

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

Patrik Guggenberger, Yale University

This version: November 2002; First version: June 2002  
Preliminary and Incomplete, Comments very welcome

## 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  $\chi^2$ , the use of  $\chi^2$  critical values can lead to extreme size-distortions in finite samples in weakly identified situations. In this paper, I therefore propose new test statistics whose sizes are robust to the strength or weakness of identification. In fact, the asymptotic null distribution of these statistics is  $\chi^2$  under partial (Phillips (1989)), weak (Staiger and Stock (1997)), and strong identification.

Both test statistics are based on Generalized Empirical Likelihood (GEL) methods. The first statistic  $GELR_\rho$  is constructed from the criterion function of the GEL estimator. The second statistic  $K_\rho^W$  generalizes the  $K$  statistic in Kleibergen (2001) from Generalized Method of Moments (GMM) to GEL. This statistic is given by a quadratic form in the first-order condition of the GEL estimator evaluated at the true parameter value. I show how both statistics can be modified to test a 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 conditional heteroskedasticity and certain features of the error distribution like asymmetry and thick tails. In over-identified problems, the statistic  $K_\rho^W$  is generally more powerful than  $GELR_\rho$ .

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 Phillips (1989) and Stock and Wright (2000) for 2SLS and GMM, the resulting estimators have non-standard asymptotic distributions and are in general inconsistent. Therefore, inference based on the classical normal approximation is inappropriate under weak identification.

*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, and  $\lambda_{\min}(A)$  denotes the smallest eigenvalue of  $A$ . 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  $\text{vec}(M)$  stands for the column vectorization of the  $k \times p$  matrix  $M$ , i.e. if  $M = (m_1, \dots, m_p)$  then  $\text{vec}(M) = (m'_1 \dots m'_p)'$ .

## 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 hypotheses 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

$$y = Y\theta_0 + X\gamma_0 + u, \quad (2.1)$$

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  $\theta_0$ .

From now on, I assume wlog CHECK THIS 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,  $\hat{\theta}_{2SLS} := (Y'P_Z Y)^{-1} Y'P_Z y$  is a consistent and asymptotically normal estimator for  $\theta_0$ . 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  $\chi^2$  random variable 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 (S), strong identification, or (W), weak identification, holds, where:

(S)  $\xi = 0$  and  $C$  is of full column rank,

(W)  $\xi = 1/2$ .

Note that Assumption 1(W) 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.

Assumption 1(W) 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  $\Pi_n$  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.

In the following, by  $Y_i, V_i, Z_i, \dots$  for  $i = 1, \dots, n$ , I denote the  $i^{\text{th}}$  row of the matrix  $Y, V, Z, \dots$  written as a column vector. By  $Y_{ij}, V_{ij}, \dots$  I denote the  $j^{\text{th}}$  component of the vectors  $Y_i, V_i, \dots$

For given sample size  $n$ , define the random  $k$ -vector

$$g_{ni}(\theta) := (y_i - Y_i'\theta)Z_i. \quad (2.2)$$

I usually write  $g_i(\theta)$  for  $g_{ni}(\theta)$ . Under both cases in Assumption 1, we can write

$$g_i(\theta) = (u_i + (n^{-\xi}Z_i'C + V_i')(\theta_0 - \theta))Z_i. \quad (2.3)$$

Define the  $k \times k$  matrix

$$\Omega(\theta_1, \theta_2) := \lim_{n \rightarrow \infty} E g_i(\theta_1) g_i(\theta_2)' \text{ and write } \Omega(\theta) \text{ for } \Omega(\theta, \theta). \quad (2.4)$$

The existence of  $\Omega(\theta_1, \theta_2)$  is justified by the following moment and distributional assumptions. Let  $U := (u, V)$ .

**Assumption 2**

- (i)  $\{(U_i, Z_i) : i \geq 1\}$  are iid,
- (ii)  $E Z_i U_i' = 0$ ,
- (iii)  $E \|Z_i\|^4 < \infty$ ,  $E(Z_i Z_i') = Q_{ZZ} > 0$ ;  $E u_i^2 Z_i Z_i'$ ,  $E u_i V_{ij} Z_i Z_i'$ , and  $E V_{ij} V_{ik} Z_i Z_i'$  exist and are finite for  $j, k = 1, \dots, p$ ,
- (iv)  $\theta_0$  is in the interior of the compact set  $\Theta \subset R^p$ .

We introduce two versions of the next assumption, namely Assumption 3(W), for weak asymptotics, and Assumption 3(S), for strong asymptotics.

**Assumption 3**

- (S)  $\Omega(\theta_0) = E u_i^2 Z_i Z_i' > 0$ ,
- (W)  $\inf_{\theta \in \Theta} \lambda_{\min}(\Omega(\theta)) > 0$ .

Assumption (HOM) “conditional homoskedasticity”

**(HOM)**  $E(U_i U_i' | Z_i) = \Sigma_{UU} > 0$

is sufficient for Assumption 3, i.e. Assumptions 1( $\cdot$ ), 2, and (HOM) imply 3( $\cdot$ ), where  $\cdot$  equals either W or S.<sup>1</sup> Assumption (HOM) is used in Staiger and Stock (1997). Our assumptions are more general because they allow for cases of conditional heteroskedasticity. 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 our assumptions.

The GEL estimator  $\hat{\theta}$  of  $\theta_0$  in (2.1) exploits the moment condition  $E g_i(\theta_0) = 0$ , implied by Assumption 2. It is given by

$$\hat{\theta} := \arg \min_{\theta \in \Theta} \sup_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta, \lambda),^2 \quad (2.5)$$

---

<sup>1</sup>To prove this, note that under Assumption (HOM)  $\Omega(\theta_0) = E E(u_i^2 | Z_i) Z_i Z_i' = E(u_i^2 | Z_i) Q_{ZZ} > 0$  because  $Q_{ZZ} > 0$  and  $\Sigma_{UU} > 0$ . This proves Assumption 3(S).

Under Assumption 1(W), the terms in  $\Omega(\theta)$  that are crossproducts involving  $n^{-1/2} Z_i' C$  terms are zero, because for  $\theta \in \Theta$ ,  $E u_i n^{-1/2} Z_i' C(\theta_0 - \theta) Z_i Z_i' = O(n^{-1/2})$ ,  $E n^{-1} (Z_i' C(\theta_0 - \theta))^2 Z_i Z_i' = O(n^{-1})$ , and  $E n^{-1/2} Z_i' C(\theta_0 - \theta) V_i'(\theta_0 - \theta) Z_i Z_i' = O(n^{-1/2})$ , by Assumption 2. Then (HOM) implies that  $\Omega(\theta) = E[u_i + V_i'(\theta_0 - \theta)]^2 Z_i Z_i' = (1, (\theta_0 - \theta)') \Sigma_{UU} (1, (\theta_0 - \theta)')' Q_{ZZ}$  which is a positive definite matrix uniformly over  $\theta \in \Theta$ .

<sup>2</sup>For  $\Theta$  compact,  $\rho$ , and each  $g_i$  continuous it can be shown that an argmin  $\hat{\theta}$  really exists. In fact,  $\sup_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta, \lambda)$ , viewed as a function in  $\theta$ , can be shown to be lower semicontinuous (ls). A function  $f(x)$  is ls at  $x_0$  if for each real number  $c$  such that  $c < f(x_0)$  there exists an open neighborhood  $U$  of  $x_0$  such that  $c < f(x)$  holds for all  $x \in U$ . The function  $f$  is said to be ls if it is ls at each  $x_0$  of its domain. It is easily shown that ls functions on compact sets take on their minimum. Uniqueness of  $\hat{\theta}$ , however, is not implied. As a simple example, 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  $i = 1, \dots, n$ . In that case, if  $\hat{\theta} \in \Theta$  is an

where

$$\widehat{P}(\theta, \lambda) := 2 \sum_{i=1}^n \rho(\lambda' g_i(\theta)) / n - 2\rho_0. \quad (2.6)$$

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

**Assumption 4**

- (i)  $\rho$  is strictly concave on  $Q$ .
- (ii)  $\rho$  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  $\rho(v) = -(1+v)^2/2$ ,  $\rho(v) = \ln(1-v)$ , and  $\rho(v) = -\exp v$ . The EL and ET estimators were introduced by Owen (1988, 1990) and Kitamura and Stutzer (1997), respectively.

**2.2 Asymptotics for GEL estimators under Weak Identification**

In the following I derive the asymptotic distribution of  $\widehat{\theta}$  under weak identification, i.e. under Assumption 1(W). It is instructive to examine a simple case first, namely the case where  $\rho$  is quadratic. In that case,  $Q = R$ , and a second order Taylor expansion in  $\lambda$  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  $\lambda$  is unconstrained. The latter implies that for

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

and

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

we have

$$\widehat{P}(\theta, \lambda) = -2\widehat{g}(\theta)' \lambda - \lambda' \widehat{\Omega}(\theta) \lambda. \quad (2.7)$$

By concavity of  $\widehat{P}(\theta, \lambda)$  in  $\lambda$ , any solution  $\lambda(\theta)$  to the FOC  $0 = -\widehat{g}(\theta) - \widehat{\Omega}(\theta) \lambda$  maximizes  $\widehat{P}(\theta, \lambda)$  with respect to  $\lambda$  for fixed  $\theta$ . 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  $\rho$  is thus given by

$$\widehat{P}(\theta, \lambda(\theta)) = \widehat{g}(\theta)' \widehat{\Omega}(\theta)^- \widehat{g}(\theta). \quad (2.8)$$

---

argmin vector of  $\sup_{\lambda \in \widehat{\Lambda}_n(\theta)} \widehat{P}(\theta, \lambda)$ , then each  $\bar{\theta} \in \Theta$  with  $\bar{\theta}_1 + \bar{\theta}_2 = \widehat{\theta}_1 + \widehat{\theta}_2$  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).

The previous argument was used in NS to show that for quadratic  $\rho$  the GEL estimator formally resembles the GMM Continuous-Updating estimator defined in Hansen, Heaton, and Yaron (1996)<sup>3</sup>. Both estimators minimize a quadratic form whose weighting matrix is continuously altered as  $\theta$  changes. However, in the latter case the weighting matrix is the inverse of a consistent estimate of the covariance matrix of  $\hat{g}(\theta)$  while in the former it is usually not. In general, only in the case  $\theta = \theta_0$ , the matrix  $\hat{\Omega}(\theta)$  consistently estimates the covariance matrix of  $\hat{g}(\theta)$ . 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<sub>GMM</sub> for the GEL and GMM Continuous-Updating estimator, respectively.

I now derive the asymptotic distribution of  $\arg \min \hat{P}(\theta, \lambda(\theta))$  under local to zero asymptotics. The next lemma establishes the probability limit of  $n^{1/2}\hat{g}(\theta)$  under weak asymptotics.

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

Furthermore, Lemma 9 in the Appendix implies that  $\hat{\Omega}(\theta)$  converges uniformly to the matrix  $\Omega(\theta)$  which, under Assumption 3(W), is uniformly positive definite. Therefore  $\hat{\Omega}(\theta)$  is invertible wpa1. It follows that  $n\hat{P}(\theta, \lambda(\theta)) \Rightarrow P(\theta)$ , where

$$\begin{aligned} P(\theta, \bar{\theta}) &:= [\Psi(\theta) + Q_{ZZ}C(\theta_0 - \bar{\theta})]' \Omega(\theta)^{-1} [\Psi(\theta) + Q_{ZZ}C(\theta_0 - \bar{\theta})], \\ P(\theta) &:= P(\theta, \theta). \end{aligned} \quad (2.9)$$

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

$$\hat{\theta}_{CUE} \rightarrow_d \arg \min_{\theta \in \Theta} P(\theta).$$

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<sub>GMM</sub>. 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<sub>GMM</sub>.

I now deal with the asymptotic distribution of GEL estimators for general  $\rho$ . For nonquadratic  $\rho$  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  $\sup_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta, \lambda)$  in (2.5) is actually a maximum wpa1. Then the following definition is justified (at least wpa1):

$$\lambda(\theta) := \arg \max_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta, \lambda). \quad (2.10)$$

---

<sup>3</sup>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.

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(W), 2, 3(W), and 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  $\widehat{\theta}_{GEL} \rightarrow_d \arg \min_{\theta \in \Theta} P(\theta)$ .*

**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  $\rho$  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 immediately implies that  $E[\sup_{\theta \in \Theta} \|g_i(\theta)\|^2] < \infty$ . 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

In this subsection, we want to introduce new statistics to test the simple hypothesis  $H_0 : \theta = \theta^*$  versus  $H_1 : \theta \neq \theta^*$ , or to construct confidence intervals for  $\theta_0$ . 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 test statistics or confidence intervals for  $\theta_0$ .

When EL methods were introduced in the late eighties by Owen (1988), they were first used to construct confidence regions for means of *iid* random variables. Our first test statistic is based on a direct generalization to GEL of the well known EL result that  $-2 \ln R(\theta_0)$  converges in distribution to a chi squared random variable, where  $R(\theta) := \sup_{\pi_1, \dots, \pi_n} \{ \prod_{i=1}^n n\pi_i \mid \sum_{i=1}^n \pi_i g(X_i, \theta) = 0, \pi_i > 0, \sum_{i=1}^n \pi_i = 1 \}$  is the empirical likelihood ratio (see Owen (1988) p.237, 238 or Owen (1990)).

Define

$$GELR_\rho(\theta) := n\widehat{P}(\theta, \lambda(\theta)). \quad (2.11)$$

For CUE, equation (2.8) shows that when  $\theta \neq \theta_0$ , we get  $GELR_\rho(\theta) \rightarrow_d \chi^2(k)$  for  $\xi > 1/2$  in Assumption 1, and  $GELR_\rho(\theta)$  diverges to infinity at a rate of  $n^{1-2\xi}$  for

$\xi < 1/2$ . I thus derive the asymptotic distribution of  $GELR_\rho(\theta_0)$  for  $\xi = 0$  under a local rather than a fixed alternative. We therefore introduce the following addition to Assumption 1, the local alternative case.

**Assumption 1(S-LA)** Assumption 1(S) holds and for some fixed  $d \in R^p$ , we have  $y = Y(\theta_0 + n^{-1/2}d) + u$ .

**Corollary 3** *Suppose Assumptions 2 and 4.*

(i) *Under Assumptions 1(S-LA), and 3(S), we have*

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

where the noncentrality parameter  $\delta$  is given by  $\delta := \|\Omega(\theta_0)^{-1/2}Q_{ZZ}Cd\|^2$ .

(ii) *Under Assumptions 1(W), and 3(W), for fixed  $\theta \in \Theta$ , we have*

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

where the noncentrality parameter  $\delta$  is given by  $\delta := \|\Omega(\theta)^{-1/2}Q_{ZZ}C(\theta_0 - \theta)\|^2$ .

