Queen's Economics Department Working Paper No. 1257

Confidence Sets Based on Inverting Anderson-Rubin Tests

Russell Davidson
McGill University

James G. MacKinnon
Queen's University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

1-2011

Confidence Sets Based on Inverting Anderson-Rubin Tests

by

Russell Davidson

Department of Economics
McGill University
Montreal, Quebec, Canada
H3A 2T7

Russell.Davidson@mcgill.ca

and

James G. MacKinnon

Department of Economics
Queen's University
Kingston, Ontario, Canada
K7L 3N6

jgm@econ.queensu.ca

Abstract

Economists are often interested in the coefficient of a single endogenous explanatory variable in a linear simultaneous equations model. One way to obtain a confidence set for this coefficient is to invert the Anderson-Rubin test. The "AR confidence sets" that result have correct coverage under classical assumptions. In this paper, however, we show that AR confidence sets also have many undesirable properties. Their coverage conditional on quantities that the investigator can observe, notably the Sargan statistic, can be far from correct. It is well known that they can be unbounded when the instruments are weak. Even when they are bounded, their length may be very misleading. We argue that, at least when the instruments are not so weak that inference is hopeless, it is much better to obtain confidence intervals by bootstrapping either the IV or LIML t statistic on the coefficient of interest in a particular way that we propose.

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada, the Canada Research Chairs program (Chair in Economics, McGill University), and the Fonds Québécois de Recherche sur la Société et la Culture.

January 4, 2011.

1. Introduction

Classical confidence intervals are, at least implicitly, defined by "inverting" a test. A confidence set at level $1 - \alpha$, which may or may not be a single bounded interval, is simply the set of parameter values for which a test at level α does not reject the null hypothesis. This seems to imply that inverting an exact test must lead to a confidence set that has good properties. However, as we show in this paper, that is not necessarily so.

In the linear simultaneous-equations model with weak instruments, the asymptotic distributions of t statistics often provide poor guides to their finite-sample distributions; see Staiger and Stock (1997). As a consequence, confidence intervals based on inverting t tests often have very poor coverage. One proposed solution to this problem is to invert tests which have better finite-sample properties. Several papers, including Dufour (1997), Zivot, Startz, and Nelson (1998), and Dufour and Taamouti (2005), therefore suggest inverting the test of Anderson and Rubin (1949), which is exact under classical assumptions. We shall refer to the resulting confidence set as an "AR confidence set."

In this paper, we argue that, although AR confidence sets have correct unconditional coverage, at least under classical assumptions, they have many undesirable properties. Although some of these properties have previously been studied, notably by Zivot, Startz, and Nelson (1998), we offer some new theoretical results together with supporting simulation evidence. AR confidence sets do not have correct coverage conditional on the type of confidence set that actually occurs. Moreover, when they are bounded, their length depends on the value of the Sargan statistic for the validity of the overidentifying restrictions. Therefore, any AR confidence set that is actually observed does not have correct coverage. It can be empty, misleadingly short, misleadingly long, or unbounded.

Having correct coverage unconditionally, while desirable, is by itself not very useful. One can always create a $(1 - \alpha)\%$ confidence set with the correct unconditional coverage by setting it equal to the empty set with probability α and the real line with probability $1 - \alpha$. But such a straw-man confidence set provides no useful information. Unfortunately, when the instruments are weak, the AR confidence set may not be much more informative than this straw-man one. Even when they are strong, it never has the correct conditional coverage.

Forchini and Hillier (2003) have argued that the AR statistic is not in fact pivotal, because it does not depend on the parameter of interest everywhere in the parameter space, and that confidence sets based on it are therefore invalid. Our paper is concerned with the more detailed properties of AR confidence sets, but some of the issues that arise below are related to this important point.

There are actually two different problems with AR confidence sets. The first problem is that they may be unbounded. This problem can arise whenever the instruments in a linear simultaneous-equations model are weak, and it can also affect confidence sets based on inverting other tests. See Dufour (1997) and Zivot, Startz, and Nelson

(1998). The second problem is that they may be empty or extremely short. This problem can arise whenever we invert a test that has more than one degree of freedom. In the Appendix, we show that it can occur when we invert an F test in the classical normal linear regression model.

In the next section, we introduce Anderson-Rubin confidence sets and show that there are four types of them. In Section 3, we explore the important relationship between AR confidence sets and the Sargan statistic for overidentification. Construction of an AR confidence set is similar to inverting an F test, and this is discussed in the Appendix. In Section 4, we use simulation experiments to study the properties of AR confidence sets. In Section 5, we describe a procedure that can be used to obtain confidence sets by inverting bootstrap t tests. Simulation evidence shows that it works very well, provided the instruments are not so weak that reliable inference is basically impossible. We argue that confidence sets based on this bootstrap procedure are clearly superior to AR confidence sets. Although they may be unbounded, they cannot be empty or extremely short.

