Testing overidentifying restrictions with a restricted parameter space

Philipp Ketz

September 26, 2019

We show that the standard test for testing overidentifying restrictions, which compares the J-statistic (Hansen, 1982) to $\chi^2$ critical values, does not control asymptotic size when the true parameter vector is allowed to lie on the boundary of the (optimization) parameter space. We also propose a modified J-statistic that, under the null hypothesis, is asymptotically $\chi^2$ distributed, such that the resulting test does control asymptotic size.

Keywords: Asymptotic size, boundary, overidentifying restrictions.

JEL classification: C12, C52
1 Introduction

Economic models often imply certain restrictions on the distribution of the data. When the number of restrictions, $H$, exceeds the number of unknown parameters, $p$, the additional, so-called overidentifying, restrictions allow to test the model’s validity, at least partially. In the context of the Generalized Method of Moments (GMM), this is often done using the J-statistic (Hansen, 1982), i.e., the GMM objective function evaluated at its minimizer. In standard settings, its asymptotic null distribution is given by a $\chi^2$ distribution with the degrees of freedom equal to the number of overidentifying restrictions, $H - p$. One of the assumptions underlying this standard result is that the true parameter vector lies in the interior of the (optimization) parameter space, such that the J-statistic asymptotically behaves like a quadratic function evaluated at an *unconstrained* minimizer. If, however, the true parameter vector lies on the boundary, then the J-statistic asymptotically behaves like a quadratic function evaluated at a *constrained* minimizer.\footnote{The same holds true under drifting sequences of true parameters that are such that the true parameter vector is close to the boundary relative to the sample size.}\footnote{Strictly speaking, the asymptotic behavior of the J-statistic also depends on the shape of the parameter space around the true parameter vector. See Section 2, in particular Assumption 5, for more details.} Therefore, ceteris paribus the J-statistic is (weakly) greater when the true parameter vector lies on the boundary. As a result, the standard overidentification test that relies on $\chi^2(H - p)$ critical values does not control asymptotic size, i.e., the test’s asymptotic size exceeds the nominal level, when the true parameter vector is allowed to lie on the boundary. In practice, this increases the likelihood of mistakenly discarding a valid model. A prominent example of where this may be an issue is the random coefficients logit model (Berry, Levinsohn, and Pakes, 1995), where the boundary of the parameter space is of considerable empirical relevance; see Ketz (2019) for examples.\footnote{While overidentifying restrictions are seldom tested in the context of the random coefficients logit model, some authors do so without mentioning the potential consequences of a restricted parameter space, see e.g., Bonnet and Dubois (2010).} In this paper, we build on the insight in Ketz (2018) that (locally) ignoring the restrictions on the parameter space can be useful for testing. In particular, we propose a modified J-statistic, which is given by a quadratic approximation of the objective function evaluated at its minimizer.\footnote{Note that this amounts to (locally) ignoring the restrictions on the parameter space, as this minimizer can lie outside the parameter space.} The use of a quadratic approximation of the objective function is motivated by the possibility that the latter may not be defined outside the parameter space, as e.g., the case in the random coefficients logit model. We show that, under the null hypothesis, the modified J-statistic is asymptotically $\chi^2(H - p)$ distributed, i.e., it has the standard asymptotic null distribution, irrespective of where the true parameter vector lies. An immediate consequence is that the corresponding test that uses $\chi^2(H - p)$ critical values is asymptotically similar and, thus, controls...
asymptotic size. In addition, the test has the advantage of being easy to implement, as it avoids the potentially computationally costly adaptation of critical values, e.g., using the least favorable configuration approach (see e.g., Andrews and Cheng, 2012) or Bonferroni-based approaches (see e.g., McCloskey, 2017), that would be required to ensure that a test based on the standard J-statistic controls asymptotic size.

The results in this paper build on the results in Andrews (1999, 2002), where a quadratic expansion of the objective function is used to derive the asymptotic distribution of extremum estimators when the true parameter vector lies on the boundary of the parameter space. In particular, we apply the results in Andrews (1999, 2002) to derive (features of) the asymptotic null distribution of the standard J-statistic when the true parameter vector lies on the boundary. This complements Andrews (2001), who derives the asymptotic null distribution of (different versions of) the trinity of test statistics when parameter vectors in the null hypothesis lie on the boundary and when, in addition, nuisance parameters may appear under the alternative, but not under the null, hypothesis. Furthermore, the quadratic expansion of the objective function used in Andrews (1999, 2002) is leveraged in the construction of the proposed test statistic.