In particular, if  $d = 0$ , then independent of the value of  $\xi \in [0, \infty)$ , we have

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

MENTION TESTS OF IMBENS AND SPADY (2002).

The corollary provides a straightforward method for constructing confidence sets and perform hypothesis tests involving  $\theta_0$  that are asymptotically valid. For example, to test the hypothesis  $H_0 : \theta = \theta^*$  versus  $H_1 : \theta \neq \theta^*$  at significance level  $r$ , reject the hypothesis iff  $GELR_\rho(\theta^*)$  exceeds  $\chi_r^2(k)$ , the  $(1 - r)$   $\chi^2(k)$  critical value. We obtain  $(1 - r)$  confidence regions for  $\theta_0$  by inverting the just described test, i.e. by  $\{\theta \in \Theta \mid GELR_\rho(\theta) \leq \chi_r^2(k)\}$ .

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 is inconsistent. The noncentrality parameter  $\delta$  of the asymptotic  $\chi^2$  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  $\xi$ , the test has correct asymptotic size with weak and strong identification.

For the CUE<sub>GMM</sub> Stock and Wright (2000, Theorem 2) derive the asymptotic distribution of the analogue to  $GELR_\rho(\theta_0)$  under weak asymptotics. Qin and Lawless (1994, Theorem 2) propose the statistic  $2 \ln R(\hat{\theta}) - 2 \ln R(\theta^*)$  to test the hypothesis  $H_0 : \theta = \theta^*$ . They show that under strong identification the statistic is asymptotically distributed as  $\chi^2(p)$ . However, due to its dependence on  $\hat{\theta}$ , the test statistic leads to size-distortion under weak identification. I now propose a test statistic with  $\chi^2(p)$  asymptotic distribution that overcomes the problem of size-distortion under weak identification.

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 over-identified situations. MENTION Kleibergen (2002b). Kleibergen (2001, 2002a) 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  $\theta_0$ .

Applying Kleibergen's (2001) idea to GEL, I construct a quadratic form from the GEL FOC condition for  $\theta$  evaluated at the true value  $\theta_0$ . If the minimum of the objective function  $\widehat{P}(\theta, \lambda(\theta))$  is obtained in the interior of the parameter space  $\Theta$ , the following FOC has to hold

$$\lambda(\theta)' \sum_{i=1}^n \rho_1(\lambda(\theta)' g_i(\theta)) G_i(\theta) / n = 0, \quad (2.12)$$

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 (2.10) above (for a proof see NS, Section 2.2 or equation (4.8) 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 (2.12) can thus be written as  $\lambda(\theta)' D_\rho(\theta)$ .

Under Assumption 1(W), for CUE we have seen above that  $\lambda(\theta) = -\widehat{\Omega}(\theta)^{-1} \widehat{g}(\theta)$  wpa1 and thus by Lemmas 1 and 9  $n^{1/2} \lambda(\theta_0) \rightarrow_d N(0, \Omega(\theta_0)^{-1})$ . 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 expression 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

$$K_\rho^W(\theta) := n \widehat{g}(\theta)' \Omega(\theta)^{-1/2} P_{\Omega(\theta)^{-1/2} D_\rho(\theta)} \Omega(\theta)^{-1/2} \widehat{g}(\theta), \quad (2.13)$$

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

$$K_\rho^L(\theta) := n \lambda(\theta)' \Omega(\theta)^{1/2} P_{\Omega(\theta)^{-1/2} D_\rho(\theta)} \Omega(\theta)^{1/2} \lambda(\theta). \quad (2.14)$$

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(S). In that case, we know from NS that  $\widehat{\theta}$  is  $n^{1/2}$ -consistent. Therefore, if the FOC (2.12) hold at  $\widehat{\theta}$ , then, at least

asymptotically, they also hold at the true value  $\theta_0$ . 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  $\theta_0$ .

Under weak identification, i.e. the case in Assumption 1(W), the argument has to be modified. As proved above,  $\widehat{\theta}$  is no longer consistent for  $\theta_0$ . Therefore, the fact that the FOC hold at  $\widehat{\theta}$  does not imply automatically that they have to hold at the true value  $\theta_0$ , not even approximately or asymptotically. However, as shown in Lemma 11 below, under weak identification the FOC  $n^{-1} \sum_{i=1}^n \rho_1(\lambda' g_i(\theta)) g_i(\theta) = 0$  not only hold at  $\widehat{\theta}$  wpa1 but hold uniformly over  $\theta \in \Theta$  wpa1. Therefore, the FOC is not a condition that asymptotically pins down the true value  $\theta_0$ , but a condition that holds asymptotically for all  $\theta \in \Theta$ . Under weak identification, we therefore should not expect that hypothesis tests for  $\theta_0$  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 Assumption 1(S) and (W). 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, 2002a).

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 identification I thus derive the asymptotic distribution for  $\theta_0$  under a local alternative given in Assumption 1(S-LA).

Note that in the linear model we have  $G_i(\theta) = G_i = -n^{-\xi} Z_i Z_i' C - Z_i V_i' = -Z_i Y_i'$ .

**Corollary 4** *Suppose Assumptions 2 and 4.*

(i) *Suppose Assumptions 1(S-LA), and 3(S). Then we have*

$$K_\rho^W(\theta_0) \rightarrow_d \chi^2(p, \delta),$$

with noncentrality parameter  $\delta := \|\Omega(\theta_0)^{-1/2} Q_{ZZ} C d\|^2$ . In particular, if  $d = 0$ , we have

$$K_\rho^W(\theta_0) \rightarrow_d \chi^2(p).$$

(ii) *Suppose Assumptions 1(W), 3(W), and that  $E(V_i V_i' \otimes Z_i Z_i')$  has full column rank. Fix  $\theta \in \Theta$ . Then,*

$$K_\rho^W(\theta) \rightarrow_d (W(\theta) + \zeta)' (W(\theta) + \zeta),$$

where  $\zeta \sim N(0, I_p)$ , where the nonstandard distribution of the random  $p$ -vector  $W(\theta)$  is defined in (4.17), and where  $\zeta$  and  $W(\theta)$  are independent. Because  $W(\theta_0) \equiv 0$ , for  $\theta = \theta_0$ , we have

$$K_\rho^W(\theta_0) \rightarrow_d \chi^2(p).$$

*The same results hold for the statistic  $K_\rho^L(\theta_0)$ .*

The corollary shows that under weak identification, hypothesis tests based on the statistics  $K_\rho^W(\theta)$  and  $K_\rho^L(\theta)$  are inconsistent. The reason is that for  $\theta \neq \theta_0$ ,  $W(\theta)$  is a

random variable with a certain distribution that does not depend on  $n$ , in particular  $\|W(\theta)\|$  does not converge to infinity for  $n \rightarrow \infty$ .

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

$$\widehat{\Omega}(\theta)^{-1} = \left( \sum_{i=1}^n g_i(\theta)g_i(\theta)' / n \right)^{-1}. \quad (2.15)$$

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. Note that if we use the above estimator  $\widehat{\Omega}(\theta)^{-1}$  then, in the case  $k = p = 1$ ,  $K_\rho^W(\theta)$  reduces to  $n\widehat{g}(\theta)'\widehat{\Omega}(\theta)^{-1}\widehat{g}(\theta) = GELR_{CUE}(\theta)$ .

Kleibergen's  $K$  statistic does not coincide with  $K_\rho^W$  for quadratic  $\rho$ . The  $K$  statistic uses the FOC for  $CUE_{GMM}$  defined in Hansen, Heaton, and Yaron (1996) while  $K_\rho^W$  uses the FOC for GEL for quadratic  $\rho$ . It was already mentioned earlier that the two estimators do in general not coincide.

## 2.4 Some extensions

### 2.4.1 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 under this setup.

**Assumption 1(WS)**  $\Pi_n = (\Pi_A, \Pi_B)$ ,  $C = (C_A, C_B)$  fixed,  $\Pi_A = n^{-1/2}C_A$ ,  $\Pi_B = C_B$ . The matrices  $\Pi_A$  and  $C_A$  are  $R^{k \times p_A}$ , and  $\Pi_B$  and  $C_B$  are  $R^{k \times p_B}$ , where  $p_A + p_B = p$ , and  $p_A$  and  $p_B \geq 1$ . The matrix  $C_B$  has full column rank.

Conformably with  $\Pi_n$ , I write  $Y = (Y_A, Y_B)$ ,  $V_i = (V_{iA}', V_{iB}')'$ ,  $\theta = (\alpha', \beta)'$ ,  $\widehat{\theta} = (\widehat{\alpha}', \widehat{\beta})'$ , and  $\theta_0 = (\alpha_0', \beta_0)'$ . We have to modify Assumption 3 according to the new setup in Assumption 1.

**Assumption 3(WS)**  $\inf_{\alpha \in \{\alpha \in R^{p_A} | (\alpha', \beta_0)' \in \Theta\}} \lambda_{\min}\{\Omega(\alpha', \beta_0)'\} > 0$ .

Assumption 1(WS) 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  $\alpha_0$  as weakly identified and  $\beta_0$  as strongly identified. Assumption 1(WS) contains as a particular case the partially identified model of Phillips (1989). In fact, choosing  $p_A$  and setting  $C_A = 0$ , we obtain a model in which  $C$  has any desired (less than full) rank.

The next theorem establishes the asymptotic behavior of  $\widehat{\theta}$ .

**Theorem 5** *Assume 1(WS), 2, and 4. Assume the following stronger version of Assumption 3(WS),  $\inf_{\alpha_1, \alpha_2 \in \{\alpha \in R^{p_A} | (\alpha', \beta_0)' \in \Theta\}} \lambda_{\min}\{\Omega((\alpha_1', \beta_0)', (\alpha_2', \beta_0)')\} > 0$ . Then*

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

Assume  $\Theta = A \times B$ , where  $A \subset R^{p_A}$  and  $B \subset R^{p_B}$  are compact. Let  $B_{\beta_0} := \{b \in R^{p_B} | \exists b_0 \in B, b = b_0 - \beta_0\}$ . Then, for  $(\alpha, b) \in A \times B_{\beta_0}$ , and  $\theta_{\alpha b} := (\alpha', \beta'_0 + n^{-1/2}b')$

$$(ii) n\widehat{P}(\theta_{\alpha b}, \lambda(\theta_{\alpha b})) \Rightarrow P_{\alpha b} := P((\alpha', \beta'_0)', (\alpha', \beta'_0 + b)').$$

Therefore, assuming that a unique  $\arg \min_{(\alpha, b) \in A \times B_{\beta_0}} P_{\alpha b}$  exists, we have

$$(\widehat{\alpha}', n^{1/2}(\widehat{\beta} - \beta_0)')' \rightarrow_d (\alpha^{*'}, \beta^{*'})' := \arg \min_{(\alpha, b) \in A \times B_{\beta_0}} P_{\alpha b}.$$

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 analogous 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.  $p_A = 0$ , Theorem 5(i) has a more precise formulation. As shown in NS, the estimator  $\widehat{\theta}$  is consistent and asymptotically normal. The distribution of the estimator when  $p_B = 0$  was given above in Theorem 2. Therefore, together with the result in Theorem 5, all possible combinations of  $p_A$  and  $p_B$  are covered.

I now derive the asymptotic distribution of the test statistics  $GELR_\rho$ ,  $K_\rho^L$ , and  $K_\rho^W$  under Assumption 1(WS). We allow for a local alternative in the strongly and a fixed alternative in the weakly identified parameters.

**Assumption 1(WS-LA)** Assumption 1(WS) holds and for some fixed  $b \in R^{p_B}$ , we have  $y = Y(\theta_0 + n^{-1/2}(0', b)') + u$ .

**Corollary 6** Suppose Assumptions 1(WS-LA), 2, 3(WS), and 4 hold. Let  $\alpha \in R^{p_A}$  s.t.  $(\alpha', \beta'_0)' \in \Theta$ .

(i) Then we have

$$GELR_\rho((\alpha', \beta'_0)') \rightarrow_d \chi^2(k, \delta),$$

where the noncentrality parameter  $\delta$  is given by  $\delta := \|\Omega((\alpha', \beta'_0)')^{-1/2} Q_{ZZ} C((\alpha_0 - \alpha)', b')'\|^2$ . In particular, if  $b = 0$  we have

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

(ii) Assume  $E(V_{iA} V'_{iA} \otimes Z_i Z'_i)$  has full column rank. Then we have

$$K_\rho^W((\alpha', \beta'_0)') \rightarrow_d (W(\alpha, \beta_0) + \zeta)'(W(\alpha, \beta_0) + \zeta),$$

where  $\zeta \sim N(0, I_p)$ , where the nonstandard distribution of the random  $p$ -vector  $W(\alpha, \beta_0)$  is defined in (4.17), and where  $\zeta$  and  $W(\alpha, \beta_0)$  are independent. If  $\alpha = \alpha_0$  and  $b = 0$ , we have  $W(\alpha, \beta_0) \equiv 0$ . Therefore, if  $b = 0$  it follows that

$$K_\rho^W(\theta_0) \rightarrow_d \chi^2(p).$$

The same result holds for the statistic  $K_\rho^L$ .

## 2.4.2 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 test statistics  $GELR_\rho$ ,  $K_\rho^W$ , and  $K_\rho^L$  to a setup that allows for inference for subvectors.

Let  $\theta_0 = (\alpha'_0, \beta'_0)'$ , where  $\alpha_0$  and  $\beta_0$  are  $p_A$  and  $p_B$ -dimensional vectors, respectively, where  $p_A + p_B = p$ . We are interested in inference on  $\beta_0$ . The idea is to simply replace  $\alpha_0$  in the test statistics  $GELR_\rho$ ,  $K_\rho^W$ , and  $K_\rho^L$  by an estimate  $\hat{\alpha}$ . Let  $\beta^*$  be a fixed hypothesized  $p_B$ -vector. Then  $Eg_i((\alpha'_0, \beta^{*'})') = 0$  if  $\beta^* = \beta_0$ . Define the GEL estimator  $\hat{\alpha}(\beta^*)$  for  $\alpha_0$  by

$$\hat{\alpha}(\beta^*) := \arg \min_{\alpha \in \{\alpha \in R^{p_A} : (\alpha', \beta^{*'})' \in \Theta\}} \sup_{\lambda \in \hat{\Lambda}_n(\alpha', \beta^{*'})'} \hat{P}((\alpha', \beta^{*'})', \lambda). \quad (2.16)$$

I usually write  $\hat{\alpha}$  for  $\hat{\alpha}(\beta^*)$  if it is clear which  $\beta^*$  is meant. We need that  $\hat{\alpha}$  is consistent for  $\alpha_0$  if  $\beta^* = \beta_0$ . For the subvector versions of the test statistics we therefore have to make the assumption that the parameters that are not involved in the hypothesis test, i.e.  $\alpha_0$ , are strongly identified. On the other hand, the components of the subvector  $\beta_0$  can be weakly or strongly identified. Wlog, we assume that the first  $p_{B_1}$  components of  $\beta_0$  are weakly and the remaining  $p_{B_2}$  components are strongly identified, where  $p_{B_1} + p_{B_2} = p_B$ .