2. Anderson-Rubin Confidence Sets

We deal with the two-equation linear model

$$\boldsymbol{y}_1 = \beta \boldsymbol{y}_2 + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}_1 \tag{1}$$

$$\boldsymbol{y}_2 = \boldsymbol{W}\boldsymbol{\pi} + \boldsymbol{u}_2 = \boldsymbol{Z}\boldsymbol{\pi}_1 + \boldsymbol{W}_2\boldsymbol{\pi}_2 + \boldsymbol{u}_2. \tag{2}$$

Here $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are n–vectors of observations on endogenous variables, $\boldsymbol{Z}$ is an $n \times k$ matrix of observations on exogenous variables, and $\boldsymbol{W}$ is an $n \times l$ matrix of exogenous instruments with the property that $\mathcal{S}(\boldsymbol{Z})$, the subspace spanned by the columns of $\boldsymbol{Z}$, lies in $\mathcal{S}(\boldsymbol{W})$, the subspace spanned by the columns of $\boldsymbol{W}$. The $n \times (l-k)$ matrix $\boldsymbol{W}_2$ is constructed in such a way that $\mathcal{S}(\boldsymbol{Z}, \boldsymbol{W}_2) = \mathcal{S}(\boldsymbol{W})$. Equation (1) is a structural equation, and equation (2) is a reduced-form equation.

The disturbance vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are assumed to be serially uncorrelated and homoskedastic, with mean zero and contemporaneous covariance matrix

$$\boldsymbol{\Sigma} \equiv \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

For the AR test to be exact, we also need the disturbances to be normally distributed. We assume that the model is overidentified, which implies that $l > k+1$. The number of overidentifying restrictions is $l - k - 1$.

The Anderson-Rubin statistic for a test of the hypothesis that $\beta = \beta_0$ is

$$\mathrm{AR}(\beta_0) = \frac{n-l}{l-k} \frac{(\boldsymbol{y}_1 - \beta_0\boldsymbol{y}_2)^\top \boldsymbol{P}_1 (\boldsymbol{y}_1 - \beta_0\boldsymbol{y}_2)}{(\boldsymbol{y}_1 - \beta_0\boldsymbol{y}_2)^\top \boldsymbol{M}_{\boldsymbol{W}} (\boldsymbol{y}_1 - \beta_0\boldsymbol{y}_2)}, \tag{3}$$

where $\boldsymbol{M}_{\boldsymbol{W}} \equiv \mathbf{I} - \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top = \mathbf{I} - \boldsymbol{P}_{\boldsymbol{W}}$, $\boldsymbol{M}_{\boldsymbol{Z}} \equiv \mathbf{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top = \mathbf{I} - \boldsymbol{P}_{\boldsymbol{Z}}$, and $\boldsymbol{P}_1 \equiv \boldsymbol{M}_{\boldsymbol{Z}} - \boldsymbol{M}_{\boldsymbol{W}} = \boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{Z}}$. Under the null hypothesis, the AR statistic (3)

is distributed as $F(l-k, n-l)$. This statistic is, of course, minimized at the LIML estimator $\hat{\beta}_{\text{LIML}}$.

Let q be the $1-\alpha$ quantile of the $F(l-k, n-l)$ distribution. Then β_0 belongs to the confidence set at level $1-\alpha$ if and only if $\text{AR}(\beta_0) \leq q$. This inequality can be reformulated as

$$(\boldsymbol{y}_2^\top \boldsymbol{A} \boldsymbol{y}_2)\beta_0^2 - 2(\boldsymbol{y}_1^\top \boldsymbol{A} \boldsymbol{y}_2)\beta_0 + \boldsymbol{y}_1^\top \boldsymbol{A} \boldsymbol{y}_1 \geq 0, \tag{4}$$

where $\boldsymbol{A} = c\boldsymbol{M}_{\boldsymbol{W}} - \boldsymbol{P}_1$, with $c = q(l-k)/(n-l)$. Zivot, Startz, and Nelson (1998) study this inequality in some detail and obtain the result that the AR confidence set is unbounded whenever the F statistic for $\boldsymbol{\pi}_2 = \boldsymbol{0}$ in (2) is less than q. It is worth going through the argument that leads to this important result, because it also shows that there are four types of AR confidence set and explains the circumstances in which they occur.