The rest of the paper is organized as follows. Section 2 introduces the framework and the testing problem. Section 3 formally establishes the results concerning the (tests based on the) standard J-statistic and the modified J-statistic, while Section 4 concludes. Throughout, the linear instrumental variables (IV) model serves as a running example. All proofs are collected in Section A of the supplementary material (SM).

Throughout this paper, “≡” denotes “equals by definition.” Let \( M_A = I_k - A(A'A)^{-1}A' \) denote the projection matrix onto the space orthogonal to the column space of the \( k \times l \) matrix \( A \), where \( I_k \) denotes the \( k \)-dimensional identity matrix. Let \( \lambda_{\min}(A) \) denote the smallest eigenvalue of the (square) matrix \( A \). For any set \( A \subset \mathbb{R} \), let \( A^k \equiv A \times \cdots \times A \) with \( k \) copies. Furthermore, \( \text{int}(A) \), \( \text{bd}(A) \), and \( \text{cl}(A) \) denote the interior, the boundary, and the closure of the set \( A \subset \mathbb{R}^p \), respectively. Let \( 0_k \equiv (0, \ldots, 0)' \) with \( k \) entries and let \( \| \cdot \| \) denote the Euclidean norm. Lastly, \( \xrightarrow{p} \) and \( \xrightarrow{d} \) denote convergence in probability and distribution, respectively, while all limits are taken as “\( n \to \infty \).”

2 Testing problem

Throughout this paper, we borrow notation from Andrews and Cheng (2012, 2014). Let

\[
G_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} g(\theta, w_i)
\]
denote the sample moments, where \( g(\theta, w_i) \) denotes a \( H \)-dimensional vector function that depends on the \( p \)-dimensional vector \( \theta \) and the data vector \( w_i \), with \( H \geq p \). The GMM objective function is given by

\[
Q_n(\theta) \equiv G_n(\theta)'W_n(\theta)G_n(\theta)/2,
\]

where \( W_n(\theta) \) denotes a symmetric weighting matrix that may depend on \( \theta \). Define an estimator, \( \hat{\theta}_n \), as any random variable that satisfies \( \hat{\theta}_n \in \Theta \) and

\[
Q_n(\hat{\theta}_n) = \inf_{\theta \in \Theta} Q_n(\theta) + o_p(1),
\]

where \( \Theta \subset \mathbb{R}^p \) is compact and denotes the optimization parameter space. The true distribution of the data \( \{w_i\}_{i=1}^n \) is denoted \( F_\gamma \) for some parameter \( \gamma \in \Gamma \). In what follows, \( P_\gamma \) and \( E_\gamma \) denote the probability and expectation under \( F_\gamma \), respectively. The parameter space for the true parameter \( \gamma, \Gamma \), is assumed to be of the following form

\[
\Gamma = \{ \gamma = (\theta, \phi) : \theta \in \Theta, \phi \in \Phi(\theta) \},
\]

where \( \Theta \subset \bar{\Theta} \) and \( \Phi(\theta) \subset \Phi \forall \theta \in \Theta \) for some compact metric space \( \Phi \) with a metric that induces weak convergence of \( (w_i, w_{i+m}) \) for all \( i, m \geq 1 \). Throughout this paper, we refer to \( \theta \), when \( \gamma = (\theta, \phi) \in \Gamma \) is the true parameter, as the true parameter vector.

Let \( G(\theta; \gamma^*) = E_\gamma^*g(\theta, w_i) \) for \( \gamma^* = (\theta^*, \phi^*) \in \Gamma \). For expositional purposes, we assume the data \( \{w_i\}_{i=1}^n \) to be identically, but not necessarily independently, distributed such that the testing problem of interest can be written as follows

\[
H_0 : \gamma \in \Gamma_0 \text{ vs. } H_1 : \gamma \in \Gamma \setminus \Gamma_0,
\]

where \( \gamma \) is the true parameter and where

\[
\Gamma_0 = \{ \gamma \in \Gamma : G(\theta; \gamma) = 0_H \}
\]

is assumed to be compact and referred to as the null parameter space. In words, we want to test whether the (population) moment conditions hold at the true parameter vector.\(^7\) As our focus lies on asymptotic size properties, we, in what follows,

\(^5\)Under appropriate dependence conditions, \( g(\theta, w_i) \) can also be replaced by \( g_i(\theta) \) where the latter depends on \( \{w_j\}_{j=1}^i \), cf. footnote 32 in Andrews and Cheng (2012).