**Assumption 1(S $_\alpha$ )** Let  $\Pi_n = (\Pi_A, \Pi_{B_1}, \Pi_{B_2})$ ,  $C = (C_A, C_B) = (C_A, C_{B_1}, C_{B_2})$  fixed,  $\Pi_A = C_A$ ,  $\Pi_{B_1} = n^{-1/2}C_{B_1}$ , and  $\Pi_{B_2} = C_{B_2}$ . The matrices  $\Pi_A$  and  $C_A$  are  $R^{k \times p_A}$ ,  $\Pi_{B_1}$  and  $C_{B_1}$  are  $R^{k \times p_{B_1}}$ , and  $\Pi_{B_2}$  and  $C_{B_2}$  are  $R^{k \times p_{B_2}}$ , where  $p_A + p_B = p$  and  $p_{B_1} + p_{B_2} = p_B$ . The matrix  $C_A$  has full column rank.

Decompose  $\beta_0 = (\beta'_{01}, \beta'_{02})'$ , where  $\beta_{01}$  and  $\beta_{02}$  are  $p_{B_1}$  and  $p_{B_2}$ -dimensional vectors, respectively. Define  $V_{iA}$  and  $Y_{iA}$  to be the subvectors of  $V_i$  and  $Y_i$ , respectively, that consist of their first  $p_A$  components. Analogously,  $V_{iB}$  and  $Y_{iB}$  denote the last  $p_B$  components of these vectors. Furthermore, I use the following notation. For a  $p_{B_1}$ -vector  $b_1 \in B_1 := \{b \in R^{p_{B_1}} | (\alpha'_0, b', \beta'_{02})' \in \Theta\}$  let

$$\hat{\theta}_{b_1} := (\hat{\alpha}', b'_1, \beta'_{02})' \text{ and } \theta_{b_1} := (\alpha'_0, b'_1, \beta'_{02})'.$$

Similarly, for a  $p_B$ -vector  $b \in B := \{b \in R^{p_B} | (\alpha'_0, b')' \in \Theta\}$  let

$$\hat{\theta}_b := (\hat{\alpha}', b')' \text{ and } \theta_b := (\alpha'_0, b')'.$$

I introduce a modified version of Assumption 3.

**Assumption 3(S $_\alpha$ )**  $\inf_{b_1 \in B_1} \lambda_{\min} \Omega(\theta_{b_1}) > 0$ .

We now state the asymptotics for the  $GELR_\rho$  subvector test statistic under a fixed alternative for the weakly identified components  $\beta_{01}$  of  $\beta_0$  and a local alternative for the strongly identified components  $\beta_{02}$  of  $\beta_0$ .

**Assumption 1(S $_\alpha$ -LA)** Assumption 1(S $_\alpha$ ) holds and for some fixed  $b_2 \in R^{p_{B_2}}$ , we have  $y = Y(\theta_0 + n^{-1/2}(0', 0', b_2)') + u$ .

**Theorem 7** Assume  $1 \leq p_B < p$ . Suppose Assumptions 1(S $_\alpha$ -LA), 2, 3(S $_\alpha$ ), and 4 hold, and fix  $b_1 \in B_1$ .

Then we have

$$GELR_\rho(\widehat{\theta}_{b_1}) \rightarrow_d \chi^2(k - p_A, \delta),$$

where the noncentrality parameter  $\delta$  is given by

$$\delta := \|M_{\Omega(\theta_{b_1})^{-1/2}Q_{ZZ}C_A}\Omega(\theta_{b_1})^{-1/2}Q_{ZZ}C_B((\beta_{01} - b_1)', b_2)'\|^2.$$

In particular, if  $b_2 = 0$  we have

$$GELR_\rho(\widehat{\theta}_{\beta_{01}}) \rightarrow_d \chi^2(k - p_A).$$

I now generalize the statistics  $K_\rho^W$  and  $K_\rho^L$  to the subvector case.

Suppose  $\beta^* = \beta_0$ . We show in the Appendix that  $\lambda(\widehat{\theta}_{\beta_0}) := \arg \max_{\lambda \in \widehat{\Lambda}_n(\widehat{\theta}_{\beta_0})} \widehat{P}(\widehat{\theta}_{\beta_0}, \lambda)$  exists wpa1, and that  $n^{1/2}\lambda(\widehat{\theta}_{\beta_0}) \rightarrow_d N(0, M(\beta_0))$ , where

$$M(\beta^*) := \Omega(\theta_{\beta^*})^{-1/2}M_{\Omega(\theta_{\beta^*})^{-1/2}EG_{A_i}}\Omega(\theta_{\beta^*})^{-1/2} \quad (2.17)$$

and where  $G_{A_i} := (\partial g_i(\theta)/\partial \alpha) = -Z_i Y'_{iA}$ .

The last  $p_B$  columns of the FOC (2.12), evaluated at  $\widehat{\theta}_{\beta_0}$ , read

$$\lambda(\widehat{\theta}_{\beta_0})' \sum_{i=1}^n \rho_1(\lambda(\widehat{\theta}_{\beta_0})' g_i(\widehat{\theta}_{\beta_0})) G_{B_i} / n = 0, \quad (2.18)$$

where  $G_{B_i} := (\partial g_i(\theta)/\partial \beta) = -Z_i Y'_{iB}$ . For  $\widehat{\alpha} = \widehat{\alpha}(\beta^*)$  let

$$D_{\rho\beta^*}(\widehat{\alpha}) := \sum_{i=1}^n \rho_1(\lambda(\widehat{\theta}_{\beta^*})' g_i(\widehat{\theta}_{\beta^*})) G_{B_i} / n.$$

As for the test statistics for the full parameter vector, the following subvector test statistics are motivated by the FOC (2.18).

$$\begin{aligned} K_\rho^L(\widehat{\theta}_{\beta^*}) &:= n\lambda(\widehat{\theta}_{\beta^*})' W_\rho(\widehat{\theta}_{\beta^*}) \lambda(\widehat{\theta}_{\beta^*}), \\ K_\rho^W(\widehat{\theta}_{\beta^*}) &:= n\widehat{g}(\widehat{\theta}_{\beta^*})' \Omega(\widehat{\theta}_{\beta^*})^{-1} W_\rho(\widehat{\theta}_{\beta^*}) \Omega(\widehat{\theta}_{\beta^*})^{-1} \widehat{g}(\widehat{\theta}_{\beta^*}), \\ \text{where } W_\rho(\widehat{\theta}_{\beta^*}) &:= D_{\rho\beta^*}(\widehat{\alpha})(D_{\rho\beta^*}(\widehat{\alpha})' M(\beta^*) D_{\rho\beta^*}(\widehat{\alpha}))^{-1} D_{\rho\beta^*}(\widehat{\alpha})'. \end{aligned}$$

We now state the asymptotics for the subvector test statistics  $K_\rho^L$  and  $K_\rho^W$ .

**Theorem 8** Assume  $1 \leq p_B < p$ . Suppose Assumptions 1( $S_\alpha$ -LA), 2, 3( $S_\alpha$ ), and 4 hold. Assume that  $(C_A, C_{B_2})$  and  $E(V_{iB_1}V'_{iB_1} \otimes Z_iZ'_i)$  have full column rank  $p_A + p_{B_2}$  and  $kp_{B_1}$ , respectively. Fix  $b_1 \in B_1$ .

Then,

$$K_\rho^W(\widehat{\theta}_{b_1}) \rightarrow_d (W(b_1, b_2) + \zeta)'(W(b_1, b_2) + \zeta),$$

where  $\zeta \sim N(0, I_{p_B})$ , where the nonstandard distribution of the random  $p_B$ -vector  $W$  is defined in (4.24), and where  $\zeta$  and  $W$  are independent. If  $b_2 = 0$  and  $b_1 = \beta_{01}$ , we have  $W(b_1, b_2) \equiv 0$ . Therefore, if  $b_2 = 0$  it follows that

$$K_\rho^W(\widehat{\theta}_{\beta_0}) \rightarrow_d \chi^2(p_B).$$

The same results hold for the statistic  $K_\rho^L$ .

To use the result in the theorem for hypothesis testing or the construction of confidence intervals for  $\beta_0$ , we have to replace the unknown quantities  $M(\beta^*)$  and  $\Omega(\widehat{\theta}_{\beta^*})$  in the statistics  $K_\rho^L(\widehat{\theta}_{\beta^*})$  and  $K_\rho^W(\widehat{\theta}_{\beta^*})$  by estimators that are consistent under the null. The matrix  $\Omega(\widehat{\theta}_{\beta_0})$  can be consistently estimated by  $\sum_{i=1}^n g_i(\widehat{\theta}_{\beta_0})g_i(\widehat{\theta}_{\beta_0})'/n$ . Using this estimate and replacing  $EG_{Ai}$  by its sample average, we obtain a consistent estimate of  $M(\beta_0)$ .

If  $M(\beta_0)$  and  $\Omega(\widehat{\theta}_{\beta_0})$  are replaced by these estimators, then the two test statistics  $K_\rho^L(\widehat{\theta}_{\beta_0})$  and  $K_\rho^W(\widehat{\theta}_{\beta_0})$  are again numerically identical in the CUE case,  $\rho(v) = -(1+v)^2/2$ .

Even though it is difficult to derive the asymptotic distribution of the test statistics without assuming strong identification of  $\alpha_0$ , it is obvious that the statistics  $K_\rho^L(\widehat{\theta}_{\beta_0})$  and  $K_\rho^W(\widehat{\theta}_{\beta_0})$  would no longer converge to a  $\chi^2(p_B)$  random variable. The reason is that in general the quantities  $n^{1/2}\lambda(\widehat{\theta}_{\beta_0})$  in  $K_\rho^L(\widehat{\theta}_{\beta_0})$  and  $n^{1/2}\widehat{g}(\widehat{\theta}_{\beta_0})$  in  $K_\rho^W(\widehat{\theta}_{\beta_0})$  no longer converge to a normal distribution because of their dependence on  $\widehat{\alpha}$ , which in direct consequence of Theorem 5 has a nonstandard asymptotic distribution if  $\alpha_0$  is not strongly identified. In that case, the subvector version of Kleibergen's (2002a)  $K$  statistic experiences the same problem and would not have an asymptotic  $\chi^2(p_B)$  either. In fact, for  $CUE_{GMM}$ ,  $\widehat{\alpha}$  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 for  $\beta_0$  that is similar independent of the strength or weakness of identification of  $\alpha_0$ . The main advantage of the subvector test statistics introduced in this section is that they have asymptotically accurate sizes independent of whether  $\beta_0$  is weakly or strongly identified. This property is not shared by the classical tests based on the Wald, LR, and LM statistics. In general, they only have correct sizes asymptotically if the whole vector  $\theta_0$  is strongly identified.

### 3 Monte Carlo Experiment

To assess the finite sample performance of the hypothesis tests introduced in Corollaries 3 and 4, I conduct a Monte Carlo experiment. The data generating process

(DGP) is given by model (2.1)

$$\begin{aligned} y &= Y\theta_0 + u, \\ Y &= Z\Pi + V. \end{aligned} \tag{3.19}$$

Interest focuses on testing the null hypothesis  $H_0 : \theta_0 = 0$  versus the alternative hypothesis  $H_1 : \theta_0 \neq 0$ . There is one endogenous variable,  $p = 1$  and  $Z \sim N(0, I_k \otimes I_n)$  where  $k$  is the number of instruments and  $n$  the sample size. In the just-identified case,  $k = 1$ , let  $\Pi = \Pi_1$  and in the over-identified case,  $k > 1$ , let  $\Pi = (\Pi_1, 0)'$ , for a  $(k - 1)$ -vector of zeros  $0$ ; i.e. in the over-identified case we add on a number of irrelevant instruments.

### 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 Design (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  $\sigma$ . All other designs are constructed from Design (I) by modifying the structural error distribution  $u$ .

In 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  $\chi^2(r)$  random variable independent of  $u_i$  and  $V_i$ , i.e. in Design (II) the error in the structural equation has a  $t$ -distribution with  $r$  degrees of freedom. We take  $r = 2$ .

Design (III) modifies Design (I) by exchanging  $u_i$  by  $u_i^2 - 1$  i.e. in Design (III) 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 density function for  $u_i$  has peaks at  $-2$  and  $+2$ .

In addition, we examine the impact of conditional heteroskedasticity in the data on the performance of the test statistics. In Designs (I<sub>HET</sub>)-(IV<sub>HET</sub>) I therefore modify Designs (I)-(IV) by replacing  $u_i$  by  $u_i = ||Z_i||u_i$ .

### Test statistics

The following test statistics are used.

I calculate three statistics  $GELR_\rho(\theta)$  from (2.11), for  $\rho(v) = -(1+v)^2/2$  (CUE),  $\rho(v) = \ln(1-v)$  (EL), and  $\rho(v) = -\exp v$  (ET).

I calculate three statistics for each of  $K_\rho^W(\theta)$  and  $K_\rho^L(\theta)$  defined in (2.13) and (2.14) with the same choices for  $\rho$  as for  $GELR_\rho(\theta)$  and where  $\Omega(\theta)$  is replaced by the consistent estimator  $\hat{\Omega}(\theta)$ , see (2.15). Recall that for CUE,  $K_\rho^W(\theta)$  and  $K_\rho^L(\theta)$  are then numerically identical.

The asymptotic null distributions of these statistics are given in Corollaries 3 and 4.

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

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

where  $s_{uu}(\theta) := (y - Y\theta)'M_Z(y - Y\theta)/(n - k)$ . Under the null  $AR(\theta)$  has an asymptotic  $\chi^2(k)$  distribution.

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

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

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  $s_{uV}(\theta) := (y - Y\theta)'M_Z Y/(n - k)$ . Under the null  $K(\theta)$  has an asymptotic  $\chi^2(p)$  distribution. In the just-identified case  $k = p = 1$ , the  $AR$  and  $K$  statistics coincide, see Kleibergen (2002a).

