta_0) \to_p Q_{zz}\Pi$.

(ii) The main part of this proof is used to show that $D_\rho(\theta)$ is asymptotically independent of $\lambda(\theta)$ and $\widehat{g}(\theta)$. To show that, I first develop the asymptotic joint distribution of $D_\rho(\theta)$ and $\widehat{g}(\theta)$. From now on, I write $D(\theta)$ for $D_\rho(\theta)$.

Expanding the FOC for λ, $n^{-1}\sum_{i=1}^n \rho_1(\lambda(\theta)'g_i(\theta))g_i(\theta) = 0$ for fixed $\theta \in \Theta$, as in the beginning of the proof of Theorem 2 using the same notation, leads to

$$\lambda(\theta) = -\widehat{\Omega}^{-1}_{\widetilde{\lambda}\theta}\widehat{g}(\theta), \text{ where } \widehat{\Omega}^{-1}_{\widetilde{\lambda}\theta} \to_p \Omega(\theta)^{-1}. \tag{4.4}$$

Expanding $\rho_1(\lambda(\theta)'g_i(\theta))$ up to first order and replacing $\lambda(\theta)$ by (4.4) we have for a certain mean value $\overline{\lambda}$

$$\begin{aligned} vec(D(\theta)) &= \sum_{i=1}^n [-vec(G_i)/n + (\rho_2(\overline{\lambda}'g_i(\theta))/n)\lambda(\theta)'g_i(\theta)vec(G_i)] \\ &= \sum_{i=1}^n [-vec(G_i)/n - \{\rho_2(\overline{\lambda}'g_i(\theta))vec(G_i)g_i(\theta)'/n\}\widehat{\Omega}^{-1}_{\widetilde{\lambda}\theta}\widehat{g}(\theta)]. \end{aligned}$$

Furthermore, by the definition of G_i, continuity of $\rho_2(\cdot)$ in 0, and Assumption 2

$$\sum_{i=1}^n \{\rho_2(\overline{\lambda}'g_i(\theta))vec(G_i)g_i(\theta)'/n\} = -\sum_{i=1}^n \{vec(-Z_iV_i')g_i(\theta)'/n\} + o_p(1) =$$

$$E\{vec(Z_iV_i')(u_iZ_i' + V_i'(\theta_0 - \theta)Z_i')\} + o_p(1) = \Omega_\theta + o_p(1),$$

where the $kp \times k$ matrix Ω_θ has been implicitly defined. Combining the two previous equations and using Lemma 1(ii) yields

$$\begin{aligned} n^{1/2}vec(D(\theta)) &= vec(Q_{ZZ}C) + [\sum_{i=1}^{n} n^{-1/2}vec(Z_iV_i')] - \Omega_\theta\Omega(\theta)^{-1}n^{1/2}\widehat{g}(\theta) + o_p(1) \\ &\to_d vec(Q_{ZZ}C) + (I_{kp}, -\Omega_\theta\Omega(\theta)^{-1})N(m, \Delta(\theta)), \end{aligned} \tag{4.5}$$

where $m := (0', (Q_{ZZ}C(\theta_0 - \theta))')'$, where 0 is a kp vector of zeros, and where the $(p+1)k \times (p+1)k$ matrix $\Delta(\theta)$ is given by

$$\Delta(\theta) := \begin{pmatrix} \Sigma_{VV} \otimes Q_{ZZ} & \Omega_\theta \\ \Omega_\theta' & \Omega(\theta) \end{pmatrix}.$$