The discriminant of the quadratic equation obtained by replacing the inequality in (4) by an equality is

$$D \equiv 4(\boldsymbol{y}_1^\top \boldsymbol{A} \boldsymbol{y}_2)^2 - 4\boldsymbol{y}_1^\top \boldsymbol{A} \boldsymbol{y}_1 \, \boldsymbol{y}_2^\top \boldsymbol{A} \boldsymbol{y}_2. \tag{5}$$

If $D < 0$, the equation has no real roots, so that the inequality (4) is either always or never satisfied. It is always satisfied if the coefficient of β_0^2 is positive, since the left-hand side tends to $+\infty$ as $|\beta_0| \to \infty$. In this case, the confidence set is the entire real line. However, it is never satisfied if $\boldsymbol{y}_2^\top \boldsymbol{A} \boldsymbol{y}_2 < 0$, which implies that the confidence set is empty.

If $D > 0$, the equation has two real roots. If $\boldsymbol{y}_2^\top \boldsymbol{A} \boldsymbol{y}_2 < 0$, the quadratic function of β_0 on the left-hand side of (4) tends to $-\infty$ as $\beta_0 \to \infty$. It has a single maximum. The inequality (4) is therefore satisfied between these roots, so that the interval between them is the confidence set. If $\boldsymbol{y}_2^\top \boldsymbol{A} \boldsymbol{y}_2 > 0$, the quadratic has a single minimum, and (4) is satisfied in the set composed of the disjoint union of the open infinite interval from the upper root to $+\infty$ and that from the lower root to $-\infty$.

Whether $D < 0$ or $D > 0$, the confidence set is unbounded whenever $\boldsymbol{y}_2^\top \boldsymbol{A} \boldsymbol{y}_2 > 0$. This condition can be rewritten as

$$c\, \boldsymbol{y}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2 - \boldsymbol{y}_2 \boldsymbol{P}_1 \boldsymbol{y}_2 > 0.$$

Using the definition of c and a little algebra allows us to rewrite this inequality as

$$\frac{\boldsymbol{y}_2 \boldsymbol{P}_1 \boldsymbol{y}_2 / (l-k)}{\boldsymbol{y}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2 / (n-l)} < q. \tag{6}$$

The quantity on the left-hand side of (6) is the ordinary F statistic for $\boldsymbol{\pi}_2 = \boldsymbol{0}$ in equation (2), and q is the critical value for a test at level α based on this statistic, which tests the null hypothesis that the structural equation (1) is not identified. Thus, as Zivot, Startz, and Nelson (1998) showed, the AR confidence set is unbounded (with or without a hole in the middle) whenever we cannot reject the hypothesis that the

instruments that are not also explanatory variables (namely, the columns of $\boldsymbol{W}_2$) have no explanatory power for $\boldsymbol{y}_2$.

There is no point calculating an AR confidence set whenever the inequality (6) holds, because a set that consists of the entire real line, perhaps with a hole in the middle, tells us nothing useful about the value of β. In contrast to the confidence set, the identifiability test statistic does provide valuable information, since it provides a natural measure of the strength of the instruments; see Stock and Yogo (2005).

We have seen that there are four types of AR confidence set. The set is a bounded interval when $D > 0$ and the test statistic on the left-hand side of (6) is significant. It is empty when $D < 0$ and this identifiability test statistic is significant. It is the entire real line when $D < 0$ and the test statistic is insignificant, and it is the disjoint union of two open intervals when $D > 0$ and the test statistic is insignificant. The fact that some types of AR confidence set are unbounded when the instruments are sufficiently weak can be viewed as a consequence of a fundamental result of Dufour (1997), who showed that no valid confidence set which is almost surely bounded exists in the neighborhood of a point where the parameter is not identified.

Figure 1 illustrates all four types of interval by graphing $\mathrm{AR}(\beta_0)$ against β_0. The dashed horizontal line is the critical value, q. Two variants of the bounded interval case are shown. In one of these, the interval is very short, and in the other it is quite long. What type of interval we obtain depends on α. In particular, the probability that the interval is an empty set diminishes as α becomes smaller and q consequently becomes larger. Note that all five intervals in the figure are for samples drawn from the same data-generating process, for which the instruments are moderately weak.