I include Moreira's conditional likelihood ratio test (see Section 3 in Moreira (2002) for motivation). For the model (3.19) with only one endogenous variable it can be described as follows. Define

$$LR_M : = \frac{1}{2}[\bar{S}'\bar{S} - \bar{T}'\bar{T} + \{(\bar{S}'\bar{S} + \bar{T}'\bar{T})^2 - 4(\bar{S}'\bar{S}\bar{T}'\bar{T} - (\bar{S}'\bar{T})^2)\}^{1/2}], \quad (3.20)$$

$$\text{where } \bar{S} : = (Z'Z)^{-1/2}S(b_0'\hat{\Lambda}b_0)^{-1/2}, \quad \bar{T} := (Z'Z)^{-1/2}T(a_0'\hat{\Lambda}^{-1}a_0)^{-1/2},$$

$$\text{where } S : = Z'(y - Y\theta_0), \quad T := Z'(y, Y)\hat{\Lambda}^{-1}a_0,$$

$$\text{where } a_0 : = (\theta_0, 1)', \quad b_0 := (1, -\theta_0)', \quad \text{and } \hat{\Lambda} := (y, Y)'M_Z(y, Y)/(n - k).$$

Moreira (2002) suggests a simulation method to find the critical value for  $LR_M$  conditional on  $\bar{T}'\bar{T} = \bar{t}'\bar{t}$ . The method leads to a hypothesis test with exact size in the normal model with known reduced form covariance matrix  $\Lambda$ . The simulation method works as follows:

Simulate  $R$  values from

$$LR := \frac{1}{2}[Q_1 + Q_{k-1} - \bar{T}'\bar{T} + \{(Q_1 + Q_{k-1} + \bar{T}'\bar{T})^2 - 4Q_{k-1}\bar{T}'\bar{T}\}^{1/2}],$$

where  $Q_1$  and  $Q_{k-1}$  are independent realizations from a  $\chi^2(1)$  and  $\chi^2(k - 1)$  distribution, respectively. If  $k = 1$  let  $Q_{k-1} \equiv 0$ . For a fixed size  $\alpha$ , let  $c(\alpha)$  be the  $(1 - \alpha)$  quantile of the  $R$  realizations of  $LR$ . Reject the null, iff  $LR_M > c(\alpha)$ .

Finally, I include two versions of the two-stage least squares Wald statistic, see for example Wooldridge (2002, p. 98, 100), one assuming homoskedastic errors and one that is robust to conditional heteroskedasticity

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

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

where  $\hat{\theta} := (Y'P_Z Y)^{-1}Y'P_Z y$ , and  $W := \hat{\sigma}^2(Y'P_Z Y)^{-1}$ , where  $\hat{\sigma}^2 := (n - k)^{-1} \sum_{i=1}^n \hat{u}_i^2$  and  $\hat{u}_i := y_i - Y_i'\hat{\theta}$ . Finally,  $W_{HET}$  is an estimate of the covariance matrix of  $\hat{\theta}$  that is robust to conditional heteroskedasticity

$$W_{HET} := (n/(n - k))(Y'P_Z Y)^{-2} \left( \sum_{i=1}^n \hat{u}_i^2 (P_Z Y)_i^2 \right).$$

With one endogenous variable, the Wald statistics are asymptotically distributed as  $\chi^2(1)$ .

To calculate  $GELR_\rho(\theta)$ ,  $K_\rho^W(\theta)$ , and  $K_\rho^L(\theta)$  for EL and ET, I have to solve the globally concave maximization problem  $\max_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta, \lambda)$  numerically. To do so I implement a variant of the Newton-Raphson algorithm. I start the algorithm by setting  $\lambda$  equal to the zero vector. In each iteration the algorithm tries several shrinking stepsizes in the search direction and accepts the first one that increases the function value compared to the previous value for  $\lambda$ . This procedure enforces an “uphill climbing”-feature of the algorithm.

### Size comparison

I calculate actual sizes for all the above statistics for data generating processes (DGP) in (3.19)<sup>4</sup> that correspond to all 54 possible combinations of

$$\begin{aligned} n &= 50, 100, 250 \\ k &= 1, 5, 10 \\ \sigma &= 0, .5, .99 \\ \Pi_1 &= .1, 1, \end{aligned}$$

and Designs (I)-(IV) and ( $I_{HET}$ )-(IV $_{HET}$ ). Sizes are calculated at the 5% asymptotic critical values using  $R = 3,000$  samples from the DGP. I also use 3,000 realizations from a  $\chi^2(1)$  and  $\chi^2(k-1)$  distribution to simulate the critical values in Moreira’s  $LR_M$  statistic. For the results that I actually report in the tables below I used 10,000 samples. The cases  $\Pi_1 = .1$  and 1 are referred to as the “weak instrument” and “strong instrument” case, respectively. With the constant  $\sigma$  we vary the degree of endogeneity of  $Y$ . While for  $\sigma = 0$ ,  $Y$  is exogenous,  $Y$  is strongly endogenous for  $\sigma = .99$ . We include the just-identified case,  $k = 1$ , and two over-identified-cases,  $k = 5$  and 10.

I first describe the results for Design (I). TABLE 1 contains size results for all combinations of  $(n, k, \sigma, \Pi_1)$  for all test statistics except  $AR$ ,  $GELR_{ET}$ ,  $K_{ET}^W$ , and  $K_{ET}^L$ . For  $k = 1$ ,  $AR$  coincides with  $K$  and for  $k > 1$  I find that  $K$  has in most cases better size properties than  $AR$ . The qualitative features of the size results for  $GELR_{ET}$ ,  $K_{ET}^W$ , and  $K_{ET}^L$  are identical to their  $EL$  counterparts.

I first discuss the separate effects of  $n, k, \sigma$ , and  $\Pi_1$  on the size results.

---

<sup>4</sup>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 (3.19) using Design (I) with  $n = 100$  and  $\sigma = .99$ . I simulate a new matrix  $Z$  with each sample of the DGP (3.19). As a consequence, my results do not coincide with the results that Kleibergen (2002a) reports.

To investigate the sensitivity of the results in Kleibergen (2002a) to the choice of  $Z$ , I iterated Kleibergen’s (2002a) 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 (3.19). 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 (2002a) generates, he reports power of about 93%, see his Figure 1, p.1793.

The most important finding is that the size results of all statistics except  $2SLS$  show little or no dependence on  $\Pi_1$ . However, the size results of  $2SLS$  depend crucially on the strength or weakness of identification. While for  $\Pi_1 = 1$ ,  $2SLS$  has reliable size properties for many cases, with weak instruments sizes range over the entire interval from 0 to 100%.

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

The size results of  $AR$ , and  $GELR_\rho$  do not depend on the value of  $\sigma$ . The slight dependence of the size results in TABLE 1 on  $\sigma$  results from the use of different samples. For all the remaining statistics except for  $2SLS$ ,  $\sigma$  does not affect the size properties in a clear pattern and there is little dependence of  $\sigma$  on the results. However for  $2SLS$ , increasing  $\sigma$ , leads to severe over-rejection if combined with over-identification especially when the instruments are weak.

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

I now compare the performance of the statistics to each other. The  $2SLS$  statistics should not be used with weak instruments or in strongly endogenous over-identified problems when the sample size is small. In all other cases,  $2SLS$  has very competitive size properties. Using  $2SLS_{HET}$  instead of  $2SLS_{HOM}$  usually slightly increases the rejection rates. The statistics  $GELR_{EL}$  and  $K_{EL}^L$  severely over-reject in over-identified problems when the sample size is small. Overall, the statistics  $K_{CUE}^W$ ,  $K_{EL}^W$ ,  $GELR_{CUE}$ ,  $K$ , and  $LR_M$  lead to the best size results. Across the 54 combinations the sizes of  $K_{CUE}^W$  are in the interval [1.4,5.3]. For  $K_{EL}^W$ ,  $GELR_{CUE}$ ,  $K$ , and  $LR_M$  the corresponding intervals are [3.7,6.3], [1.4,5.3], [4.9,8.5], and [4.7,9.3]. While the former two statistics tend to underreject, especially in over-identified situations, the latter two usually slightly over-reject. In 26 of the 54 cases, the size of  $K_{EL}^W$  comes closest to the 5% significance level across all the statistics. The corresponding numbers for  $K_{CUE}^W$ ,  $GELR_{CUE}$ ,  $K$ , and  $LR_M$  are 5, 5, 19, and 13. Based on the size results of Design (I),  $K_{EL}^W$  has a slight advantage over the remaining statistics.

I now discuss the size results for Design ( $I_{HET}$ ). TABLE 2 contains the same information for Design ( $I_{HET}$ ) that TABLE 1 contains for Design (I). Most findings are similar to the ones discussed for Design (I) and I only discuss the new features that arise when moving from Design (I) to ( $I_{HET}$ ).

The statistics  $2SLS_{HOM}$ ,  $K$ ,  $LR_M$  perform uniformly worse than in Design (I). The tests based on these statistics severely over-reject, especially in the just-identified case. The performance does not improve when  $n$  increases. Rejection rates of the three tests across the 54 combinations are in the intervals [.9,100], [7.5,26.9], and [7.4-26.8], respectively.

In contrast, the size properties of  $2SLS_{HET}$  and of the statistics based on GEL methods, are not negatively influenced by conditional heteroskedasticity. This is to

be expected based on the theory section of the paper that does not assume conditional homoskedasticity. Of course,  $2SLS_{HET}$  still suffers from weak identification and  $GELR_{EL}$  and  $K_{EL}^L$  perform poorly in over-identified situation for small  $n$ . Rejection rates of the three test statistics  $K_{CUE}^W$ ,  $K_{EL}^W$ , and  $GELR_{CUE}$  across the 54 combinations are in the intervals  $[1.1,5.0]$ ,  $[1.4,5.0]$ , and  $[3.5-6.5]$ , respectively.

In summary, the only statistics with accurate size properties across all combinations of Designs (I) and ( $I_{HET}$ ) are  $K_{EL}^W$ ,  $K_{CUE}^W$ , and  $GELR_{CUE}$ . Based on the above results we find that  $K_{EL}^W$  enjoys a slight advantage over the other two. From the 108 cases in TABLES 1 and 2 the size of  $K_{EL}^W$  is closest to 5% in 74 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. One striking difference however occurs for  $2SLS$  under weak identification with  $\chi^2(1)$  and bimodal errors. Rejection rates across the 54 combinations of Design (III) and (IV) for  $2SLS_{HOM}$  are in the intervals  $[.1,7.1]$  and  $[0.0,5.4]$ . While with normal errors and weak identification  $2SLS$  severely over-rejects, with these errors distributions it severely under-rejects.

Due to the similarity of the overall picture, I do not include additional tables for Designs (II)-(IV) and ( $II_{HET}$ )-(IV $_{HET}$ ).

The overall conclusion of this section is that  $K_{EL}^W$ ,  $K_{CUE}^W$ , and  $GELR_{CUE}$  have reliable size properties across all designs independent of the strength or weakness of identification and independent of heteroskedasticity.  $2SLS$  performs very poorly with weak instruments. Using  $2SLS_{HET}$  instead of  $2SLS_{HOM}$  significantly improves the size properties when there is conditional heteroskedasticity in the data and only slightly worsens the size properties when there is not. The statistics  $K$  and  $LR_M$  perform well in the homoskedastic cases but poorly in cases with heteroskedasticity.

### Power comparison

I calculate power curves of the above statistics for DGP in (3.19) that correspond to all 16 possible combinations of  $n = 100, 250$ ,  $k = 5, 10$ ,  $\sigma = .5, .99$  and  $\Pi_1 = .1, 1$  for each of the three error distributions in Designs (I)-(III). Except for  $LR_M$ , I report size-corrected power curves at the 5% significance level, 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 straightforward, and I therefore use power curves for  $LR_M$  that have not been size-corrected.

Across almost all scenarios the statistics  $K_{CUE}^W$ ,  $K_{EL}^W$ , and  $K_{ET}^W$  have very similar performance and therefore I only report results for  $K_{EL}^W$ . I do not report power results for the statistics  $K_{EL}^L$  and  $K_{ET}^L$  because, as we have seen above, their size properties can be quite poor for the sample sizes we consider. In the case  $k = 1$ ,  $AR$  and  $K$  are numerically identical. In the over-identified cases  $K$  generally performs better than  $AR$ . I therefore do not report results for  $AR$  but refer to Kleibergen (2002a) for the comparison of  $K$  and  $AR$ . Similarly,  $GELR_{CUE}$  is numerically identical to  $K_{\rho}^W$  for  $k = 1$  but leads to a less powerful test for  $k > 1$ . The statistics  $GELR_{\rho}$  for  $\rho = EL$  and  $ET$  have rather unreliable size properties for the sample sizes we consider. Therefore I do not report detailed results for  $GELR_{\rho}$ .

Detailed results are discussed for the statistics  $K_{EL}^W$ ,  $K$ ,  $LR_M$ , and  $2SLS_{HET}$ .

I use 1,000 samples from the DGP in (3.19) for various values of the true value  $\theta_0$  and test the hypothesis that  $\theta_0 = 0$ . With weak identification, I take  $\theta_0$  values in the interval  $[-4,4]$  and with strong identification in  $[-.4,.4]$ . I use 1,000 realizations from a  $\chi^2(1)$  and  $\chi^2(k-1)$  random variable to simulate the cut-off values for the  $LR_M$  statistic. For the results that I actually report in the figures below, I use 10,000 samples from (3.19).

I first discuss general features of the power curves under strong and weak identification.

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 the true value  $\theta_0$ . In Designs (I) and (II),  $2SLS_{HET}$  is usually biased, taking on its lowest value at a negative  $\theta$  value.

With weak identification, the power curves of  $K_{EL}^W$ ,  $K$ , and  $LR_M$  are generally very flat across all  $\theta_0$  values, hardly exceeding the significance level of the test. However, in Design (I) with  $\sigma = .99$ , while being flat at about 5% for positive  $\theta_0$  values, the power curves reach a sharp peak of almost 100% at about  $\theta_0 = -1$ . For negative  $\theta_0$  values with  $|\theta_0| > 1$  power quickly falls back to lower values, reaching about 20% at  $\theta_0 = -4$ .

In contrast to the power curves of  $K_{EL}^W$ ,  $K$ , and  $LR_M$ , the power curve of  $2SLS_{HET}$  retains its U-shaped form for  $\Pi_1 = .1$ . In many cases, the power curve reaches values close to 100% when  $|\theta_0|$  is close to 4.

Next I discuss the impact of  $n$ ,  $k$ , and  $\sigma$  on the power results.

As to be expected the power curves take on bigger values when  $n$  is increased from 100 to 250. This holds uniformly across statistics and designs.

...

### Acknowledgements

Financial support for this research was provided by a Carl Arvid Anderson Prize Fellowship. 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. I gratefully acknowledge the help of my advisors Donald Andrews, Peter Phillips, and Joseph Altonji. I would also like to thank John Chao, Frank Kleibergen, and Motohiro Yogo for helpful discussion and correspondence, and Vadim Marner for help with the simulation section.

## 4 Appendix of Proofs

I introduce one more version of Assumptions 1 and 3 that includes all versions so far introduced as subcases.