Notice that by Assumptions 2(ii) and (iii) and Lemma 1(i) the matrices $\Sigma_{VV} \otimes Q_{ZZ}$ and $\Omega(\theta)$ have full rank. Thus $\Delta(\theta)$ has full rank. Equation (4.5) implies that

$$n^{1/2}vec(D(\theta)) \to_d N(m, Cov), \tag{4.6}$$

where $m := vec(Q_{ZZ}C) - \Omega_\theta\Omega(\theta)^{-1}Q_{ZZ}C(\theta_0 - \theta)$ and where $Cov := \Sigma_{VV} \otimes Q_{ZZ} - \Omega_\theta\Omega(\theta)^{-1}\Omega_\theta'$ is the full rank covariance matrix. Using (4.5), it is easily seen that $n^{1/2}vec(D(\theta))$ is asymptotically uncorrelated with and, thus, is asymptotically independent of $n^{1/2}\widehat{g}(\theta)$ and, hence, by (4.4) also asymptotically independent of $n^{1/2}\lambda(\theta)$. Therefore, by (4.4) and Lemma 1(ii) we have that asymptotically

$$(D(\theta)'\Omega(\theta)^{-1}D(\theta))^{-1/2}D(\theta)'n^{1/2}\lambda(\theta) \to_d W(\theta) + \zeta, \tag{4.7}$$

where the random p-vector $W(\theta)$ has a distribution given by the limiting distribution of $-(D(\theta)'\Omega(\theta)^{-1}D(\theta))^{-1/2}D(\theta)'\ \Omega(\theta)^{-1}Q_{ZZ}C(\theta_0 - \theta)$, where $\zeta \sim N(0, I_p)$, and where $W(\theta)$ and ζ are independent. Denote by $\overline{D}(\theta)$ the limiting normal distribution of $n^{1/2}D(\theta)$, see (4.6). I claim that $W(\theta)$ is distributed as

$$W(\theta) \sim -(\overline{D}(\theta)'\Omega(\theta)^{-1}\overline{D}(\theta))^{-1/2}\overline{D}(\theta)'\Omega(\theta)^{-1}Q_{ZZ}C(\theta_0 - \theta). \tag{4.8}$$

This follows by the Continuous Mapping Theorem once I show that the function $g : R^{k\times p} \to R^{p\times k}$ defined by $g(M) := (M'\Omega(\theta)^{-1}M)^{-1/2}M'$ for $M \in R^{k\times p}$ is continuous on a set $C \subset R^{k\times p}$ with $\Pr(\overline{D}(\theta) \in C) = 1$. Now, g is continuous at each M with full column rank. It is therefore enough to show that $\overline{D}(\theta)$ has full column rank a.s.. Define $O := \{o \in R^{kp} | \exists \widetilde{o} \in R^{k\times p}$, s.t. $o = vec(\widetilde{o})$ and $\widetilde{o}$ has linearly dependent columns$\}$. Clearly, O is closed and therefore Lebesgue-measurable. Also O has empty interior and thus has measure 0. From above we know that $vec(\overline{D}(\theta))$ has a full rank variance covariance matrix. This implies that for any measurable set O with empty interior we have $\Pr(vec(\overline{D}(\theta)) \in O) = 0$, in particular for the O above. This proves the claim.

Equation (4.7) immediately implies the result for $K_\rho^L(\theta)$. By (4.4) and Lemma 1(i), the asymptotic distribution of $K_\rho^W(\theta)$ then follows. □

Proof of Theorem 5. For the result involving $K^L_{\rho\beta_0}(\widehat{\alpha})$, it is enough to show that $D_{\rho\beta_0}(\widehat{\alpha})$ converges in probability to a matrix of full rank. We have

$$D_{\rho\beta_0}(\widehat{\alpha}) = \textstyle\sum_{i=1}^n \rho_1(\lambda(\widehat{\alpha})' h_i(\widehat{\alpha}))(-Z_i Z_i' \Pi_{p_1} - Z_i V_{ip_1}')/n.$$

Using the proof of Lemma 9 one can show that $\max_{i=1,\ldots,n} h_i(\widehat{\alpha}) = o_p(n^{1/2})$. To reach that conclusion we need that $\alpha_0 - \widehat{\alpha} = O_p(1)$. In fact, by NS Theorem 3.2., we even know that $\alpha_0 - \widehat{\alpha} = O_p(n^{-1/2})$ and $\lambda(\widehat{\alpha}) = O_p(n^{-1/2})$. The latter implies that $\max_{i=1,\ldots,n} |-1 + \rho_1(\lambda(\widehat{\alpha})' h_i(\widehat{\alpha}))| = o_p(1)$. Using Assumption 2, it then follows that $D_{\rho\beta_0}(\widehat{\alpha}) \to_p Q_{ZZ}\Pi_{p_1}$ which by assumption has full rank.