Unconditionally, the AR confidence set always has the correct coverage. However, once we observe what type of set it is, that is no longer the case. By construction, the empty set undercovers, and the real line overcovers. The bounded interval and the disjoint interval can either overcover or undercover. As the figure illustrates, the bounded interval can be very much too short. Thus we cannot interpret an observed AR confidence set, even a bounded interval, in the way we would like to interpret a confidence interval. On average, at least when the model is well identified, bounded intervals must overcover, in order to offset the failure of empty sets to cover at all. But there will always be bounded intervals like the one shown in the top panel of Figure 1 which give the misleading impression that we have estimated β much more accurately than is actually the case.

3. Relations with the Sargan Test

The Sargan statistic for overidentifying restrictions (Sargan, 1958) is most commonly computed as $1/\hat{\sigma}_1^2$ times the minimized value of the IV criterion function, that is,

$$\frac{1}{\hat{\sigma}_1^2}(\boldsymbol{y}_1 - \hat{\beta}_{\mathrm{IV}}\boldsymbol{y}_2)^\top \boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{y}_1 - \hat{\beta}_{\mathrm{IV}}\boldsymbol{y}_2) = \frac{1}{\hat{\sigma}_1^2}(\boldsymbol{y}_1 - \hat{\beta}_{\mathrm{IV}}\boldsymbol{y}_2)^\top \boldsymbol{P}_1(\boldsymbol{y}_1 - \hat{\beta}_{\mathrm{IV}}\boldsymbol{y}_2), \qquad (7)$$

where $\hat{\beta}_{\mathrm{IV}}$ is the IV (or two-stage least squares) estimate of β, and the estimated variance $\hat{\sigma}_1^2$ denotes $n^{-1}\hat{\boldsymbol{u}}_1^\top \boldsymbol{M}_{\boldsymbol{Z}}\hat{\boldsymbol{u}}_1$, with $\hat{\boldsymbol{u}}_1 \equiv \boldsymbol{y}_1 - \hat{\beta}_{\mathrm{IV}}\boldsymbol{y}_2$. The equality in (7) follows from the fact that

$$(\boldsymbol{M}_{\boldsymbol{Z}} - \boldsymbol{M}_{\boldsymbol{W}})(\boldsymbol{y}_1 - \hat{\beta}_{\mathrm{IV}}\boldsymbol{y}_2) = (\mathbf{I} - \boldsymbol{M}_{\boldsymbol{W}})(\boldsymbol{y}_1 - \hat{\beta}_{\mathrm{IV}}\boldsymbol{y}_2) = \boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{y}_1 - \hat{\beta}_{\mathrm{IV}}\boldsymbol{y}_2),$$

because $\boldsymbol{Z}$ must be orthogonal to the IV residuals.

It is evident that the numerator of the expression on the right-hand side of equation (7) would be identical to the numerator of the AR statistic (3) if $\hat{\beta}_{\mathrm{IV}}$ were replaced by β_0. The latter will always be larger than the former, because $\hat{\beta}_{\mathrm{IV}}$ minimizes the numerator. That is why the AR statistic has $l - k$ degrees of freedom in the numerator, while the Sargan statistic (which, of course, is not exact) has $l - k - 1$. It is not hard to show that the numerator of (3) can be rewritten as

$$(\boldsymbol{y}_1 - \hat{\beta}_{\mathrm{IV}}\boldsymbol{y}_2)^\top \boldsymbol{P}_1(\boldsymbol{y}_1 - \hat{\beta}_{\mathrm{IV}}\boldsymbol{y}_2) + (\hat{\beta}_{\mathrm{IV}}\boldsymbol{y}_2 - \beta_0\boldsymbol{y}_2)^\top \boldsymbol{P}_1(\hat{\beta}_{\mathrm{IV}}\boldsymbol{y}_2 - \beta_0\boldsymbol{y}_2). \tag{8}$$

The first term in (8) is the numerator of the Sargan statistic (7). Thus, if the Sargan and AR statistics had the same denominator, the latter would always be larger than the former. This is not always true in finite samples, because the denominators are not the same, although they both estimate σ_1^2 consistently under the null. But there is inevitably a very strong tendency for large values of the Sargan statistic to be associated with large values of the AR statistic.