\(^6\)Here, we implicitly assume that the parameter vector that enters the objective function, \( \theta \), also governs the distribution of the data, i.e., \( \phi \) alone does not fully specify the distribution of the data; see Example 1 for an illustrative example.

\(^7\)Note that, given our terminology, the moment conditions can be misspecified at the true parameter (vector), i.e., \( G(\theta; \gamma) \neq 0_H \), which differs from some other sources, where the true parameter vector by definition satisfies the moment conditions.

\(^8\)To see in how far this corresponds to testing the overidentifying restrictions, note that Assumptions
only make assumptions pertaining to the null parameter space, $\Gamma_0$, without further specifying $\Gamma$. For the following example(s), it is convenient to define $\Phi_0(\theta)$ such that $\Gamma_0 = \{\gamma = (\theta, \phi) : \theta \in \Theta, \phi \in \Phi_0(\theta)\}$.

**Example 1—Linear IV model:** Let $z_i \in \mathbb{R}^d$ denote the instruments and let $y_i = x_i^\prime \theta + u_i$, where $y_i \in \mathbb{R}$ denotes the outcome variable, $x_i \in \mathbb{R}^p$ the potentially endogenous regressors, and $u_i \in \mathbb{R}$ the error term. Take $g(\theta, w_i) = z_i(y_i - x_i^\prime \theta)$ and consider the “standard” (efficient) two-step GMM estimator, i.e., take $\mathcal{W}_n \equiv \mathcal{W}_n(\theta) = (\frac{1}{n} \sum_{i=1}^{n} g(\theta_n, w_i)g(\theta_n, w_i))^{-1}$, where $\theta_n \in \Theta$ denotes a first-stage estimator obtained using $\mathcal{W}_n(\theta) = (\frac{1}{n} \sum_{i=1}^{n} z_i z_i^\prime)^{-1}$. For the sake of concreteness, assume that $\Theta = [0, \hat{\epsilon}] \times [-\hat{\epsilon}, \hat{\epsilon}]^{p-1}$ for some $0 < \hat{\epsilon} < \infty$. Furthermore, take $\tilde{\Theta} = [0, \hat{\epsilon}]$ for some $\hat{\epsilon} \leq \hat{\epsilon} < \infty$, i.e., think of $\theta_1$ as restricted below by zero, but unrestricted from above, and of $(\theta_2, \ldots, \theta_p)^\prime$ as unrestricted. For example, $y_i$ may be demand, $x_{i1}$ the price of a giffen good, and $(x_{i2}, \ldots, x_{ip})^\prime$ a set of controls. Assume that the data vector $w_i = (z_i^\prime, x_i^\prime, y_i)^\prime$ is iid across $i$ and let $\phi \in \Phi$ denote the distribution of $\{z_i, x_i, u_i\}$, where $\Phi$ is a compact metric space with some metric that induces weak convergence, such as the uniform metric. The null parameter space, $\Gamma_0$, is defined through

$$\Phi_0(\theta) = \{\phi \in \Phi : \Omega \equiv E_\phi z_i z_i^\prime u_i^2; W \equiv E_\phi z_i z_i^\prime; D \equiv E_\phi z_i x_i^\prime; \lambda_{\min}(\Omega) \geq \delta; \lambda_{\min}(W) \geq \delta; \lambda_{\min}(D^\prime D) \geq \delta; E_\phi z_i u_i = 0_H; E_\phi \|z_i\|^{4+\epsilon} + E_\phi \|x_i\|^{4+\epsilon} + E_\phi |u_i|^{4+\epsilon} \leq C\}$$

for some $0 < \delta, C < \infty$. Note that, here, $\Phi_0 \equiv \Phi_0(\theta)$ does not depend on $\theta$.