For the result involving $K^W_{\rho\beta_0}(\widehat{\alpha})$, I use once more the FOC for $\lambda(\widehat{\alpha})$, $n^{-1}\sum_{i=1}^n \rho_1(\lambda(\widehat{\alpha})' h_i(\widehat{\alpha})) h_i(\widehat{\alpha})$ that holds wpa1. A first-order Taylor expansion of $\rho_1(\lambda(\widehat{\alpha})' h_i(\widehat{\alpha}))$ implies that wpa1

$$\lambda(\widehat{\alpha}) = (n^{-1}\sum_{i=1}^n \rho_1(\lambda(\widehat{\alpha})' h_i(\widehat{\alpha})) h_i(\widehat{\alpha}) h_i(\widehat{\alpha})')^{-1} \widehat{h}(\widehat{\alpha}).$$

By the uniform WLLN and $n^{1/2}$-consistency of $\widehat{\alpha}$ the matrix in the last equation converges in probability to $-(E h_i(\alpha_0) h_i(\alpha_0)')^{-1} = -(\sigma_{uu} Q_{ZZ})^{-1}$. The result for $K^W_{\rho\beta_0}(\widehat{\alpha})$ then follows from the result involving $K^L_{\rho\beta_0}(\widehat{\alpha})$. □

Proof of Lemma 6. (i) In addition to the terms in the proof of Lemma 1(i) we have to deal with the crossproduct terms involving $Z_i' C_B(\beta_0 - \beta_i)$ for $i = 1, 2$. We have

$$n^{-1}\sum_{i=1}^n (n^{-1/2} Z_i' C_A(\alpha_0 - \alpha_1) Z_i' C_B(\beta_0 - \beta_2)) Z_i Z_i' = O_p(n^{-1/2}),$$
$$n^{-1}\sum_{i=1}^n ((u_i + V_i'(\theta_0 - \theta_1)) Z_i' C_B(\beta_0 - \beta_2)) Z_i Z_i' = o_p(1)$$

uniformly over $\theta_1, \theta_2 \in \Theta$ and thus the only nontrivial additional summand is given by $n^{-1}\sum_{i=1}^n (Z_i' C_B(\beta_0 - \beta_1))(Z_i' C_B(\beta_0 - \beta_2)) Z_i Z_i'$ which by the WLLN converges uniformly over $\beta_1, \beta_2 \in B$ to $E((Z_i' C_B(\beta_0 - \beta_1))(Z_i' C_B(\beta_0 - \beta_2)) Z_i Z_i')$.

(ii) This is a simple application of the WLLN together with the compactness of Θ. □

Proof of Theorem 7. (i) I first show consistency of $\widehat{\beta}$. If we can establish that $||\widehat{g}(\widehat{\theta})|| = o_p(1)$ consistency of $\widehat{\beta}$ follows immediately from Lemma 6(ii) and the fact that $Q_{ZZ} C_B$ has full rank. The next two claims which are adapted from NS (see their Lemmas A2 and A3) establish the desired result. For ease of notation, in the next claim I write $\theta = (\alpha, \beta)$ for a sequence $\theta_n = (\alpha_n, \beta_n)$.