In order to analyze the statistical properties of the AR confidence set and its relationship to the Sargan statistic, we need to specify a data-generating process. Following Davidson and MacKinnon (2008), we use the DGP:

$$\begin{aligned} \boldsymbol{y}_1 &= \beta\boldsymbol{y}_2 + \boldsymbol{u}_1, \\ \boldsymbol{y}_2 &= a\boldsymbol{w} + \boldsymbol{u}_2, \end{aligned} \tag{9}$$

where $\boldsymbol{w} \in \mathcal{S}(\boldsymbol{W})$ is an n–vector with $\|\boldsymbol{w}\|^2 = 1$, and

$$\begin{aligned} \boldsymbol{u}_1 &= r\boldsymbol{v}_1 + \rho\boldsymbol{v}_2, \\ \boldsymbol{u}_2 &= \boldsymbol{v}_2, \end{aligned} \qquad \begin{bmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \end{bmatrix} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}), \quad r^2 + \rho^2 = 1. \tag{10}$$

The fact that there is just a single instrument $\boldsymbol{w}$ in the DGP is entirely consistent with there being l of them in equation (2). What matters is the total explanatory power of all the instruments for $\boldsymbol{y}_2$. According to (9), all of this explanatory power comes from the vector $\boldsymbol{w}$, and the other columns of $\boldsymbol{W}$ are simply noise. Since it is only $\mathcal{S}(\boldsymbol{W})$ that matters, we are perfectly free to perform a linear transformation on $\boldsymbol{W}$ that makes this the case.

The instrument vector $\boldsymbol{w}$ is normalized to have squared length unity. By employing this normalization, we are implicitly using weak-instrument asymptotics; see Staiger and Stock (1997). The strength of the instruments is measured by the parameter a.

The square of this parameter is the so-called scalar concentration parameter; see Phillips (1983, p. 470) and Stock, Wright, and Yogo (2002). For simplicity, all variances have also been normalized to unity.

As we have seen, the AR confidence set is a bounded interval if and only if $D > 0$ and $\boldsymbol{y}_2^\top \boldsymbol{A} \boldsymbol{y}_2 < 0$. In this case, the length of the interval is the distance between the two roots of the quadratic equation (4), which is $-\sqrt{D}/\boldsymbol{y}_2^\top \boldsymbol{A} \boldsymbol{y}_2$. Under the DGP (9), the limit of this ratio as $a \to \infty$ is zero. The quantity that has a non-trivial limit as $a \to \infty$ is thus the length of the interval times a. It can be shown that this limit is the square root of

$$c(\boldsymbol{y}_1 - \beta \boldsymbol{y}_2)^\top \boldsymbol{M}_{\boldsymbol{W}} (\boldsymbol{y}_1 - \beta \boldsymbol{y}_2) - (\boldsymbol{y}_1 - \beta \boldsymbol{y}_2)^\top (\boldsymbol{P}_1 - \boldsymbol{P}_{\boldsymbol{w}})(\boldsymbol{y}_1 - \beta \boldsymbol{y}_2), \tag{11}$$

where $\boldsymbol{P}_{\boldsymbol{w}} \equiv \boldsymbol{w}^\top (\boldsymbol{w}^\top \boldsymbol{w})^{-1} \boldsymbol{w}^\top$. The first term in (11) is c times a random variable that follows the $\chi^2(n-l)$ distribution. The second term is an independent random variable that follows the $\chi^2(l-k-1)$ distribution. Of course, both of these quantities would have to be multiplied by σ_1^2 if we had not set it to unity. The distribution of the second term, and its independence from the first term, both follow from the fact that the matrix $\boldsymbol{P}_1 - \boldsymbol{P}_{\boldsymbol{w}} = \boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_{\boldsymbol{w}}$ projects onto the $l - k - 1$ components of $\boldsymbol{W}$ that do not lie in $\mathcal{S}(\boldsymbol{w}, \boldsymbol{Z})$.

Expression (11) is random and may be either positive or negative. It is most likely to be negative when α is large, so that q, the $1 - \alpha$ quantile of $F(l-k, n-l)$, is small. There is evidently a non-empty confidence set only when it is positive. Since we are considering the limit as $a \to \infty$, there is no danger that the set will be unbounded.

It is interesting to see how expression (11) is related to a slightly modified version of the Sargan statistic (7). The modified statistic is

$$S = \frac{\hat{\boldsymbol{u}}^\top \boldsymbol{P}_1 \hat{\boldsymbol{u}}}{\check{\sigma}_1^2}, \quad \check{\sigma}_1^2 = \frac{1}{n-l} \hat{\boldsymbol{u}}^\top \boldsymbol{M}_{\boldsymbol{W}} \hat{\boldsymbol{u}}, \tag{12}$$

where $\hat{\boldsymbol{u}} = \boldsymbol{y}_1 - \hat{\beta}_{\mathrm{IV}} \boldsymbol{y}_2$. This differs from (7) because, instead of using the usual variance estimate $\hat{\sigma}_1^2$, it uses the same one as the AR statistic for testing $\beta = \hat{\beta}_{\mathrm{IV}}$.