**Example 2—Random coefficients logit model:** The most prevalent version of the random coefficients logit model assumes that the random coefficients are independently normally distributed, such that $\theta$ is given by a $(K_1 \times 1)$ vector of means, say $\mu$, and a $(K_2 \times 1)$ vector of variances, say $\sigma^2$. Here, $\tilde{\Theta}$ is given by $[-\hat{\epsilon}, \hat{\epsilon}]^{K_1} \times [0, \hat{\epsilon}]^{K_2}$ for some $0 < \hat{\epsilon} < \infty$ and $\Theta$ may be taken equal to $[-\hat{\epsilon}, \hat{\epsilon}]^{K_1} \times [0, \hat{\epsilon}]^{K_2}$ for some $0 < \hat{\epsilon} < \hat{\epsilon}$. In the interest of space, details are relegated to Section C of the SM. We note, however, that in the random coefficients logit model $g(\theta, w_i)$ and, thus, $Q_n(\theta)$ are not defined for $\theta \not\in \tilde{\Theta}$.

In what follows, we present the assumptions under which the results in this paper are derived. They allow for non-smooth objective functions as in Pakes and Pollard (1989) and are similar to the assumptions in Andrews (2002), allowing for the true parameter vector to be (near or) at the boundary and for the objective function to not be defined

1 and 3(ii)-(iii) imply that $G_\phi^W G(\theta; \gamma) = 0_p$ for all $\gamma \in \Gamma_0$ for which $\theta \in \text{int}(\Theta)$, such that, at least for those $\gamma$, $G(\theta; \gamma) = 0_H \Leftrightarrow M_{\theta^H/2} W^{1/2} G(\theta; \gamma) = 0_H$.

We assume $\hat{\epsilon} < \infty$ in accordance with the high-level assumption that $\Theta$ is compact. Note, however, that this assumption could be relaxed in the context of the linear model.
outside the parameter space. The assumptions are stated for drifting sequences of true parameters (in the null parameter space) following Andrews and Cheng (2012, 2014), allowing us to apply the high-level results in Andrews, Cheng, and Guggenberger (2019) concerning asymptotic size. Define

$$\Gamma_0(\gamma^*) = \{ \gamma_n = (\theta_n, \phi_n) \in \Gamma_0 : n \geq 1 \} : \gamma_n \to \gamma^* \in \Gamma_0 \}.$$ 

Throughout this paper, we use the terminology “under \(\{\gamma_n\} \in \Gamma_0(\gamma^*)\)” to mean “when the true parameters are \(\{\gamma_n\} \in \Gamma_0(\gamma^*)\) for any \(\gamma^* \in \Gamma_0\).”

Assumption 1 ensures consistency of \(\hat{\theta}_n\) and, thus, rules out weak instruments or, more generally, weak identification (see e.g., Staiger and Stock, 1997; Stock and Wright, 2000). Note that the results in this paper do not apply when weak identification is allowed for.

**Assumption 1.**

1. Under \(\{\gamma_n\} \in \Gamma_0(\gamma^*)\), \(\sup_{\theta \in \Theta} \|G_n(\theta) - G(\theta; \gamma^*)\| \overset{p}{\rightarrow} 0\) and \(\sup_{\theta \in \Theta} \|W_n(\theta) - W(\theta; \gamma^*)\| \overset{p}{\rightarrow} 0\), where \(W(\theta; \gamma^*)\) denotes some nonrandom function.

2. Each element of \(G(\theta; \gamma^*)\) has continuous left/right \((l/r)\) partial derivatives on \(\bar{\Theta}\), \(\forall \gamma^* \in \Gamma_0\), with the \(H \times p\) matrix of \(l/r\) partial derivatives denoted \(G_\theta(\theta; \gamma^*)\).

3. \(W(\theta; \gamma^*)\) is continuous in \(\theta\) on \(\bar{\Theta}\), \(\forall \gamma^* \in \Gamma_0\).

4. \(Q(\theta; \gamma^*) \equiv G(\theta; \gamma^*)'W(\theta; \gamma^*)G(\theta; \gamma^*)\) is uniquely minimized by \(\theta^*\) over \(\bar{\Theta}\), \(\forall \gamma^* \in \Gamma_0\).

Assumptions 1(i)-(iii) can often be verified using a uniform law of large numbers, see e.g., Andrews (1992). Note that the results in this paper also extend to Minimum Distance estimation under suitable adaptation of the assumptions, cf. Andrews and Cheng (2014).

Assumption 2 is a stochastic equicontinuity assumption used to obtain a quadratic approximation of the objective function, cf. Pakes and Pollard (1989) and Andrews (2002). A simple sufficient condition is available when the sample moment has continuous \((l/r)\) partial derivatives, see e.g., Assumption 2° in Ketz (2019).