**Assumption 1** Let  $\Pi_n = (\Pi_1, \Pi_2, \Pi_3, \Pi_4) = (n^{-1/2}C_1, C_2, n^{-1/2}C_3, C_4)$ , where  $C_i$  for  $i = 1, \dots, 4$  are fixed matrices of dimensions  $k \times p_i$ , where  $p_1 + \dots + p_4 = p$ . Let  $y = Y(\theta_0 + n^{-1/2}(0', 0', 0', \theta'_4)') + u$  for a fixed  $p_4$ -vector  $\theta_4$ .

Decompose  $\theta_0$  conformably as  $\theta_0 = (\theta'_{01}, \theta'_{02}, \theta'_{03}, \theta'_{04})'$ .

**Assumption 3**  $\inf_{\{d_1 \in R^{p_1}, d_3 \in R^{p_3} | (d'_1, \theta'_{02}, d'_3, \theta'_{04})' \in \Theta\}} \lambda_{\min}\{\Omega(d'_1, \theta'_{02}, d'_3, \theta'_{04})'\} > 0$ .

For fixed  $d_{34} \in \{d_{34} \in R^{p_3+p_4} | \exists d_{12} \in R^{p_1+p_2}, (d'_{12}, d'_{34})' \in \Theta\}$  define

$$\widehat{\theta}_{12}(d_{34}) := \arg \min_{\{d_{12} \in R^{p_1+p_2} | \theta := (d'_{12}, d'_{34})' \in \Theta\}} \sup_{\lambda \in \widehat{\Lambda}_n(\theta)} \widehat{P}(\theta, \lambda) \quad (4.1)$$

Assumptions 1(S-LA), (W), (WS-LA), and (S $_{\alpha}$ -LA) correspond to  $p_3 = p$ ,  $p_4 = p$ ,  $p_3 + p_4 = p$ , and  $p_2 = 0$  in Assumption 1, respectively. Assumption 1(WS) in the context of Theorem 5 corresponds to  $p_1 + p_2 = p$ .

We start off with some lemmas that are needed throughout the Appendix. The first one establishes uniform convergence of  $\widehat{\Omega}(\theta)$  to  $\Omega(\theta)$ .

**Lemma 9** *Suppose Assumptions 1 and 2. Then  $\sup_{\theta, \bar{\theta} \in \Theta} \|\widehat{\Omega}(\theta, \bar{\theta}) - \Omega(\theta, \bar{\theta})\| \rightarrow_p 0$ , in particular,  $\sup_{\theta \in \Theta} \|\widehat{\Omega}(\theta) - \Omega(\theta)\| \rightarrow_p 0$ .*

**Proof.** Using (2.2), Assumptions 1 and 2, and the WLLN we have that uniformly over  $\theta, \bar{\theta} \in \Theta$

$$\widehat{\Omega}(\theta, \bar{\theta}) \rightarrow_p \lim_{n \rightarrow \infty} E[(Z'_i \Pi_n + V'_i)(\theta_0 - \theta) + u_i][(Z'_i \Pi_n + V'_i)(\theta_0 - \bar{\theta}) + u_i] Z_i Z'_i.$$

The last expression equals  $\Omega(\theta, \bar{\theta})$ . Note that the assumption of compactness of  $\Theta$  is crucial for the uniformity result.  $\square$

The next Lemmas are modified versions of Lemmas A1-A3 in NS. The modifications are necessary because I work with weakly and strongly identified parameters and because my moment assumptions are slightly weaker, see Remark 1. The lemmas are needed for every theorem in the paper.

**Lemma 10** *Suppose Assumptions 1 and 2. Let  $f_{ni} := \sup_{\theta \in \Theta} \|g_{ni}(\theta)\|$ ,  $c_n := n^{-1/2} \max_{1 \leq i \leq n} f_{ni}$ , and  $\Lambda_n := \{\lambda \mid \|\lambda\| \leq n^{-1/2} c_n^{-1/2}\}$ . Then*

(i)  $\sup_{\theta \in \Theta, \lambda \in \Lambda_n, 1 \leq i \leq n} |\lambda' g_i(\theta)|$  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  $\max_{1 \leq i \leq n} |W_i| = o(n^{1/2})$ , see Owen (1990, Lemma 3) for a proof. By definition of  $g_{ni}(\theta)$  in (2.2), Assumption 1, and the triangle inequality, we have for  $\theta_n := \theta_0 - \theta + n^{-1/2}(0', 0', 0', \theta'_4)'$

$$\max_{1 \leq i \leq n} f_{ni} \leq \max_{1 \leq i \leq n} \sup_{\theta \in \Theta} (\|u_i Z_i\| + \|Z_i Z'_i \Pi_n \theta_n\| + \|Z_i V_i \theta_n\|). \quad (4.2)$$

By Assumption 2, I can apply the above result to each of the three summands in (4.2). Therefore,  $\max_{1 \leq i \leq n} f_{ni} = o(n^{1/2})$  and thus  $c_n = o(1)$ . Part (i) then follows from

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

which also immediately implies (ii).  $\square$

Let  $(\theta_{02n}) \subset R^{p_2}$  and  $\theta_{nd\bar{d}} := (d', \theta'_{02n}, \bar{d}', \theta'_{04})' \in R^p$  and  $\Theta_n := \{\theta_{nd\bar{d}} | d \in R^{p_1}, \bar{d} \in R^{p_3}, \theta_{nd\bar{d}} \in \Theta\}$ .

**Lemma 11** *Suppose Assumptions 1-4. Assume  $\hat{g}(\theta) = O_p(n^{-1/2})$  uniformly over  $\theta \in \Theta_n$  and  $\theta_{02n} \rightarrow_p \theta_{02}$ . Then  $\lambda(\theta) := \arg \max_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta, \lambda)$  exists uwp1,  $\lambda(\theta) = O_p(n^{-1/2})$  uniformly over  $\theta \in \Theta_n$ , and  $\sup_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta, \lambda) = O_p(n^{-1})$  uwp1, where “uwp1” stands for “uniformly over  $\theta \in \Theta_n$  uwp1”.*

**Proof.** For  $\theta \in \Theta_n$ , define  $\lambda_\theta := \arg \max_{\lambda \in \Lambda_n} \hat{P}(\theta, \lambda)$ , where  $\Lambda_n$  is defined in Lemma 10. This definition is justified uwp1 because a continuous function takes on its maximum on a compact set and by Lemma 10,  $\hat{P}(\theta, \lambda)$  (as a function in  $\lambda$  for fixed  $\theta$ ) is  $C^2$  uwp1 on some open neighborhood of  $\Lambda_n$ . I now show that actually  $\hat{P}(\theta, \lambda_\theta) = \sup_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta, \lambda)$  which then proves the first part of the lemma. By a second order Taylor expansion around  $\lambda = 0$ , there is  $\lambda_\theta^*$  on the line segment  $\overline{0\lambda_\theta}$ , such that for some positive constants  $C_1$  and  $C_2$

$$\begin{aligned} 0 = \hat{P}(\theta, 0) &\leq \hat{P}(\theta, \lambda_\theta) = -2\lambda_\theta' \hat{g}(\theta) + \lambda_\theta' \left[ \sum_{i=1}^n \rho_2(\lambda_\theta^{*'} g_i(\theta)) g_i(\theta) g_i(\theta)' / n \right] \lambda_\theta \\ &\leq -2\lambda_\theta' \hat{g}(\theta) - C_1 \lambda_\theta' \hat{\Omega}(\theta) \lambda_\theta \leq 2\|\lambda_\theta\| \|\hat{g}(\theta)\| - C_2 \|\lambda_\theta\|^2 \end{aligned} \quad (4.3)$$

uwp1, where the second to last inequality follows from the fact that by Lemma 10, continuity of  $\rho_2(\cdot)$  at zero, and  $\rho_2 = -1$ , we have  $\max_{1 \leq i \leq n} \rho_2(\lambda_\theta^{*'} g_i(\theta)) < -1/2$  uwp1. The last inequality follows from  $\theta_{02n} \rightarrow_p \theta_{02}$ , Lemma 9 and Assumption 3, which imply that the smallest eigenvalue of  $\hat{\Omega}(\theta)$  is positive and bounded away from zero uwp1. Now, (4.3) implies that  $(C_2/2)\|\lambda_\theta\| \leq \|\hat{g}(\theta)\|$  uwp1, the latter being  $O_p(n^{-1/2})$  uniformly over  $\theta \in \Theta_n$  by assumption. It follows that  $\lambda_\theta \in \text{int}(\Lambda_n)$  uwp1. To prove this, let  $\varepsilon > 0$ . Because  $\lambda_\theta = O_p(n^{-1/2})$  uniformly over  $\theta \in \Theta_n$ , there exist  $M_\varepsilon < \infty$  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  $\theta \in \Theta_n$ . Because  $c_n = o(1)$ , we can choose  $n(\varepsilon) > n_\varepsilon$  so big that  $c_n^{-1/2} > M_\varepsilon$  for all  $n \geq n(\varepsilon)$ . Then  $\Pr(\lambda_\theta \in \text{int}(\Lambda_n)) = \Pr(\|n^{1/2}\lambda_\theta\| < c_n^{-1/2}) \geq \Pr(\|n^{1/2}\lambda_\theta\| \leq M_\varepsilon) > 1 - \varepsilon$  for  $n \geq n(\varepsilon)$  uniformly over  $\theta \in \Theta_n$ .

Hence, the FOC for an interior maximum  $\partial \hat{P}(\theta, \lambda) / \partial \lambda = 0$  holds at  $\lambda = \lambda_\theta$  uwp1. By Lemma 10 we know that  $\lambda_\theta \in \hat{\Lambda}_n(\theta)$  uwp1 and thus by concavity of  $\hat{P}(\theta, \lambda)$  (as a function in  $\lambda$  for fixed  $\theta$ ) and convexity of  $\hat{\Lambda}_n(\theta)$  it follows that  $\lambda_\theta = \arg \max_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}(\theta, \lambda)$  which implies the first part of the lemma. Because  $\lambda_\theta = O_p(n^{-1/2})$  uniformly over  $\theta \in \Theta_n$ , the second part and by (4.3) the third part follows too.  $\square$

Let  $\widehat{\theta}_{d_3} := \widehat{\theta}_{12}((d'_3, \theta'_{04})')$  for  $d_3 \in \Theta_3 := \{d_3 \in R^{p_3} | \exists d_1 \in R^{p_1} (d'_1, \theta'_{02}, d'_3, \theta'_{04})' \in \Theta\}$ , where  $\widehat{\theta}_{12}((d'_3, \theta'_{04})')$  is defined in (4.1).

**Lemma 12** *Suppose Assumptions 1-4. Then  $\widehat{g}(\widehat{\theta}_{d_3}) = O_p(n^{-1/2})$  uwp1, where “uwp1” stands for “uniformly over  $d_3 \in \Theta_3$  wpa1”.*

**Proof.** Define  $\underline{\lambda} := -n^{-1/2} \widehat{g}(\widehat{\theta}_{d_3}) / \|\widehat{g}(\widehat{\theta}_{d_3})\|$ . Note that  $\underline{\lambda} \in \Lambda_n$  uwp1, see Lemma 10. 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  $C_2$

$$\begin{aligned} \widehat{P}(\widehat{\theta}_{d_3}, \underline{\lambda}) &= -2\underline{\lambda}' \widehat{g}(\widehat{\theta}_{d_3}) + \underline{\lambda}' \left[ \sum_{i=1}^n \rho_2(\widetilde{\lambda}' g_i(\widehat{\theta}_{d_3})) g_i(\widehat{\theta}_{d_3}) g_i(\widehat{\theta}_{d_3})' / n \right] \underline{\lambda} \\ &\geq 2n^{-1/2} \|\widehat{g}(\widehat{\theta}_{d_3})\| - C_1 \underline{\lambda}' \left[ \sum_{i=1}^n g_i(\widehat{\theta}_{d_3}) g_i(\widehat{\theta}_{d_3})' / n \right] \underline{\lambda} \\ &\geq 2n^{-1/2} \|\widehat{g}(\widehat{\theta}_{d_3})\| - C_2 n^{-1} \end{aligned} \quad (4.4)$$

uwp1, where the first inequality follows from Lemma 10 which implies that  $\min_{i=1, \dots, n} \rho_2(\widetilde{\lambda}' g_i(\widehat{\theta}_{d_3})) \geq -1.5$  uwp1. The last inequality follows by the uniform convergence result in Lemma 9 and boundedness of  $\Theta$  which imply that the largest eigenvalue of  $n^{-1} \sum_{i=1}^n g_i(\widehat{\theta}_{d_3}) g_i(\widehat{\theta}_{d_3})'$  is bounded above uwp1. Choose a  $d_1 \in R^{p_1}$  s.t.  $\theta_{d_1 d_3} := (d'_1, \theta'_{02}, d'_3, \theta'_{04})' \in \Theta$ . The definition of  $\widehat{\theta}_{d_3}$  implies

$$\widehat{P}(\widehat{\theta}_{d_3}, \underline{\lambda}) \leq \sup_{\lambda \in \widehat{\Lambda}_n(\widehat{\theta}_{d_3})} \widehat{P}(\widehat{\theta}_{d_3}, \lambda) \leq \sup_{\lambda \in \widehat{\Lambda}_n(\theta_{d_1 d_3})} \widehat{P}(\theta_{d_1 d_3}, \lambda) = O_p(n^{-1}) \quad (4.5)$$

uwp1. The last equality follows from Lemma 11 (for the case where  $\theta_{02n} \equiv \theta_{02}$ ) and noting that  $\sup_{\{d_1 \in R^{p_1}, d_3 \in R^{p_3} | \theta_{d_1 d_3} \in \Theta\}} \|\widehat{g}(\theta_{d_1 d_3})\| = O_p(n^{-1/2})$ , which holds by the CLT. Finally, combining equations (4.4) and (4.5) implies  $n^{-1/2} \|\widehat{g}(\widehat{\theta}_{d_3})\| = O_p(n^{-1})$  uwp1.  $\square$

**Proof of Lemma 1.** Under Assumption 1(W),  $g_i(\theta)$  is given by equation (2.3) with  $\xi = 1/2$ . Because  $\sup_{\theta \in \Theta} \|n^{-1} \sum_{i=1}^n Z_i' C(\theta_0 - \theta) Z_i - Q_{ZZ} C(\theta_0 - \theta)\| \rightarrow_p 0$ , 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:

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

Furthermore,  $\Theta$  is compact by assumption. The proposition on p.2251 in Andrews (1994) can thus be applied which yields the desired result.  $\square$

**Proof of Theorem 2.** Write “uwp1” for “uniformly over  $\theta \in \Theta$  wpa1”. By Lemmas 11 and 12 for the case  $p_4 = p$ , the FOC with respect to  $\lambda$ ,  $n^{-1} \sum_{i=1}^n \rho_1(\lambda' g_i(\theta)) g_i(\theta) = 0$  has to hold at  $\widetilde{\lambda} = \lambda(\theta)$  uwp1. Expanding the FOC in  $\lambda$  around 0, we get for some mean value  $\widetilde{\lambda}$  on the line segment  $\overline{0\lambda(\theta)}$