The numerator of S is

$$\hat{\boldsymbol{u}}^\top \boldsymbol{P}_1 \hat{\boldsymbol{u}} = (\boldsymbol{y}_1 - \hat{\beta}_{\mathrm{IV}} \boldsymbol{y}_2)^\top \boldsymbol{P}_1 (\boldsymbol{y}_1 - \hat{\beta}_{\mathrm{IV}} \boldsymbol{y}_2) = \boldsymbol{y}_1^\top \boldsymbol{P}_1 (\boldsymbol{y}_1 - \hat{\beta}_{\mathrm{IV}} \boldsymbol{y}_2), \tag{13}$$

where the second equality follows from the moment condition $\boldsymbol{y}_2^\top \boldsymbol{P}_1 (\boldsymbol{y}_1 - \hat{\beta}_{\mathrm{IV}} \boldsymbol{y}_2) = 0$ that defines $\hat{\beta}_{\mathrm{IV}}$. This moment condition implies that

$$\hat{\beta}_{\mathrm{IV}} = \frac{\boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_1}{\boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2}. \tag{14}$$

Substituting (14) into the rightmost expression in (13) yields

$$\begin{aligned} \hat{\boldsymbol{u}}^\top \boldsymbol{P}_1 \hat{\boldsymbol{u}} &= \boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_1 - \boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_2 (\boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2)^{-1} \boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_1 \\ &= \boldsymbol{y}_1^\top (\boldsymbol{P}_1 - \boldsymbol{P}_{\boldsymbol{P}_1 \boldsymbol{y}_2}) \boldsymbol{y}_1 = \boldsymbol{u}_1^\top (\boldsymbol{P}_1 - \boldsymbol{P}_{\boldsymbol{P}_1 \boldsymbol{y}_2}) \boldsymbol{u}_1, \end{aligned}$$

where $\boldsymbol{P}_{\boldsymbol{P}_1\boldsymbol{y}_2}$ projects orthogonally on to $\mathcal{S}(\boldsymbol{P}_1\boldsymbol{y}_2)$. Thus from (12) we have

$$\check{\sigma}_1^2 S = \boldsymbol{u}_1^\top(\boldsymbol{P}_1 - \boldsymbol{P}_{\boldsymbol{P}_1\boldsymbol{y}_2})\boldsymbol{u}_1 = \boldsymbol{u}_1^\top\boldsymbol{P}_1\boldsymbol{u}_1 - \frac{(\boldsymbol{u}_1^\top\boldsymbol{P}_1\boldsymbol{y}_2)^2}{\boldsymbol{y}_2^\top\boldsymbol{P}_1\boldsymbol{y}_2}. \tag{15}$$

If we replace $\boldsymbol{y}_2$ by $a\boldsymbol{w} + \boldsymbol{u}_2$ and retain only the leading-order terms as $a \to \infty$, the term that is subtracted in the rightmost expression here tends to $(\boldsymbol{w}^\top\boldsymbol{u}_1)^2 = \boldsymbol{u}_1^\top\boldsymbol{P}_{\boldsymbol{w}}\boldsymbol{u}_1$, where the equality follows from the fact that $\boldsymbol{w}^\top\boldsymbol{w} = 1$. Thus, in the limit,

$$\check{\sigma}_1^2 S = \boldsymbol{u}_1^\top(\boldsymbol{P}_1 - \boldsymbol{P}_{\boldsymbol{w}})\boldsymbol{u}_1 = (\boldsymbol{y}_1 - \beta\boldsymbol{y}_2)^\top(\boldsymbol{P}_1 - \boldsymbol{P}_{\boldsymbol{w}})(\boldsymbol{y}_1 - \beta\boldsymbol{y}_2). \tag{16}$$

It is easy to see that $\hat{\beta}_{\mathrm{IV}}$ tends to β as $a \to \infty$. From (14),

$$\hat{\beta}_{\mathrm{IV}} = \frac{(a\boldsymbol{w} + \boldsymbol{u}_2)^\top\boldsymbol{P}_1\big(\beta(a\boldsymbol{w} + \boldsymbol{u}_2) + \boldsymbol{u}_1\big)}{(a\boldsymbol{w} + \boldsymbol{u}_2)^\top\boldsymbol{P}_1(a\boldsymbol{w} + \boldsymbol{u}_2)} = \beta + \frac{(a\boldsymbol{w} + \boldsymbol{u}_2)^\top\boldsymbol{P}_1\boldsymbol{u}_1}{(a\boldsymbol{w} + \boldsymbol{u}_2)^\top\boldsymbol{P}_1(a\boldsymbol{w} + \boldsymbol{u}_2)}.$$