**Assumption 2.** Under \(\{\gamma_n\} \in \Gamma_0(\gamma^*)\),

$$\sup_{\theta \in \Theta; \|\theta - \theta_n\| \leq \epsilon_n} \frac{\sqrt{n}\|G_n(\theta) - G(\theta; \gamma^*) - G_n(\theta_n) + G(\theta_n; \gamma^*)\|}{1 + \|\sqrt{n}(\theta - \theta_n)\|} = o_p(1)$$

for all sequences of positive scalar constants \(\{\epsilon_n : n \geq 1\}\) for which \(\epsilon_n \to 0\).

---

10A function is said to have left/right partial derivatives if it has partial derivatives at each interior point, partial derivatives at each boundary point with respect to coordinates that can be perturbed to the left and the right, and left (right) partial derivatives at each boundary point with respect to coordinates that can be perturbed only to the left (right), see Section 3.3 in Andrews (1999).
Assumption 3 requires a central limit theorem to hold for the (scaled) sample moments. Furthermore, it requires that the model is first-order locally identified.\(^{11}\)

**Assumption 3.**

(i) Under \(\{ \gamma_n \} \in \Gamma_0(\gamma^*)\), \(\sqrt{n} G_n(\theta_n) \xrightarrow{d} N(0_H, \Omega(\gamma^*))\) for some symmetric and positive-definite matrix \(\Omega(\gamma^*)\).

(ii) \(G_\theta \equiv G_\theta(\theta; \gamma)\) has full column rank, \(\forall \gamma \in \Gamma_0\).

(iii) \(W \equiv W(\theta; \gamma)\) is nonsingular, \(\forall \gamma \in \Gamma_0\).

Note that the “centering” in Assumption 3(i) is implicit, as \(G(\theta_n; \gamma_n) = 0\) for \(\gamma_n \in \Gamma_0\).

The next assumption states that an optimal weighting matrix is used.

**Assumption 4.** \(W = \Omega(\gamma)^{-1}\) for all \(\gamma \in \Gamma_0\).

**Example 1 continued:** The verification of Assumptions 1–4, given the conditions in (2), is provided in Section B of the SM. Here, we have \(G(\theta^*; \gamma^*) = E_{\phi^*} z_i x_i' \theta^* - \theta\), \(W^{-1} = \Omega(\gamma^*) = E_{\phi^*} z_i z_i' u_i^2\), and \(G_\theta = -E_{\phi^*} z_i x_i'\) for \(\gamma^* \in \Gamma_0\).

### 3 Results

Hansen’s J-statistic (Hansen, 1982) for testing (1) is given by

\[
J_n \equiv 2n Q_n(\hat{\theta}_n).
\]

The corresponding “standard” overidentification test’s null rejection probability is given by \(P_\gamma(J_n > \chi^2_{1-\alpha}(H-p))\) for \(\gamma \in \Gamma_0\), where \(\chi^2_{1-\alpha}(H-p)\) denotes the \(1 - \alpha\) quantile of a \(\chi^2(H-p)\). We approximate the (exact) size of the test, \(\sup_{\gamma \in \Gamma_0} P_\gamma(J_n > \chi^2_{1-\alpha}(H-p))\), by its asymptotic size,

\[
\text{AsySz}_{J} \equiv \limsup_{n \to \infty} \sup_{\gamma \in \Gamma_0} P_\gamma(J_n > \chi^2_{1-\alpha}(H-p)).
\]

Under \(\{ \gamma_n \} \in \Gamma_0(\gamma^*)\) and Assumptions 1–3, the objective function admits the following quadratic expansion (cf. Andrews, 1999, 2002)

\[
Q_n(\theta) = G_n'(\theta_n) W G_n(\theta_n)/2 + G_n'(\theta_n) W G_\theta(\theta - \theta_n) + (\theta - \theta_n)' G_\theta W G_\theta(\theta - \theta_n)/2 + R_n(\theta),
\]

where the remainder, \(R_n(\theta)\), satisfies

\[
\sup_{\theta \in \Theta; \|\sqrt{n}(\theta - \theta_n)\| \leq \epsilon} |R_n(\theta)| = o_p(1/n)
\]

\(^{11}\)See Kitamura (2006) for an analysis of specification tests (in the linear model) under rank deficiency.
for all constants $0 < \epsilon < \infty$.