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

where the matrix  $\widehat{\Omega}_{\widehat{\lambda}\theta}$  has been implicitly defined. Because  $\lambda(\theta) = O_p(n^{-1/2})$  uniformly over  $\theta \in \Theta$ , Lemma 10 implies that  $\max_{i,\theta \in \Theta} |\rho_2(\widehat{\lambda}' g_i(\theta)) + 1| \rightarrow_p 0$ . By Lemma 9, it then follows that  $\widehat{\Omega}_{\widehat{\lambda}\theta}$  converges in probability to  $\Omega(\theta)$  uniformly over  $\theta \in \Theta$  and by Assumption 3(W)  $\widehat{\Omega}_{\widehat{\lambda}\theta}$  is thus invertible uwpa1. Therefore  $\lambda(\theta) = -\widehat{\Omega}_{\widehat{\lambda}\theta}^{-1} \widehat{g}(\theta)$  uwpa1. Inserting this into a second order Taylor expansion for  $\widehat{P}(\theta, \lambda)$  (with mean value  $\lambda^*$ , like in (4.3) above) we find that

$$\widehat{P}(\theta, \lambda(\theta)) = 2\widehat{g}(\theta)' \widehat{\Omega}_{\widehat{\lambda}\theta}^{-1} \widehat{g}(\theta) - \widehat{g}(\theta)' \widehat{\Omega}_{\widehat{\lambda}\theta}^{-1} \widehat{\Omega}_{\lambda^*\theta} \widehat{\Omega}_{\widehat{\lambda}\theta}^{-1} \widehat{g}(\theta). \quad (4.6)$$

Lemma 1 now implies that  $n\widehat{P}(\theta, \lambda(\theta))$  converges weakly to  $P(\theta)$ . The second part of the theorem then follows from Lemma 3.2.1 in van der Vaart and Wellner (1996, p.286).  $\square$

**Proof of Corollaries 3 and 4.** These corollaries are special cases of Corollary 6.  $\square$

**Proof of Theorem 5.** (i) I first show consistency of  $\widehat{\beta}$ . Note that by the WLLN and compactness of  $\Theta$ , we have  $\sup_{\theta \in \Theta} \|\widehat{g}(\theta) - Q_{ZZ} C_B(\beta_0 - \beta)\| \rightarrow_p 0$ . Therefore,  $\|\widehat{g}(\widehat{\theta})\| = o_p(1)$  is a sufficient condition for consistency of  $\widehat{\beta}$  because  $Q_{ZZ} C_B$  has full rank. By Lemma 12 applied to the case  $p_1 + p_2 = p$ , we even have  $\|\widehat{g}(\widehat{\theta})\| = O_p(n^{-1/2})$ .

Next I establish  $n^{1/2}$ -consistency for  $\widehat{\beta}$ , following a standard procedure, see the proof of Theorem 3.2 in NS. Given consistency of  $\widehat{\beta}$ , Lemma 11 implies that the FOC

$$n^{-1} \sum_{i=1}^n \rho_1(\lambda' g_i(\theta)) g_i(\theta) = 0 \quad (4.7)$$

has to hold at  $(\widehat{\theta}, \widehat{\lambda})$  wpa1, where  $\widehat{\lambda} := \lambda(\widehat{\theta})$  and where  $\lambda(\theta)$ , for given  $\theta \in \Theta$ , has been defined in (2.10). Lemmas 9-12 imply that  $n^{-1} \sum_{i=1}^n \rho_2(\widehat{\lambda}' g_i(\widehat{\theta})) g_i(\widehat{\theta}) g_i(\widehat{\theta})'$  converges in probability to  $-\Omega(\widehat{\theta})$ , which by Assumption 3(WS) and consistency of  $\widehat{\beta}$  is negative definite wpa1 and thus nonsingular wpa1. Therefore, the implicit function theorem implies that there is a neighborhood of  $\widehat{\theta}$  where the solution to the FOC, say  $\widehat{\lambda}(\theta)$ , is continuously differentiable wpa1. The envelope theorem then implies

$$n^{-1} \sum_{i=1}^n \rho_1(\widehat{\lambda}' g_i(\widehat{\theta})) (\partial g_i / \partial \theta)'(\widehat{\theta}) \widehat{\lambda} = 0 \quad (4.8)$$

wpa1. A mean-value expansion of (4.7) about  $(\theta, \lambda) = (\theta_0, 0)$  yields (where  $g_i(\theta)$  inside  $\rho_1$  is kept constant at  $g_i(\widehat{\theta})$ )

$$-\widehat{g}(\theta_0) + n^{-1} \sum_{i=1}^n [\rho_1(\widehat{\lambda}' g_i(\widehat{\theta})) (\partial g_i / \partial \theta)(\bar{\theta}) (\widehat{\theta} - \theta_0) + \rho_2(\widehat{\lambda}' g_i(\widehat{\theta})) g_i(\bar{\theta}) g_i(\widehat{\theta})' \widehat{\lambda}] = 0, \quad (4.9)$$

where  $(\bar{\theta}, \bar{\lambda})$  are mean-values between  $(\theta_0, 0)$  and  $(\widehat{\theta}, \widehat{\lambda})$  that may be different for each row. Combining the  $p$  rows of (4.8) with the  $k$  rows of (4.9) I get

$$\begin{pmatrix} 0 \\ -\widehat{g}(\theta_0) \end{pmatrix} + M \begin{pmatrix} \widehat{\theta} - \theta_0 \\ \widehat{\lambda} \end{pmatrix} = 0, \quad (4.10)$$

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

$$\frac{1}{n} \sum_{i=1}^n \rho_1(\bar{\lambda}' g_i(\hat{\theta})) (\partial g_i(\theta)/\partial \theta)(\bar{\theta}) \rightarrow_p (0, Q_{ZZ} C_B). \quad (4.11)$$

By Lemma 9 and consistency of  $\hat{\beta}$

$$n^{-1} \sum_{i=1}^n \rho_2(\bar{\lambda}' g_i(\hat{\theta})) g_i(\bar{\theta}) g_i(\hat{\theta})' \rightarrow_p \Omega_{\bar{\alpha}\hat{\alpha}} := -\Omega((\bar{\alpha}', \beta_0)', (\hat{\alpha}', \beta_0)'), \quad (4.12)$$

the latter matrix being random and by assumption nonsingular. Denote by  $M_{\bar{\alpha}\hat{\alpha}}$  the  $(p_B+k) \times (p_B+k)$  bottom-right submatrix of  $M$  that corresponds to the parameters  $\bar{\beta}$  and  $\bar{\lambda}$ . If by  $\overline{M_{\bar{\alpha}\hat{\alpha}}}$  we denote its probability limit, equations (4.11) and (4.12) imply that

$$\overline{M_{\bar{\alpha}\hat{\alpha}}} = \begin{pmatrix} 0 & C_B' Q_{ZZ} \\ Q_{ZZ} C_B & \Omega_{\bar{\alpha}\hat{\alpha}} \end{pmatrix}.$$

It follows that

$$\overline{M_{\bar{\alpha}\hat{\alpha}}}^{-1} = \begin{pmatrix} -\Sigma_{\bar{\alpha}\hat{\alpha}} & H_{\bar{\alpha}\hat{\alpha}} \\ H_{\bar{\alpha}\hat{\alpha}}' & P_{\bar{\alpha}\hat{\alpha}} \end{pmatrix},$$

where

$$\begin{aligned} \Sigma_{\bar{\alpha}\hat{\alpha}} &:= (C_B' Q_{ZZ} \Omega_{\bar{\alpha}\hat{\alpha}}^{-1} Q_{ZZ} C_B)^{-1}, \quad H_{\bar{\alpha}\hat{\alpha}} := \Sigma_{\bar{\alpha}\hat{\alpha}} C_B' Q_{ZZ} \Omega_{\bar{\alpha}\hat{\alpha}}^{-1}, \quad \text{and} \\ P_{\bar{\alpha}\hat{\alpha}} &:= \Omega_{\bar{\alpha}\hat{\alpha}}^{-1} - \Omega_{\bar{\alpha}\hat{\alpha}}^{-1} Q_{ZZ} C_B \Sigma_{\bar{\alpha}\hat{\alpha}} C_B' Q_{ZZ} \Omega_{\bar{\alpha}\hat{\alpha}}^{-1}. \end{aligned} \quad (4.13)$$

It follows that  $M_{\bar{\alpha}\hat{\alpha}}$  is nonsingular wpa1. Equation (4.10) then implies that

$$n^{1/2}(\hat{\beta}' - \beta_0', \hat{\lambda}')' = M_{\bar{\alpha}\hat{\alpha}}^{-1}(0', n^{1/2}[\hat{g}(\theta_0) - \frac{1}{n} \sum_{i=1}^n \rho_1(\bar{\lambda}' g_i(\hat{\theta})) (\partial g_i(\theta)/\partial \alpha)(\hat{\alpha} - \alpha_0)]')',$$

By the CLT and the WLLN  $n^{1/2}\hat{g}(\theta_0)$  and  $n^{-1/2} \sum_{i=1}^n Z_i(Z_i' n^{-1/2} C_A + V_{iA}')$  are  $O_p(1)$ . The same is true for  $(\hat{\alpha} - \alpha_0)$  by compactness of  $\Theta$  and for  $M_{\bar{\alpha}\hat{\alpha}}^{-1}$  by  $M_{\bar{\alpha}\hat{\alpha}}^{-1} \rightarrow_p \overline{M_{\bar{\alpha}\hat{\alpha}}}^{-1}$ . The previous equation thus implies that

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

which establishes part (i) of the theorem.

(ii) By  $n^{1/2}$ -consistency of  $\hat{\beta}$ , it follows that  $\|\hat{g}(\hat{\theta}) - \hat{g}(\theta_{\alpha b})\| = O_p(n^{-1/2})$  uniformly over  $(\alpha, b) \in A \times B_{\beta_0}$ , where  $\theta_{\alpha b} := (\alpha', \beta_0' + n^{-1/2} b')'$ . Therefore, Lemma 12 implies that  $\hat{g}(\theta_{\alpha b}) = O_p(n^{-1/2})$  uniformly over  $A \times B_{\beta_0}$ . By Lemma 11  $\lambda(\theta_{\alpha b}) := \arg \max_{\lambda \in \hat{\Lambda}_n(\theta_{\alpha b})} \hat{P}(\theta_{\alpha b}, \lambda)$  exists and  $\lambda(\theta_{\alpha b}) = O_p(n^{-1/2})$  uniformly over  $(\alpha, b) \in A \times B_{\beta_0}$ . The former statement implies that the FOC  $n^{-1} \sum_{i=1}^n \rho_1(\lambda' g_i(\theta)) g_i(\theta) = 0$  has to hold at  $\lambda = \lambda(\theta_{\alpha b})$  and  $\theta = \theta_{\alpha b}$ . Expanding the FOC and using the same steps and notation as in the proof of Theorem 2, we obtain  $\lambda(\theta_{\alpha b}) = -\hat{\Omega}_{\lambda\theta_{\alpha b}}^{-1} \hat{g}(\theta_{\alpha b})$  and upon inserting this into a second order Taylor expansion of  $\hat{P}(\theta, \lambda)$  we get

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

The matrices  $\widehat{\Omega}_{\lambda\theta_{\alpha b}}$  and  $\widehat{\Omega}_{\lambda^*\theta_{\alpha b}}$  converge uniformly to  $\Omega(\alpha, \beta_0)$ . Also note that like in Lemma 1 we have

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

and therefore that

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

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

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

**Proof of Corollary 6.** (i) Let  $\theta_{\alpha\beta_0} := (\alpha', \beta_0)'$ . By Lemmas 11 and 12 for the case  $p_3 + p_4 = p$ , we know that  $\lambda(\theta_{\alpha\beta_0}) := \arg \sup_{\lambda \in \widehat{\Lambda}_n(\theta_{\alpha\beta_0})} \widehat{P}(\theta_{\alpha\beta_0}, \lambda)$  exists uniformly over  $\{\alpha \in R^{p_A} | (\alpha', \beta_0)' \in \Theta\}$  wpa1. Using Assumption 3(W) and the same steps and notation as in the proof of Theorem 2, leads to

$$\widehat{P}(\theta_{\alpha\beta_0}, \lambda_{\alpha\beta_0}) = 2\widehat{g}(\theta_{\alpha\beta_0})'\widehat{\Omega}_{\lambda\theta_{\alpha\beta_0}}^{-1}\widehat{g}(\theta_{\alpha\beta_0}) - \widehat{g}(\theta_{\alpha\beta_0})'\widehat{\Omega}_{\lambda\theta_{\alpha\beta_0}}^{-1}\widehat{\Omega}_{\lambda^*\theta_{\alpha\beta_0}}\widehat{\Omega}_{\lambda\theta_{\alpha\beta_0}}^{-1}\widehat{g}(\theta_{\alpha\beta_0}),$$

where both  $\widehat{\Omega}_{\lambda\theta_{\alpha\beta_0}}$  and  $\widehat{\Omega}_{\lambda^*\theta_{\alpha\beta_0}}$  converge in probability to  $\Omega(\theta_{\alpha\beta_0})$ . Finally,

$$n^{1/2}\widehat{g}(\theta_{\alpha\beta_0}) \rightarrow_d N(q, \Omega(\theta_{\alpha\beta_0})) \quad (4.14)$$

for

$$q := Q_{ZZ}C((\alpha_0 - \alpha)', b)'$$

from which the result follows.

(ii) By an argument as in part (i),  $n^{1/2}\lambda(\theta_{\alpha\beta_0}) = -\Omega(\theta_{\alpha\beta_0})^{-1}n^{1/2}\widehat{g}(\theta_{\alpha\beta_0}) + o_p(1)$  and therefore the statement of the theorem involving  $K_\rho^L(\theta_{\alpha\beta_0})$  follows immediately from the one for  $K_\rho^W(\theta_{\alpha\beta_0})$ . Therefore, I only deal with the statistic  $K_\rho^W(\theta_{\alpha\beta_0})$  which can be rewritten as

$$n\widehat{g}(\theta_{\alpha\beta_0})'\Omega(\theta_{\alpha\beta_0})^{-1}D^*(D^{*\prime}\Omega(\theta_{\alpha\beta_0})^{-1}D^*)^{-1}D^{*\prime}\Omega(\theta_{\alpha\beta_0})^{-1}\widehat{g}(\theta_{\alpha\beta_0}),$$

where  $\Delta := \text{diag}(n^{1/2}, \dots, n^{1/2}, 1, \dots, 1)$  is a diagonal  $p \times p$ -matrix whose first  $p_A$  diagonal elements equal  $n^{1/2}$  and the remaining  $p_B$  elements equal 1, and where  $D^* := D_\rho(\theta_{\alpha\beta_0})\Delta$  denotes the renormalized  $D_\rho(\theta_{\alpha\beta_0})$  matrix. If  $D_\rho(\theta_{\alpha\beta_0})$  is not renormalized, then the first  $p_A$  columns of  $D_\rho(\theta_{\alpha\beta_0})$  converge in probability to zero implying that the matrix  $D_\rho(\theta_{\alpha\beta_0})'\Omega(\theta_{\alpha\beta_0})^{-1}D_\rho(\theta_{\alpha\beta_0})$  is not invertible asymptotically.