Since the second term in the rightmost expression here is $O(a)/O(a^2) = O(a^{-1})$, that expression vanishes as $a \to \infty$. The consistency of $\hat{\beta}_{\mathrm{IV}}$ implies that

$$\hat{\boldsymbol{u}}^\top\boldsymbol{M}_{\boldsymbol{W}}\hat{\boldsymbol{u}}/(n-l) = (\boldsymbol{y}_1 - \beta\boldsymbol{y}_2)^\top\boldsymbol{M}_{\boldsymbol{W}}(\boldsymbol{y}_1 - \beta\boldsymbol{y}_2)/(n-l) + O(a^{-1})$$

as $a \to \infty$. Thus the first term in (11) can be replaced by

$$q(l-k)\hat{\boldsymbol{u}}^\top\boldsymbol{M}_{\boldsymbol{W}}\hat{\boldsymbol{u}} = q(l-k)\check{\sigma}_1^2.$$

Similarly, by (16), the second term can be replaced by $\check{\sigma}_1^2 S$. We conclude that, in the limit as $a \to \infty$, the length of the bounded AR interval, if it exists, is simply

$$\check{\sigma}_1\big(q(l-k) - S\big)^{1/2}. \tag{17}$$

This is a deterministic function of $\check{\sigma}_1$ and S, which is proportional to the former and nonlinear in the latter. As S increases, the interval becomes shorter and eventually ceases to exist.

Although this result is strictly true only in the limit, it may be expected to provide a good guide whenever a is reasonably large, that is, whenever the instruments are reasonably strong. It implies that, when the AR confidence set is a bounded interval, its coverage will vary inversely with the magnitude of the Sargan statistic. Thus an investigator who obtains a bounded interval and observes that the Sargan statistic is particularly large or small may reasonably infer that the interval is likely to undercover in the former case and overcover in the latter. This may be especially problematic in practice if, as will very often be the case, the overidentifying restrictions are not quite satisfied. In consequence, observed Sargan statistics may well tend to be larger than they should be by chance, and bounded AR intervals consequently shorter.

The fundamental reason for the result that the AR confidence set depends on the value of the Sargan statistic is that the AR statistic has more than one degree of freedom. Something very similar to this result is true whenever we construct a confidence interval by inverting a test with more than one degree of freedom. In the Appendix, as an example, we show what happens when one constructs a

confidence interval by inverting an F test. In that case, there turns out to be no need to consider a limiting argument. This makes it clear that the only reason we needed a limiting argument to obtain (17) is that the Sargan statistic does not have an exact distribution when a is finite.

4. Properties of AR Confidence Sets

In this section, we use simulation experiments to study various properties of AR confidence sets, including their conditional coverage. We generate artificial data from the DGP specified by (9) and (10). Because this DGP uses weak instrument asymptotics, the sample size does not matter much once it exceeds a threshold size. In Davidson and MacKinnon (2010), we found that the performance of various test statistics for β changed very little once n exceeded 400. We therefore set $n = 400$ in all our experiments. For each DGP, we generated 500,000 simulated datasets.

The key parameters in our experiments are a, ρ, and $l - k$. To save space, we report results only for $l - k = 7$, which means that the model is moderately overidentified. Results for substantially smaller or larger values of $l - k$ might look quite different, but that would primarily be because a needs to increase with $l - k$ in order to keep the strength of the instruments constant. The basic structure of the results does not seem to change much with $l - k$.

Figure 2 shows how the frequencies of the four types of 95% AR confidence set depend on a and ρ. The figure has four panels, which correspond to four different values of a. The value of ρ, which varies from 0.00 to 0.99 by increments of 0.01, is on the horizontal axis. Negative values are not included, because the figures would simply be symmetric around $\rho = 0$.

When $a = 1$, the instruments are extremely weak, and when $a = 8$ they are minimally strong. In the former case, the 95% AR confidence set is unbounded about 90% of the time. For most values of ρ, the unbounded set is usually the entire real line. However, as ρ becomes larger, the case of two unbounded segments becomes more common, until it almost completely drives out the real-line case when $\rho = 0.99$. The results for $a = 2$ are similar to those for $a = 1$, except that the bounded interval becomes somewhat more common (but it still occurs less than 25% of the time), and the two unbounded sets become somewhat less common.

The results change dramatically when we move from $a = 2$ to $a = 4$. The 95% AR confidence set is now bounded more than 80% of the time, and the empty set is a good deal more common than it was before. Finally, when $a = 8$, there is just a handful of unbounded confidence sets in 50 million replications, and the bounded interval occurs between 97.1% and 97.4% of the time. The empty set occurs very slightly more often as ρ increases.