Claim 1: Let $\theta = (\alpha, \beta) \in \Theta$ be a sequence for which $\widehat{g}(\alpha, \beta) = O_p(n^{-1/2})$. Then $\lambda(\theta) := \arg\max_{\lambda \in \widehat{\Lambda}_n(\theta)} \widehat{P}(\theta, \lambda)$ exists wpa1, $\lambda(\theta) = O_p(n^{-1/2})$, and $\sup_{\lambda \in \widehat{\Lambda}_n(\theta)} \widehat{P}(\theta, \lambda) = O_p(n^{-1})$.

Proof: The same proof as for Lemma 10 can be used where now I appeal to Lemma 6 rather than Lemma 1. The last statement follows from equation (4.2).

Claim 2: $||\widehat{g}(\widehat{\theta})|| = O_p(n^{-1/2})$.

Proof: Define $\underline{\lambda}:= -n^{-1/2}\widehat{g}(\widehat{\theta})/||\widehat{g}(\widehat{\theta})||$. By Lemma 9, $\max_{1\leq i\leq n} |\underline{\lambda}' g_i(\widehat{\theta})| \to_p 0$ and wpa1 $\underline{\lambda}\in \widehat{\Lambda}_n(\widehat{\theta})$. By a second order Taylor expansion around $\lambda = 0$, there is $\widetilde{\lambda}$ on the line segment $\overline{0\underline{\lambda}}$, such that for some positive constants C_1 and C_2

$$\begin{aligned}\widehat{P}(\widehat{\theta},\underline{\lambda}) &= -2\underline{\lambda}'\widehat{g}(\widehat{\theta}) + \underline{\lambda}'[\sum_{i=1}^n \rho_2(\widetilde{\lambda}' g_i(\widehat{\theta}))g_i(\widehat{\theta})g_i(\widehat{\theta})'/n]\underline{\lambda} \\ &\geq 2n^{-1/2}||\widehat{g}(\widehat{\theta})|| - C_1\underline{\lambda}'[\sum_{i=1}^n g_i(\widehat{\theta})g_i(\widehat{\theta})'/n]\underline{\lambda} \\ &\geq 2n^{-1/2}||\widehat{g}(\widehat{\theta})|| - C_2 n^{-1}, \end{aligned} \tag{4.9}$$

where the first inequality follows by the fact that wpa1 for $i = 1, ..., n$ $\rho_2(\widetilde{\lambda}' g_i(\widehat{\theta})) \geq -1.5$. The last inequality follows by the uniform convergence result in Lemma 6(i), which implies that wpa1 the largest eigenvalue of $n^{-1}\sum_{i=1}^n g_i(\widehat{\theta})g_i(\widehat{\theta})'$ is bounded above. By the CLT we can apply Claim 1 to the constant sequence (α_0, β_0). The definition of $\widehat{\theta}$ then implies that

$$\widehat{P}(\widehat{\theta},\underline{\lambda}) \leq \sup_{\lambda\in\widehat{\Lambda}_n(\widehat{\theta})} \widehat{P}(\widehat{\theta},\lambda) \leq \sup_{\lambda\in\widehat{\Lambda}_n(\theta_0)} \widehat{P}(\theta_0,\lambda) = O_p(n^{-1}),$$

which together with (4.9) implies the desired result of the claim

$$n^{-1/2}||\widehat{g}(\widehat{\theta})|| = O_p(n^{-1}).$$

Next I establish $n^{1/2}$-consistency for $\widehat{\beta}$, following a standard procedure, see the proof of Theorem 3.2 in NS. By previous arguments we know that the FOC

$$n^{-1}\sum_{i=1}^n \rho_1(\lambda' g_i(\theta))g_i(\theta) = 0 \tag{4.10}$$

has to hold in $(\widehat{\theta},\widehat{\lambda})$ wpa1, where $\widehat{\lambda} = \widehat{\lambda}(\widehat{\theta})$ and $\widehat{\lambda}(\theta)$, for given $\theta\in\Theta$, has been defined above as the arg sup in (2.3). Lemma 6(i) and Lemma 9 imply that $n^{-1}\sum_{i=1}^n \rho_2(\widehat{\lambda}' g_i(\theta))$ $g_i(\theta)g_i(\theta)'$ converges uniformly to $-\Omega_\beta(\theta)$ which is uniformly negative definite and thus nonsingular. Therefore, wpa1 the implicit function theorem implies that there is a neighborhood of $\widehat{\theta}$ where the solution $\widehat{\lambda}(\theta)$ to the FOC is continuously differentiable. I can therefore apply the envelope theorem to get