First, I show that the matrix  $D^*$  is asymptotically independent of  $n^{1/2}\widehat{g}(\theta_{\alpha\beta_0})$ . Write  $g_i$  for  $g_i(\theta_{\alpha\beta_0})$ ,  $\lambda$  for  $\lambda(\theta_{\alpha\beta_0})$ , and  $\Omega$  for  $\Omega(\theta_{\alpha\beta_0})$ . Then, by Assumption 1(W) and the definition of  $\Delta$

$$D^* = -n^{-1} \sum_{i=1}^n \rho_1(\lambda' g_i) Z_i [Z_i'(C_A, C_B) + (n^{1/2} V_{iA}', V_{iB}')].$$

The term involving  $n^{1/2}V'_{iA}$  requires some more attention. By a first-order Taylor expansion of  $\rho_1(\lambda'g_i)$  about  $\lambda = 0$ , we have for some mean vector  $\bar{\lambda}$ ,  $\rho_1(\lambda'g_i) = -1 + \rho_2(\bar{\lambda}'g_i)g'_i\lambda$  and thus

$$-n^{-1/2} \sum_{i=1}^n \rho_1(\lambda'g_i)Z_iV'_{iA} = n^{-1/2} \sum_{i=1}^n Z_iV'_{iA} - [n^{-1} \sum_{i=1}^n \rho_2(\bar{\lambda}'g_i)Z_iV'_{iA}g'_i]n^{1/2}\lambda.$$

By Assumptions 2 and 4 we have

$$-n^{-1} \sum_{i=1}^n \rho_2(\bar{\lambda}'g_i)vec(Z_iV'_{iA})g'_i \rightarrow_p \Omega_\alpha := \lim_{n \rightarrow \infty} Evec(Z_iV'_{iA})g'_i \in R^{kp_A \times k}.$$

Combining (4.14) and the previous equations, we have

$$vec(D^*, n^{1/2}\hat{g}(\theta_{\alpha\beta_0})) = m + Mv + o_p(1),$$

where  $m := vec(Q_{ZZ}C, 0) \in R^{kp+k}$  and

$$M := \begin{pmatrix} I_{kp_A} & -\Omega_\alpha\Omega^{-1} \\ 0 & 0 \\ 0 & I_k \end{pmatrix}, \quad v := n^{-1/2} \sum_{i=1}^n \begin{pmatrix} vec(Z_iV'_{iA}) \\ g_i \end{pmatrix}.$$

$M$  and  $v$  have dimensions  $(kp_A + kp_B + k) \times (kp_A + k)$  and  $(kp_A + k) \times 1$ , respectively. By the CLT  $v$  has a normal limiting distribution with mean  $(0', q)'$  and full rank covariance matrix

$$\begin{pmatrix} E(V_{iA}V'_{iA} \otimes Z_iZ'_i) & \Omega_\alpha \\ \Omega'_\alpha & \Omega \end{pmatrix}.$$

Therefore

$$vec(D^*, n^{1/2}\hat{g}(\theta_{\alpha\beta_0})) \rightarrow_d N\left(m + \begin{pmatrix} -\Omega_\alpha\Omega^{-1}q \\ 0 \\ q \end{pmatrix}, \begin{pmatrix} \Psi & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Omega \end{pmatrix}\right), \quad (4.15)$$

where  $\Psi := E(V_{iA}V'_{iA} \otimes Z_iZ'_i) - \Omega_\alpha\Omega^{-1}\Omega'_\alpha$  has full rank. Equation (4.15) proves that  $D^*$  and  $n^{1/2}\hat{g}(\theta_{\alpha\beta_0})$  are asymptotically independent.

I now derive the asymptotic distribution of  $K_\rho^W(\theta_{\alpha\beta_0})$ . Denote by  $\bar{D}$  and  $\bar{g}$  the limiting normal distributions of  $D^*$  and  $n^{1/2}\hat{g}(\theta_{\alpha\beta_0})$ , respectively, see (4.15). Below I show that the function  $h : R^{k \times p} \rightarrow R^{p \times k}$  defined by  $h(D) := (D'\Omega^{-1}D)^{-1/2}D'$  for  $D \in R^{k \times p}$  is continuous on a set  $C \subset R^{k \times p}$  with  $\Pr(\bar{D} \in C) = 1$ . By the Continuous Mapping Theorem we then have

$$(D^*\Omega^{-1}D^*)^{-1/2}D^*\Omega^{-1}n^{1/2}\hat{g}(\theta_{\alpha\beta_0}) \rightarrow_d (\bar{D}'\Omega^{-1}\bar{D})^{-1/2}\bar{D}'\Omega^{-1}\bar{g}. \quad (4.16)$$

By independence of  $\bar{D}$  and  $\bar{g}$ , the latter random variable is distributed as  $W(\alpha, \beta_0) + \zeta$ , where the random  $p$ -vector  $W(\alpha, \beta_0)$  is defined as

$$W(\alpha, \beta_0) := (\bar{D}'\Omega^{-1}\bar{D})^{-1/2}\bar{D}'\Omega^{-1}q, \quad (4.17)$$

where  $\zeta \sim N(0, I_p)$ , and where  $W(\alpha, \beta_0)$  and  $\zeta$  are independent.

I now show that  $h$  is continuous with  $\overline{D}$ -probability one. Note that  $h$  is continuous at each  $D$  that has full column rank. It is therefore sufficient to show that  $\overline{D}$  has full column rank a.s.. From (4.15) we know that the last  $p_B$  columns of  $\overline{D}$  equal  $Q_{ZZ}C_B$  which has full column rank by assumption. Define  $O := \{o \in R^{kp_A} | \exists \tilde{o} \in R^{k \times p_A}, \text{ s.t. } o = \text{vec}(\tilde{o}) \text{ and the } k \times p\text{-matrix } (\tilde{o}, Q_{ZZ}C_B) \text{ has linearly dependent columns}\}$ . Clearly,  $O$  is closed and therefore Lebesgue-measurable. Also  $O$  has empty interior and thus has Lebesgue-measure 0. For the first  $p_A$  columns of  $\overline{D}$ , say  $\overline{D}_{p_A}$ , we know that  $\text{vec}\overline{D}_{p_A}$  is normally distributed with full rank covariance matrix  $\Psi$ . This implies that for any measurable set  $O \subset R^{kp_A}$  with Lebesgue-measure 0, we have  $\Pr(\text{vec}(\overline{D}_{p_A}) \in O) = 0$ , in particular for the  $O$  above. This proves the continuity claim for  $h$ .

Equation (4.16) now immediately implies the asymptotic distribution of  $K_\rho^W(\theta_{\alpha\beta_0})$ .  $\square$

**Proof of Theorem 7.** First note that  $\widehat{g}(\widehat{\theta}_{b_1}) = Q_{ZZ}C_A(\alpha_0 - \widehat{\alpha}) + o_p(1)$  and Lemma 12 imply consistency of  $\widehat{\alpha}$ . Then, using Lemmas 10-12 and the same steps that led to (4.6) in the proof of Theorem 2, it follows that

$$GELR_\rho(\widehat{\theta}_{b_1}) = n^{1/2}\widehat{g}(\widehat{\theta}_{b_1})'\Omega(\theta_{b_1})^{-1}n^{1/2}\widehat{g}(\widehat{\theta}_{b_1}) + o_p(1)$$

(see equation (A.10) in NS for an analogous result under strong identification). An expansion in  $\alpha$  and steps as in the proof of Theorem 3.2. in NS (see the equation between (A.9) and (A.10)), lead to

$$\widehat{g}(\widehat{\theta}_{b_1}) = \widehat{g}(\theta_{b_1}) + (\partial\widehat{g}(\theta)/\partial\alpha)(\widehat{\alpha} - \alpha_0) = (I_k - (EG_{Ai})H)\widehat{g}(\theta_{b_1}) + o_p(n^{-1/2}), \quad (4.18)$$

where  $EG_{Ai} := E\partial g_i(\theta)/\partial\alpha = -Q_{ZZ}C_A$ ,  $H := \Sigma EG'_{Ai}\Omega(\theta_{b_1})^{-1}$ , and  $\Sigma := (EG'_{Ai}\Omega(\theta_{b_1})^{-1}EG_{Ai})^{-1}$ . By Assumption 1(S $_\alpha$ -LA) and the CLT we have

$$n^{1/2}\widehat{g}(\theta_{b_1}) \rightarrow_d N(Q_{ZZ}C_B((\beta_{01} - b_1)', b_2)', \Omega(\theta_{b_1})). \quad (4.19)$$

Note that

$$\Omega(\theta_{b_1})^{-1/2}(I_k - EG_{Ai}H)\Omega(\theta_{b_1})^{1/2} = M_{\Omega(\theta_{b_1})^{-1/2}EG_{Ai}}. \quad (4.20)$$

Therefore,  $GELR_\rho(\widehat{\theta}_{b_1}) \rightarrow_d \varsigma' M_{\Omega(\theta_{b_1})^{-1/2}EG_{Ai}} \varsigma$ , where  $\varsigma \sim N(\Omega(\theta_{b_1})^{-1/2}Q_{ZZ}C_B((\beta_{01} - b_1)', b_2)', I_k)$ , which concludes the proof.  $\square$

**Proof of Theorem 8.** As above we have

$$n^{1/2}\lambda(\widehat{\theta}_{b_1}) = -\Omega(\theta_{b_1})^{-1}n^{1/2}\widehat{g}(\widehat{\theta}_{b_1}) + o_p(1). \quad (4.21)$$

The result for  $K_\rho^W(\widehat{\theta}_{b_1})$  therefore immediately implies the result for  $K_\rho^L(\widehat{\theta}_{b_1})$  and I only deal with  $K_\rho^W(\widehat{\theta}_{b_1})$ .

As in the proof of Corollary 6(ii) renormalize by  $D^* := D_{\rho(b'_1, \beta'_{02})'}(\widehat{\alpha})\Delta$ , where  $\Delta := \text{diag}(n^{1/2}, \dots, n^{1/2}, 1, \dots, 1)$  is a diagonal  $p_B \times p_B$ -matrix whose first  $p_{B_1}$  diagonal elements equal  $n^{1/2}$  and the remaining  $p_{B_2}$  elements equal 1. We now show that  $D^*$

and  $n^{1/2}\widehat{g}(\widehat{\theta}_{b_1})$  are asymptotically independent. Proceeding exactly as in the proof of Corollary 6(ii) and using (4.18) it follows that

$$\text{vec}(D^*, n^{1/2}\widehat{g}(\widehat{\theta}_{b_1})) = m + Mv + o_p(1), \quad (4.22)$$

where

$$\begin{aligned} M & : = \begin{pmatrix} I_{kp_{B_1}} & -\Omega_{b_1}\Omega(\theta_{b_1})^{-1} \\ 0 & 0 \\ 0 & I_k \end{pmatrix} \begin{pmatrix} I_{kp_{B_1}} & 0 \\ 0 & (I_k - (EG_{A_i})H) \end{pmatrix}, \\ v & : = n^{-1/2} \sum_{i=1}^n \begin{pmatrix} \text{vec}(Z_i V'_{iB_1}) \\ g_i(\theta_{b_1}) \end{pmatrix}, m := \text{vec}(Q_{ZZ}C_B, 0), \end{aligned}$$

where  $EG_{A_i}$  and  $H$  are defined in the proof of Theorem 7 and where  $\Omega_{b_1} := \lim_{n \rightarrow \infty} E \text{vec}(Z_i V'_{iB_1}) g_i(\theta_{b_1})'$ . By the CLT,  $v$  is asymptotically normal with full rank covariance matrix

$$\text{Cov} := \begin{pmatrix} E(V_{iB_1} V'_{iB_1} \otimes Z_i Z'_i) & \Omega_{b_1} \\ \Omega'_{b_1} & \Omega(\theta_{b_1}) \end{pmatrix}.$$

For independence of  $D^*$  and  $n^{1/2}\widehat{g}(\widehat{\theta}_{b_1})$  we have to show that the right-upper block of the asymptotic covariance matrix  $M \text{Cov} M'$  of  $\text{vec}(D^*, n^{1/2}\widehat{g}(\widehat{\theta}_{b_1}))$  equals 0, i.e. we have to show that

$$[\Omega_{b_1} - \Omega_{b_1}\Omega(\theta_{b_1})^{-1}(I_k - (EG_{A_i})H)\Omega(\theta_{b_1})](I_k - (EG_{A_i})H)' = 0. \quad (4.23)$$

Let  $S := \Omega(\theta_{b_1})^{-1/2}EG_{A_i}$ . Using (4.20) it follows that the left-hand side matrix in (4.23) equals  $\Omega_{b_1}\Omega(\theta_{b_1})^{-1/2}P_S M_S \Omega(\theta_{b_1})^{1/2}$  which clearly equals 0. This proves the independence claim.