Example 1 continued: We have

$$R_n(\theta) = G_n'(\theta_n)W_nG_n(\theta_n)/2 - G_n''(\theta_n)WG_n(\theta_n)/2$$
$$+ G_n'(\theta_n)W_n(-D_n)(\theta - \theta_n) - G_n'(\theta_n)WG_\theta(\theta - \theta_n)$$
$$+ (\theta - \theta_n)(-D_n)'W_n(-D_n)(\theta - \theta_n)/2 - (\theta - \theta_n)'G_n''WG_\theta(\theta - \theta_n)/2$$

using $G_n(\theta) = G_n(\theta_n) + (-D_n)(\theta - \theta_n)$, where $D_n = \frac{1}{n} \sum_{i=1}^{n} z_i x_i'$.

Letting $I = G_n'WG_\theta$, $Z_n = I^{-1}G_n'W\sqrt{n}G_n(\theta_n)$, and $q_n(\lambda) = (\lambda + Z_n)'I(\lambda + Z_n)$, equation (3) can be rewritten as

$$Q_n(\theta) = G_n'(\theta_n)WG_n(\theta_n)/2 - \frac{1}{2n} Z_n'IZ_n + \frac{1}{2n} q_n(\sqrt{n}(\theta - \theta_n)) + R_n(\theta).$$

Since $\sqrt{n}(\hat{\theta}_n - \theta_n) = O_p(1)$ under $\{\gamma_n\} \in \Gamma_0(\gamma^*)$ and Assumptions 1–3, the asymptotic null distribution of $J_n$ is governed by

$$G_n'(\theta_n)WG_n(\theta_n)/2 - \frac{1}{2n} Z_n'IZ_n + \frac{1}{2n} q_n(\sqrt{n}(\theta - \theta_n)).$$

(5)

It follows by standard arguments that the sum of the first two summands (times $2n$) is asymptotically $\chi^2(H - p)$ distributed under $\{\gamma_n\} \in \Gamma_0(\gamma^*)$ and Assumptions 1–4. Furthermore, the sum of the first two summands (times $2n$) is asymptotically independent of $q_n(\sqrt{n}(\hat{\theta}_n - \theta_n))$, such that the standard overidentification test will not control asymptotic size if $q_n(\sqrt{n}(\hat{\theta}_n - \theta_n))$, which is nonnegative, fails to be $o_p(1)$ for some $\{\gamma_n\} \in \Gamma_0(\gamma^*)$ (and some $\gamma^* \in \Gamma_0$). A sufficient condition (for failure of $q_n(\sqrt{n}(\hat{\theta}_n - \theta_n))$ to be $o_p(1)$) is given by Assumption 5, which relies on the following definition dating back to Chernoff (1954): A set $\Lambda \subset \mathbb{R}^p$ is approximated by a cone $\Lambda \subset \mathbb{R}^p$ if\(^{12}\)

$$\inf_{\lambda \in \Lambda} \|\lambda - a\| = o(\|a\|) \text{ for } a \in A \text{ and } \inf_{a \in A} \|\lambda - a\| = o(\|\lambda\|) \text{ for } \lambda \in \Lambda.$$

Assumption 5. $\exists \gamma_0 = (\theta_0, \phi_0) \in \Gamma_0$ such that $\bar{\Theta} - \theta_0$ is approximated by a cone $\Lambda$ satisfying $\text{cl}(\Lambda) \neq \mathbb{R}^p$.

Example 1 continued: Take, for example, $\gamma_0 = (\theta_0, \phi_0)$ with $\theta_0 = 0_p$ and $\phi_0$ equal to an arbitrary element in $\Phi_0$. Then, Assumption 5 is satisfied with $\Lambda = [0, \infty) \times (-\infty, \infty)^{p-1}$.

Assumption 5 implies that $\theta_0$ is at the boundary of the optimization parameter space. Given Assumptions 1–5, the results in Andrews (1999, 2002) imply that, if the true

\(^{12}\Lambda \subset \mathbb{R} \text{ is a cone if } \lambda \in \Lambda \text{ implies } c\lambda \in \Lambda \text{ for } c \in (0, \infty)$.
parameter vector is equal to $\theta_0$, the (scaled and centered) estimator (of $\theta_0$) is not asymptotically normal. Another consequence is that the standard overidentification test does not control asymptotic size, which constitutes the first result of this paper.