$$n^{-1}\sum_{i=1}^n \rho_1(\widehat{\lambda}' g_i(\widehat{\theta}))(\partial g_i/\partial\theta)'(\widehat{\theta})\widehat{\lambda} = 0. \tag{4.11}$$

A mean-value expansion of (4.10) about $(\theta,\lambda) = (\theta_0, 0)$ yields (where $g_i(\theta)$ inside ρ_1 is kept constant at $g_i(\widehat{\theta})$)

$$-\widehat{g}(\theta_0) + n^{-1}\sum_{i=1}^n [\rho_1(\overline{\lambda}' g_i(\widehat{\theta}))(\partial g_i/\partial\theta)(\overline{\theta})(\widehat{\theta}-\theta_0) + \rho_2(\overline{\lambda}' g_i(\widehat{\theta}))g_i(\overline{\theta})g_i(\widehat{\theta})'\widehat{\lambda}] = 0, \tag{4.12}$$

where $(\overline{\theta},\overline{\lambda})$ are mean-values between $(\theta_0, 0)$ and $(\widehat{\theta},\widehat{\lambda})$ that may be different for each row. Combining the p rows of (4.11) with the k rows of (4.12) I get

$$\begin{pmatrix} 0 \\ -\widehat{g}(\theta_0) \end{pmatrix} + M \begin{pmatrix} \widehat{\theta}-\theta_0 \\ \widehat{\lambda} \end{pmatrix} = 0, \tag{4.13}$$

where the $(p+k)\times(p+k)$ matrix M has been implicitely defined. Note that $(\partial g_i/\partial\theta)(\theta) = -Z_i(Z_i'\Pi_n + V_i')$ and by Assumption 1' we thus get

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1(\overline{\lambda}'g_i(\widehat{\theta}))(\partial g_i/\partial\theta)(\overline{\theta}) \to_p (0, Q_{ZZ}C_B). \tag{4.14}$$

Also by Lemma 6(i) and consistency of $\widehat{\beta}$

$$n^{-1}\sum_{i=1}^{n}\rho_2(\overline{\lambda}'g_i(\widehat{\theta}))g_i(\overline{\theta})g_i(\widehat{\theta})' \to_p -\Omega_\beta((\overline{\alpha},\beta_0),(\widehat{\alpha},\beta_0)), \tag{4.15}$$

the latter matrix being nonsingular. Denote by $M(\alpha)$ the $(p_2+k)\times(p_2+k)$ submatrix of M that corresponds to the parameters $\widehat{\beta}$ and $\widehat{\lambda}$. If by $\overline{M(\alpha)}$ we denote its probability limit, equations (4.14) and (4.15) imply that

$$\overline{M(\alpha)} = \begin{pmatrix} 0 & C_B'Q_{ZZ} \\ Q_{ZZ}C_B & -\Omega_\beta((\overline{\alpha},\beta_0),(\widehat{\alpha},\beta_0)) \end{pmatrix}.$$

Writing Ω_α for $-\Omega_\beta((\overline{\alpha},\beta_0),(\widehat{\alpha},\beta_0))$, it follows that

$$\overline{M(\alpha)}^{-1} = \begin{pmatrix} -\Sigma_\alpha & H_\alpha' \\ H_\alpha & P_\alpha \end{pmatrix},$$

where

$$\begin{aligned} &\Sigma_\alpha := (C_B'Q_{ZZ}\Omega_\alpha^{-1}Q_{ZZ}C_B)^{-1},\ H_\alpha' := \Sigma_\alpha C_B'Q_{ZZ}\Omega_\alpha^{-1},\ \text{and} \\ &P_\alpha := \Omega_\alpha^{-1} - \Omega_\alpha^{-1}Q_{ZZ}C_B\Sigma_\alpha C_B'Q_{ZZ}\Omega_\alpha^{-1}. \end{aligned} \tag{4.16}$$