Now denote by  $\bar{D}$  and  $\bar{g}$  the limiting normal distributions of  $D^*$  and  $n^{1/2}\widehat{g}(\widehat{\theta}_{b_1})$ , implied by (4.22). Let  $M(b_1) := M((b'_1, \beta'_{02})')$ . If the function  $h : R^{k \times p_B} \rightarrow R^{p_B \times k}$  defined by  $h(D) := (D' M(b_1) D)^{-1/2} D'$  for  $D \in R^{k \times p_B}$  is continuous on a set  $C \subset R^{k \times p_B}$  with  $\Pr(\bar{D} \in C) = 1$ , then by the Continuous Mapping Theorem

$$(D^{*'} M(b_1) D^*)^{-1/2} D^{*'} \Omega(\theta_{b_1})^{-1} n^{1/2} \widehat{g}(\widehat{\theta}_{b_1}) \rightarrow_d (\bar{D}' M(b_1) \bar{D})^{-1/2} \bar{D}' \Omega(\theta_{b_1})^{-1} \bar{g}.$$

The latter is distributed as  $W(b_1, b_2) + \zeta$ , where

$$W(b_1, b_2) := (\bar{D}' M(b_1) \bar{D})^{-1/2} \bar{D}' \Omega(\theta_{b_1})^{-1/2} M_S \Omega(\theta_{b_1})^{-1/2} Q_{ZZ} C_B ((\beta_{01} - b_1)', b'_2)'. \quad (4.24)$$

Regarding continuity of  $h$  it is enough to show that with  $\bar{D}$ -probability one,  $M(b_1)\bar{D}$  has full column rank  $p_B$  or equivalently that  $\text{rank}(M_S \Omega(\theta_{b_1})^{-1/2} \bar{D}) = p_B$ . Because  $\ker M_S = S$  and  $\text{rank}(S) = p_A$ , the latter condition holds if  $(EG_{A_i}, \bar{D}) = p$ . Now  $EG_{A_i} = -Q_{ZZ}C_A$  and the last  $p_2$  columns of  $\bar{D}$ ,  $\bar{D}_{p_2}$  say, equal  $Q_{ZZ}C_{B_2}$ . By assumption, the matrix  $(EG_{A_i}, \bar{D}_{p_2})$  has rank  $p_A + p_{B_2}$  and we have to show that with  $\bar{D}$ -probability one, the first  $p_1$  columns of  $\bar{D}$ ,  $\bar{D}_{p_1}$  say, increase the rank of this matrix to  $p$ . Using (4.22), the covariance matrix of  $\bar{D}_{p_1}$  is easily shown to equal  $E(V_{iB_1} V'_{iB_1} \otimes Z_i Z'_i) - \Omega_{b_1}\Omega(\theta_{b_1})^{-1/2} M_S \Omega(\theta_{b_1})^{-1/2} \Omega'_{b_1}$  and to be of full column rank. An argument analogous to the last step in the proof of Corollary 6(ii) can then be applied to conclude the proof.  $\square$



## 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. (2002a): “Pivotal statistics for testing structural parameters in instrumental variables regression”, *Econometrica* 70(5), 1781-1805.
- Kleibergen, F. (2002b): “Two independent pivotal statistics that test location and misspecification and add-up to the Anderson-Rubin statistic”, working paper.
- 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.

- Qin J, and J. Lawless (1994): “Empirical Likelihood and general estimating equations”, *Annals of Statistics* 22(1), 300-325.
- 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**

Size results for Design (I) at 5% significance level. Strong instrument  $\Pi_1 = 1$

$n$	$k$	$\sigma$	2SLS				$GELR_\rho$				
			HOM	HET	K	$LR_M$	CUE	EL	$K_{CUE}^W$	$K_{EL}^W$	$K_{EL}^L$
50	1	.0	4.9*	6.0	5.7	5.4	4.7	6.7	4.7	4.7	8.9
		.5	5.1*	6.3	5.5	5.3	4.8	6.9	4.8	4.8	8.9
		.99	5.8	6.7	5.1*	5.1*	4.2	6.4	4.2	4.2	8.3
	5	.0	3.9	5.3*	5.9	6.2	2.8	17.8	2.6	4.2	15.1
		.5	5.8	7.1	5.3*	5.4	2.5	17.5	2.4	4.1	14.8
		.99	12.9	14.2	5.8	5.7*	2.7	17.6	2.7	4.3*	15.7
	10	.0	3.2	4.2	6.2	6.4	1.4	44.6	1.8	4.3*	27.2
		.5	8.5	10.0	5.6*	5.7	1.4	44.2	1.9	4.4*	26.3
		.99	28.4	30.5	5.8*	5.8*	1.6	45.4	1.4	3.7	25.1
100	1	.0	4.6	5.4	5.2*	5.3	4.6	5.6	4.6	4.6	6.3
		.5	5.0*	5.8	5.4	5.4	5.1	6.2	5.1	5.1	6.8
		.99	5.3	5.9	5.0*	4.9	4.5	5.6	4.5	4.5	6.3
	5	.0	4.7	5.4	5.6	5.8	3.9	10.8	3.9	5.0*	9.3
		.5	5.4	6.1	5.1*	5.3	3.6	10.3	3.5	4.7	9.5
		.99	9.2	9.7	5.6	5.2*	3.9	10.5	3.7	4.8*	9.2
	10	.0	4.2	4.8*	5.5	5.2*	2.7	21.1	2.7	4.7	14.1
		.5	7.3	8.0	5.4*	5.4*	3.0	21.7	2.5	4.4	13.3
		.99	18.6	19.8	5.3	5.1*	2.3	21.4	2.6	4.5	13.3
250	1	.0	5.0*	5.5	5.2	5.0*	5.2	5.6	5.2	5.2	5.6
		.5	5.1*	5.4	5.2	4.8	5.3	5.6	5.3	5.3	5.5
		.99	4.9*	5.4	5.2	5.2	5.1*	5.5	5.1*	5.1*	5.4
	5	.0	4.8	5.1*	5.2	5.4	4.6	7.1	4.2	4.8	6.1
		.5	5.0*	5.3	4.9	5.2	4.2	6.3	4.2	4.8	5.9
		.99	6.9	7.3	5.1*	5.2	4.6	6.7	4.3	4.9*	6.2
	10	.0	4.6	5.0*	5.2	5.1	4.3	9.9	3.7	4.9	7.6
		.5	6.0	6.2	5.0*	4.9	3.8	9.8	3.4	4.7	7.2
		.99	10.7	10.9	5.1*	4.8	4.0	9.5	3.5	4.8	7.7

Notes: 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 I** (continued)

Size results for Design (I) at 5% significance level. Weak instrument  $\Pi_1 = .1$

$n$	$k$	$\sigma$	<i>2SLS</i>				<i>GELR<math>_{\rho}</math></i>				
			<i>HOM</i>	<i>HET</i>	<i>K</i>	<i>LR<math>_M</math></i>	<i>CUE</i>	<i>EL</i>	$K_{CUE}^W$	$K_{EL}^W$	$K_{EL}^L$
50	1	.0	0.1	0.3	5.7	5.4	4.7*	6.7	4.7*	4.7*	8.9
		.5	2.2	3.0	5.5	5.3	4.8*	6.9	4.8*	4.8*	8.9
		.99	24.7	25.7	5.1*	5.1*	4.2	6.4	4.2	4.2	8.3
	5	.0	0.6	1.2	6.6	7.3	2.8	17.8	3.7	5.5*	17.1
		.5	16.5	18.9	6.8	7.3	2.5	17.5	3.7	5.4*	17.0
		.99	96.5	96.6	5.8	6.1	2.7	17.6	2.8	4.3*	15.6
	10	.0	0.9	2.1	8.5	9.2	1.4	44.6	3.1	6.0*	30.1
		.5	33.7	36.9	8.2	9.3	1.4	44.2	3.2	6.3*	30.6
		.99	100.0	100.0	6.7	7.4	1.6	45.4	2.0	4.6*	27.6
100	1	.0	0.1	0.2	5.2*	5.3	4.6	5.6	4.6	4.6	6.3
		.5	2.6	3.0	5.4	5.4	5.1*	6.2	5.1*	5.1*	6.8
		.99	18.5	19.0	5.0*	4.9	4.5	5.6	4.5	4.5	6.3
	5	.0	0.6	0.9	5.9	6.1	3.9	10.8	4.3	5.6*	10.7
		.5	17.0	18.3	5.6	6.2	3.6	10.3	4.2	5.5*	10.3
		.99	92.7	92.8	5.6	5.5	3.9	10.5	3.8	4.9*	9.2
	10	.0	1.3	2.0	6.8	6.6	2.7	21.1	3.4	6.2*	16.1
		.5	36.6	37.5	6.5	6.9	3.0	21.7	3.7	5.9*	15.7
		.99	99.8	99.8	5.5	5.4*	2.3	21.4	2.5	4.5	14.0
250	1	.0	0.3	0.3	5.2	5.0*	5.2	5.6	5.2	5.2	5.6
		.5	3.2	3.5	5.2*	4.8*	5.3	5.6	5.3	5.3	5.5
		.99	13.0	13.3	5.2	5.2	5.1*	5.5	5.1*	5.1*	5.4
	5	.0	0.7	0.8	5.1*	5.7	4.6	7.1	4.4	5.1*	6.5
		.5	15.5	16.0	5.2*	5.4	4.2	6.3	4.7	5.4	6.6
		.99	80.1	80.3	5.1*	5.3	4.6	6.7	4.3	4.9*	6.5
	10	.0	1.6	1.9	5.4*	6.0	4.3	9.9	4.1	5.4*	8.2
		.5	34.3	34.9	5.6	5.5*	3.8	9.8	4.4	5.9	8.4
		.99	99.0	99.0	5.2	4.7	4.0	9.5	3.5	5.0*	7.6

Notes: Asterisks in 5.2each 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. Strong instrument  $\Pi_1 = 1$

$n$	$k$	$\sigma$	2SLS				$GELR_\rho$					
			HOM	HET	K	$LR_M$	CUE	EL	$K_{CUE}^W$	$K_{EL}^W$	$K_{EL}^L$	
50	1	.0	24.7	7.6	26.8	26.3	3.9*	9.6	3.9*	3.9*	16.6	
		.5	23.7	7.7	26.6	26.3	3.9*	9.6	3.9*	3.9*	16.3	
		.99	22.9	8.3	26.0	26.1	3.5*	9.2	3.5*	3.5*	16.2	
	5	.0	7.7	5.8*	11.0	12.1	2.0	23.4	2.4	4.1	20.1	
		.5	9.9	7.6	10.7	11.6	2.0	22.4	2.4	3.9*	18.9	
		.99	18.1	14.3	11.2	11.4	2.1	22.9	2.6	4.1*	20.3	
	10	.0	4.7*	4.5	9.3	10.3	1.1	49.4	1.9	4.4	30.3	
		.5	10.6	10.0	8.8	9.4	1.3	49.8	1.9	4.4*	29.2	
		.99	32.1	29.9	8.8	8.9	1.4	50.3	1.4	3.7*	27.9	
100	1	.0	25.3	6.2	26.4	26.6	4.3*	7.1	4.3*	4.3*	11.1	
		.5	25.6	6.8	26.9	26.8	4.5*	8.2	4.5*	4.5*	12.1	
		.99	24.0	7.0	25.5	25.2	4.5*	7.7	4.5*	4.5*	11.2	
	5	.0	8.8	5.8	10.3	11.0	3.3	14.4	3.6	4.8*	12.3	
		.5	9.6	6.6	9.9	10.3	3.1	14.1	3.5	4.5*	12.5	
		.99	14.1	10.1	10.5	10.1	3.5	13.8	3.6	4.5*	12.7	
	10	.0	6.3	5.0*	8.3	8.1	2.4	25.1	2.7	4.6	16.5	
		.5	9.6	8.0	8.1	8.2	2.7	25.9	2.6	4.4*	15.9	
		.99	22.0	19.2	7.9	7.7	2.2	26.0	2.5	4.5*	15.8	
	250	1	.0	25.3	5.8	25.7	25.4	4.7*	6.3	4.7*	4.7*	7.7
			.5	26.4	5.8	26.6	26.0	5.0*	6.3	5.0*	5.0*	7.8
			.99	25.4	5.9	26.0	26.1	4.9*	6.3	4.9*	4.9*	7.5
5		.0	9.3	5.1*	9.9	10.2	4.1	8.5	4.0	4.6	7.6	
		.5	9.5	5.6	9.7	10.3	4.0	7.9	4.1	4.6*	7.9	
		.99	11.5	7.3	10.1	10.5	4.1	8.3	4.3	5.0*	8.1	
10		.0	6.8	4.8	7.6	7.7	3.9	12.6	3.6	5.0*	9.1	
		.5	8.4	6.3	7.5	7.5	3.5	12.0	3.3	4.7*	8.8	
		.99	13.7	10.9	7.9	7.4	3.7	11.7	3.5	4.8*	9.2	

Notes: 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** (continued)

Size results for Design ( $I_{HET}$ ) at 5% significance level. Weak instrument  $\Pi_1 = .1$

$n$	$k$	$\sigma$	<i>2SLS</i>				<i>GELR<math>_{\rho}</math></i>				
			<i>HOM</i>	<i>HET</i>	<i>K</i>	<i>LR<math>_M</math></i>	<i>CUE</i>	<i>EL</i>	$K_{CUE}^W$	$K_{EL}^W$	$K_{EL}^L$
50	1	.0	0.9	0.4	26.8	26.3	3.9*	9.6	3.9*	3.9*	16.6
		.5	4.4*	3.0	26.6	26.3	3.9	9.6	3.9	3.9	16.3
		.99	23.4	24.5	26.0	26.1	3.5*	9.2	3.5*	3.5*	16.2
	5	.0	1.4	1.5	12.2	18.5	2.0	23.4	3.9	5.6*	22.5
		.5	20.4	18.0	12.7	18.7	2.0	22.4	3.6	5.3*	22.2
		.99	94.7	93.3	18.1	21.2	2.1	22.9	2.8	4.9*	22.8
	10	.0	1.5	2.4	11.9	17.1	1.1	49.4	3.1	6.1*	33.5
		.5	36.5	35.8	12.5	17.0	1.3	49.8	3.2	6.5*	34.2
		.99	100.0	99.9	17.9	21.4	1.4	50.3	2.3	5.7*	32.2
100	1	.0	1.1	0.2	26.4	26.6	4.3*	7.1	4.3*	4.3*	11.1
		.5	6.1	2.9	26.9	26.8	4.5*	8.2	4.5*	4.5*	12.1
		.99	24.4	18.5	25.5	25.2	4.5*	7.7	4.5*	4.5*	11.2
	5	.0	1.4	0.9	10.7	17.0	3.3	14.4	4.3	5.6*	14.0
		.5	21.7	17.6	11.2	17.0	3.1	14.1	4.1	5.4*	14.1
		.99	92.0	89.0	15.0	18.1	3.5	13.8	3.5	5.0*	13.7
	10	.0	2.1	1.9	9.6	13.4	2.4	25.1	3.3	6.0*	18.9
		.5	40.0	36.5	9.2	14.4	2.7	25.9	3.5	6.0*	18.4
		.99	99.7	99.6	13.8	15.4	2.2	26.0	2.7	5.3*	18.5
250	1	.0	3.0	0.3	25.7	25.4	4.7*	6.3	4.7*	4.7*	7.7
		.5	9.3	3.2	26.6	26.0	5.0*	6.3	5.0*	5.0*	7.8
		.99	23.2	12.6	26.0	26.1	4.9*	6.3	4.9*	4.9*	7.5
	5	.0	1.8	0.9	10.1	15.8	4.1	8.5	4.3	5.2*	8.1
		.5	20.8	14.8	10.5	15.4	4.0	7.9	4.3	5.0*	7.9
		.99	81.5	76.0	12.3	14.3	4.1	8.3	4.2	5.1*	8.2
	10	.0	2.5	2.0	7.8	12.5	3.9	12.6	4.1	5.5*	9.9
		.5	38.9	33.9	8.4	11.7	3.5	12.0	4.5*	5.8	10.3
		.99	98.8	98.3	10.3	10.4	3.7	11.7	3.4	5.0*	9.3

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