**Proposition 1.** Under Assumptions 1–5, $\text{AsySz}_J > \alpha$.

One possibility to construct a test that controls asymptotic size is to adapt the critical value, using, for example, the least favorable configuration approach as in Andrews and Cheng (2012) or Bonferroni-based approaches as in McCloskey (2017). Here, we instead suggest the use of a modified J-statistic which has two advantages. First, the resulting test is easy to implement and behaves like the standard test in standard settings. Second, the resulting test is asymptotically similar, which contrasts with the alternative approaches mentioned above of which some may even yield conservative tests, i.e., tests that have asymptotic size strictly less than $\alpha$. The proposed test (statistic) is motivated by equation (5). If the first two summands (times $2n$) were observed, we could readily test (1) by comparing them, i.e., their sum, to $\chi^2_{1-\alpha}(H-p)$. Since they are, however, unobserved, we propose to use the sample analogue, i.e.,

$$J_n^M \equiv nG_n'((\hat{\theta}_n))W_nG_n((\hat{\theta}_n)) - nG_n'((\hat{\theta}_n))W_nG_{\theta,n}(\hat{\theta}_n,n)\left(G_{\theta,n}W_nG_{\theta,n}^{-1}\right)^{-1}G_{\theta,n}W_nG_n((\hat{\theta}_n),$$

where (with a slight abuse of notation) $W_n \equiv W_n((\hat{\theta}_n))$ and where $\hat{G}_{\theta,n}$ denotes a consistent estimator of $G_{\theta}$, which by Assumptions 1–3 exists. The following Proposition shows that, like its unobserved counterpart, $J_n^M$ is asymptotically $\chi^2(H-p)$ distributed under $\{\gamma_n\} \in \Gamma_0(\gamma^*)$ and Assumptions 1–4.

**Proposition 2.** Under $\{\gamma_n\} \in \Gamma_0(\gamma^*)$ and Assumptions 1–4, $J_n^M \to \chi^2(H-p)$.

Based on Proposition 2, we suggest to test the overidentifying restrictions by comparing $J_n^M$ to $\chi^2_{1-\alpha}(H-p)$, i.e., to reject $H_0$ if $J_n^M > \chi^2_{1-\alpha}(H-p)$. In practice, researchers may prefer to use the following test statistic

$$J_n^{M*} \equiv nG_n'((\hat{\theta}_n))W_nG_n((\hat{\theta}_n)) - 1((\hat{\theta}_n \in \text{bd}(\Theta)) \times nG_n'((\hat{\theta}_n))W_n\hat{G}_{\theta,n}(\hat{\theta}_n)\left(\hat{G}_{\theta,n}W_n\hat{G}_{\theta,n}^{-1}\right)^{-1}\hat{G}_{\theta,n}W_nG_n((\hat{\theta}_n),$$

which is asymptotically equivalent to $J_n^M$ (under $\{\gamma_n\} \in \Gamma_0(\gamma^*)$ and Assumptions 1–3). $J_n^{M*}$ has the advantage of reducing to $J_n$ when the estimator takes on a value in the interior of the optimization parameter space, i.e., $\hat{\theta}_n \in \text{int}(\Theta)$, such that, in those cases, researchers are, for all practical purposes, implementing the standard overidentification

---

13Note that, as alluded to in the introduction, $J_n^M$ corresponds to the sample analogue of the quadratic part of equation (3), i.e., $Q_n(\theta) - R_n(\theta)$, (times $2n$) evaluated at its minimizer (over $\mathbb{R}^p$).

14If $g(\theta, w)$ has continuous $(1/r)$ partial derivatives on $\Theta$, say $\frac{\partial^r g(\theta, w_i)}{\partial \theta^r}$, one may simply take $\hat{G}_{\theta,n} = \frac{1}{n} \sum_{i=1}^n \frac{\partial^r g(\hat{\theta}_n, w_i)}{\partial \theta^r}$. Otherwise, one can use a numerical approximation, see e.g., the estimator on p. 1043 in Pakes and Pollard (1989), cf. Lemma 1 in Andrews (2002).
test, avoiding the computation of $\hat{G}_{\theta,n}$. Similarly, note that if $Q_n(\theta)$ is defined on some $\tilde{\Theta}$ such that $\Theta \subset \text{int}(\tilde{\Theta})$, then another asymptotically equivalent test statistic (under $\{\gamma_n\} \in \Gamma_0(\gamma^*)$ and Assumptions 1–3) is given by $J_{n}^{M^{**}} \equiv 2nQ_n(\hat{\theta}_n)$, where $\hat{\theta}_n$ satisfies $Q_n(\hat{\theta}_n) = \inf_{\theta \in \tilde{\Theta}} Q_n(\theta) + o_p(1)$.