Therefore, $M(\alpha)$ is nonsingular wpa1. Equation (4.13) then implies that

$$n^{1/2}(\widehat{\beta}' - \beta_0', \widehat{\lambda}')' = -M(\alpha)^{-1}(0', -n^{1/2}\widehat{g}(\theta_0)')' = -\overline{M(\alpha)}^{-1}(0', -n^{1/2}\widehat{g}(\theta_0)')' + o_p(1).$$

In particular,

$$n^{1/2}(\widehat{\beta} - \beta_0) = -H_\alpha' n^{1/2}\widehat{g}(\theta_0) + o_p(1). \tag{4.17}$$

Applying the CLT to $n^{1/2}\widehat{g}(\theta_0)$ and noting that $H_\alpha = O_p(1)$ concludes the proof. Note that the matrix H_α is random because it depends on $\widehat{\alpha}$ and $\overline{\alpha}$. This dependence prevents $\widehat{\beta}$ from being asymptotically normal. This establishes part (i) of the theorem.

(ii) From part (i) $\forall\varepsilon > 0\ \exists M_\varepsilon < \infty$ s.t. $\Pr(n^{1/2}||\beta_0 - \widehat{\beta}|| < M_\varepsilon) > 1-\varepsilon$ for all n big enough. Let $B(M)$ be a closed ball of radius M centered at zero. Because uniformly over $(\alpha, b) \in A \times B(M_\varepsilon)$ we have $||\widehat{g}(\widehat{\theta}) - \widehat{g}(\alpha, \beta_0 + n^{-1/2}b)|| = O_p(n^{-1/2})$, Lemma 10 implies that $\widehat{g}(\alpha, \beta_0 + n^{-1/2}b)$ is itself $O_p(n^{-1/2})$ uniformly over $(\alpha, b) \in A \times B(M_\varepsilon)$. Set $\theta_{\alpha b} := (\alpha, \beta_0 + n^{-1/2}b)$. Using the same proofsteps as in Claim 1 once more, it can be shown that $\lambda(\theta_{\alpha b}) := \arg\max_{\lambda\in\widehat{\Lambda}_n(\theta_{\alpha b})}\widehat{P}(\theta_{\alpha b}, \lambda)$ exists uniformly wpa1, and $\lambda(\theta_{\alpha b}) = O_p(n^{-1/2})$ uniformly over $A \times B(M_\varepsilon)$. As in the beginning of the proof of Theorem 2 above, using the same notation, we can thus expand the FOC for a

maximum in λ to obtain $\lambda(\theta_{\alpha b}) = \widehat{\Omega}^{-1}_{\widetilde{\lambda}\theta_{\alpha b}} \widehat{g}(\theta_{\alpha b})$ and upon inserting this into a second order Taylor expansion of $\widehat{P}(\theta, \lambda)$ we get

$$\widehat{P}(\theta_{\alpha b}, \lambda(\theta_{\alpha b})) = -2\widehat{g}(\theta_{\alpha b})'\widehat{\Omega}^{-1}_{\widetilde{\lambda}\theta_{\alpha b}} \widehat{g}(\theta_{\alpha b}) + \widehat{g}(\theta_{\alpha b})'\widehat{\Omega}^{-1}_{\widetilde{\lambda}\theta_{\alpha b}} \widehat{\Omega}_{\overline{\lambda}\theta_{\alpha b}} \widehat{\Omega}^{-1}_{\widetilde{\lambda}\theta_{\alpha b}} \widehat{g}(\theta_{\alpha b}).$$