Example 1 continued: It is easy to show that $J_{n}^{M} = J_{n}^{M^*} = J_{n}^{M^{**}}$ (with $\hat{G}_{\theta,n} = -D_n$), i.e., the proposed test statistic is equal to the standard J-statistic that ignores the restrictions on the parameter space (except that the latter uses an unrestricted first-stage estimator).

More generally, the asymptotic equivalence illustrates that using $J_{n}^{M}$ amounts to locally ignoring the restrictions on the parameter space when testing overidentifying restrictions, or model validity. While the restrictions on the parameter space contain model relevant “information” that could possibly be exploited (cf. footnote 8), in particular, to partially remedy the lack of nontrivial (local asymptotic) power under certain (local) alternatives of which the proposed test, like the standard test in standard settings (see e.g., Newey, 1985; Guggenberger, 2012; Parente and Santos Silva, 2012), should suffer, the proposed test should prove useful given its ease of implementation. Relatedly, due to its resemblance to the standard J-statistic in standard settings, the proposed test statistic (based on different objective functions) can also, in a straight-forward manner, be used to test the validity of a prespecified subset of the moment conditions (cf. Newey, 1985; Eichenbaum et al., 1988), under the maintained assumption that $p_1 (p \leq p_1 < H)$ moment conditions are known to be valid.\footnote{Relatedly, note that $J_{n}^{M}$ also reduces to $J_{n}$ when $\hat{\theta}_n \in \text{int}(\tilde{\Theta})$, $W_n$ does not depend on $\theta$, and $g(\theta, w)$ has continuous (l/r) partial derivatives (and $\hat{G}_{\theta,n} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} g(\hat{\theta}_n, w_i)$).}

Let $\text{AsySz}_{J_{n}^{M}} \equiv \limsup_{n \to \infty} \sup_{\gamma \in \Gamma_0} P(\gamma) (J_{n}^{M} > \chi^2_{1-\alpha}(H - p))$ and let $\text{AsyMinRP}_{J_{n}^{M}} \equiv \liminf_{n \to \infty} \inf_{\gamma \in \Gamma_0} P(\gamma) (J_{n}^{M} > \chi^2_{1-\alpha}(H - p))$, where $\text{AsyMinRP}$ is short for “minimum asymptotic null rejection probability.” Then, we have the following Corollary.

Corollary 1. Under Assumptions 1–4, $\text{AsySz}_{J_{n}^{M}} = \text{AsyMinRP}_{J_{n}^{M}} = \alpha$.

Corollary 1 shows that the proposed test, based on $J_{n}^{M}$, is asymptotically similar and, in particular, controls asymptotic size. Note that Corollary 1 does not impose any restrictions on $\Theta$ (beyond compactness, which is only imposed to establish consistency of the estimator). In particular, Corollary 1 also applies if, in addition, Assumption 5 holds.\footnote{See Eichenbaum et al. (1988) for an empirical application where interest lies with testing the validity of a prespecified subset of the moment conditions.}
4 Conclusion

We conclude with a message of caution. As mentioned in Section 3, the proposed test should, like the standard test in standard settings, only have trivial (local asymptotic) power under certain (local) alternatives. The results in Guggenberger and Kumar (2012), then, suggest that if the test is used as a pretest in a two-step procedure, where in a second step inference on the parameter(s) of interest is conducted using, for example, the test proposed in Ketz (2018) (if the pretest does not reject), this two-step procedure will be severely size distorted—a detailed analysis is left for future research. Therefore, the warnings and recommendations in Guggenberger and Kumar (2012) also apply here.

Acknowledgements

I thank a referee for insightful comments and, in particular, for pointing out a problem with one of the proofs in a previous draft. I also thank Xavier D’Haultfoeuille, Frank Kleibergen, Adam McCloskey, and Eric Renault for helpful discussions and/or comments.

References