The matrices $\widehat{\Omega}_{\overline{\lambda}\theta_{\alpha b}}$ and $\widehat{\Omega}_{\widetilde{\lambda}\theta_{\alpha b}}$ converge uniformly to $-\Omega(\alpha, \beta_0)$. Also note that Lemma 1(ii) implies that

$$n^{1/2}\widehat{g}(\theta_{\alpha b}) = n^{1/2}\widehat{g}((\alpha, \beta_0 + b)) \Rightarrow \Psi((\alpha, \beta_0 + b)) + Q_{ZZ}C((\alpha_0 - \alpha), -b),$$

and therefore that

$$n\widehat{P}(\theta_{\alpha b}, \lambda(\theta_{\alpha b})) \Rightarrow P((\alpha, \beta_0 + b), (\alpha, \beta_0)).$$

If $P((\alpha, \beta_0 + b), (\alpha, \beta_0))$ has a unique minimum, it follows from Lemma 3.2.1 in van der Vaart and Wellner (1996, p.286) that

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

Proof of Corollary 8. Let $\mu := Q_{ZZ}C(\alpha_0 - \alpha, -b)$. $\Psi((\alpha, \beta_0 + b)) + \mu$ is distributed as $N(\mu, \Omega(\alpha, \beta_0 + b))$. Also from the definition of $\Omega(\theta)$ in (2.7) it follows that $\Omega(\alpha, \beta_0)^{-1/2}N(\mu, \Omega(\alpha, \beta_0 + b))$ is distributed as $N(\Omega(\alpha, \beta_0)^{-1/2}\mu, \eta I_k)$. By (2.8) and Theorem 7, this proves the corollary. $\square$

References

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

Andrews, D. W. K. (1994): "Empirical process methods in Econometrics", in *Handbook of Econometrics*, Vol.4, ed. by R. Engle and D.McFadden. Amsterdam: North Holland, 2247-2294.

Hansen, L. P. (1982): "Large sample properties of Generalized Method of Moment estimators", *Econometrica* 50(4), 1029-1054.

Hansen, L. P., J. Heaton, and A. Yaron (1996): "Finite-sample properties of some alternative GMM estimators", *Journal of Business and Economic Statistics* 14(3), 262-280.

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

Kleibergen, F. (2001): "Testing parameters in GMM without assuming that they are identified", working paper.

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

Moreira, M. J. (2002): "A conditional likelihood ratio test for structural models", working paper.

Nelson, C. R., and R. Startz (1990): "Some further results on the exact small sample properties of the instrumental variables estimator", *Econometrica* 58(4), 967-976.

Newey, W. K., and R. J. Smith (2001): "Higher order properties of GMM and Generalized Empirical Likelihood estimators", working paper.

Owen, A. (1988): "Empirical Likelihood ratio confidence intervals for a single functional", *Biometrika* 75(2), 237-249.

Owen, A. (1990): "Empirical Likelihood ratio confidence regions", *Annals of Statistics* 18(1), 90-120.

Pakes, A., and D. Pollard (1989): "Simulation and the asymptotics of optimization estimators", *Econometrica* 57(5), 1027-1057.

Phillips, P. C. B. (1989): "Partially identified Econometric models", *Econometric Theory* 5, 181-240.

Staiger D., and J. H. Stock (1997): "Instrumental variables regression with weak instruments", *Econometrica* 65(3), 557-586.

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

van der Vaart, A. W., and J. A. Wellner (1996): "Weak convergence and empirical processes", New York: Springer.

Wooldridge, J. (2002): "Econometric analysis of cross section and panel data", The MIT Press, Cambridge, Massachusetts.

TABLE I

Actual size of various test statistics of nominal 5% hypothesis tests

(1) $\Pi_1 = 1.0$ and $k = 5$.

Test\Design	I	II(i)	II(ii)	III	IV
$2SLS$	9.3	6.9	8.2	4.7	4.3
AR	5.8	5.8	6.5	5.7	6.1
A_{CUE}	3.7	2.6	3.3	3.0	4.5
A_{EL}	10.2	21.1	15.3	15.6	8.1
A_{ET}	8.9	13.8	11.3	11.4	7.9
LR_M	6.3	6.3	6.9	6.2	6.6
K	4.8*	5.2	5.5*	5.5	5.3*
K^W_{CUE}	3.0	3.9	3.5	3.7	3.2
K^W_{EL}	4.2	5.0*	4.5*	4.9*	4.5
K^L_{EL}	8.8	19.0	14.2	15.3	6.7
K^W_{ET}	3.6	4.3	4.0	4.2	3.8
K^L_{ET}	8.6	15.4	12.7	12.5	7.1

Notes: In this table and also in tables (ii)-(iv) below a star "*" in the size colum indicates the value closest to the 5% nominal value. The sample size is $n = 100$, and $10,000$ samples are used.

(2) $\Pi_1 = 1.0$ and $k = 1$.

Test\Design	I	II(i)	II(ii)	III	IV
$2SLS$	5.2*	5.0*	5.2*	5.3*	4.8
AR	5.2*	5.3	5.5	5.6	5.3
A_{CUE}	4.5	4.1	4.4	4.5	5.0*
A_{EL}	5.4	9.8	7.4	7.8	5.5
A_{ET}	5.4	7.7	6.7	7.0	5.6
LR_M	5.4	5.5	5.6	5.8	5.4
K	5.2*	5.3	5.5	5.6	5.3
K^W_{CUE}	4.5	4.1	4.4	4.5	5.0*
K^W_{EL}	4.5	4.1	4.4	4.5	5.0*
K^L_{EL}	6.5	17.1	11.3	12.0	5.5
K^W_{ET}	4.5	4.1	4.4	4.5	5.0*
K^L_{ET}	7.2	15.3	11.0	11.2	6.5

(3) $\Pi_1 = .1$ and $k = 5$.

Test\Design	I	II(i)	II(ii)	III	IV
$2SLS$	92.2	46.5	68.7	1.8	0.8
AR	5.8	5.8	6.5	5.7	6.1
A_{CUE}	3.7	2.6	3.3	3.0	4.5
A_{EL}	10.2	21.1	15.3	15.6	8.1
A_{ET}	8.9	13.8	11.3	11.4	7.9
LR_M	6.3	6.3	6.9	6.2	6.6
K	4.8*	5.7	6.0	5.4	5.8
K^W_{CUE}	3.2	3.7	3.8	3.7	3.8
K^W_{EL}	4.3	6.1	5.4*	5.0*	5.1*
K^L_{EL}	9.1	24.1	16.6	15.2	7.2
K^W_{ET}	3.7	4.4*	4.3	4.1	4.3
K^L_{ET}	8.7	16.0	12.9	12.4	7.7

(4) $\Pi_1 = .1$ and $k = 1$.

Test\Design	I	II(i)	II(ii)	III	IV
$2SLS$	18.3	8.1	13.6	0.3	0.2
AR	5.2*	5.3*	5.5*	5.6	5.3
A_{CUE}	4.5	4.1	4.4	4.5*	5.0*
A_{EL}	5.4	9.8	7.4	7.8	5.5
A_{ET}	5.4	7.7	6.7	7.0	5.6
LR_M	5.4	5.5	5.6	5.8	5.4
K	5.2*	5.3*	5.5*	5.6	5.3
K^W_{CUE}	4.5	4.1	4.4	4.5*	5.0*
K^W_{EL}	4.5	4.1	4.4	4.5*	5.0*
K^L_{EL}	6.5	17.1	11.3	12.0	5.5
K^W_{ET}	4.5	4.1	4.4	4.5*	5.0*
K^L_{ET}	7.2	15.3	11.0	11.2	6.5