Figure 3 shows conditional coverage for four types of confidence set for the same experiments as Figure 2. We do not bother to show coverage for the real line or the empty set. Instead, we show it for bounded intervals when the Sargan statistic, computed in the usual way as (7), either exceeds the 0.90 quantile of the $\chi^2(l-k-1)$

distribution ("S large") or falls short of the 0.50 quantile ("S modest"). Several striking results are apparent from the figure.

- When a is small, the bounded interval may either overcover slightly (when ρ is small) or undercover severely (when ρ is large and $a = 1$). When a is not small, the bounded interval always overcovers, as it must do in order to offset the undercoverage associated with the empty set.
- The two-segment confidence set undercovers when ρ is small. However, as ρ increases, its coverage increases, and it eventually overcovers. This type of confidence set does not occur when $a = 8$.
- The coverage of the bounded interval changes dramatically when we condition on the Sargan statistic. When the latter rejects at the nominal 0.10 level, the bounded interval always undercovers, often severely. In contrast, when it fails to reject at the 0.50 level, the bounded interval always overcovers except for larger values of ρ when $a = 1$. This overcoverage is generally quite extreme. For example, when $a = 8$, the 95% bounded AR interval always covers at least 99.8% of the time when the Sargan statistic fails to reject at the 0.50 level.

These results suggest that the length of a bounded AR confidence interval will generally provide a poor guide to the precision with which the parameter β is estimated. To investigate this conjecture, we calculated the dispersion of $\hat{\beta}_{\text{LIML}}$ as the difference between its 0.025 and 0.975 quantiles over the 500,000 replications. In Figure 4, we compare this with the median and with the 0.01 and 0.99 quantiles of the lengths of the 95% AR confidence sets when they are bounded intervals. Ideally, the median length of the bounded AR intervals should be very similar to the dispersion of the estimates, and the upper-tail and lower-tail quantiles of interval length should not be too much higher or lower than the median.

The results of this exercise for $a = 4$, $a = 8$, and $a = 16$ are shown in the three left-hand panels of Figure 4. We do not present results for smaller values of a because most of the AR confidence sets were unbounded (see Figure 2) and because it is unreasonable to expect any method to produce reliable results in these cases. Note that the vertical axis is logarithmic.

It is evident that the median length of the bounded 95% AR interval is generally a poor guide to the dispersion of $\hat{\beta}_{\text{LIML}}$. The former always overestimates the latter, and the problem does not go away as a becomes larger. Moreover, the length of the bounded AR intervals evidently varies greatly. When $a = 4$, the upper-tail quantile of the distribution of their lengths can be more than 80 times the dispersion of $\hat{\beta}_{\text{LIML}}$, while the lower-tail quantile can be no more than 1/4 of the dispersion. Of course, as the theory of Section 3 makes clear, there are a few bounded intervals that are just barely longer than zero, but these are evidently well to the left of the 0.01 quantile. For large a, this must occur whenever $q(l - k) - S$ in equation (17) is just barely positive.

For comparison, the left-hand panels of Figure 4 also show the median and the 0.01 and 0.99 quantiles of the lengths of the 95% Wald LIML intervals. By the latter,

we simply mean the usual confidence interval obtained by inverting a t test, which is equal to $\hat{\beta}_{\text{LIML}}$ plus or minus 1.96 standard errors. For comparability, these are only plotted for replications where there was a bounded AR interval. The Wald LIML intervals tend to be much shorter than the AR intervals. When $a = 4$, and to a lesser extent when $a = 8$, they tend to be too short. In contrast, for $a = 16$, the median length of the Wald LIML intervals is just about equal to the dispersion of the $\hat{\beta}_{\text{LIML}}$. Moreover, even their 0.99 quantile is smaller than the median length of the AR intervals in that case.

The three right-hand panels of Figure 4 show the dispersion of $\hat{\beta}_{\text{IV}}$ and the median and 0.01 and 0.99 quantiles of the lengths of the 95% Wald IV intervals. For $a = 16$, these are almost indistinguishable from the results for the Wald LIML intervals. For the other values of a, the Wald IV intervals tend to be a bit shorter than the Wald LIML intervals. In particular, the 0.99 quantiles of their lengths are always smaller.

Several things stand out when we compare the bounded AR intervals with the Wald LIML and Wald IV ones for the replications where the former exist:

- The IV estimates are substantially less dispersed than the LIML ones for $a = 4$, moderately less dispersed for $a = 8$, and slightly less dispersed for $a = 16$. This implies that the AR and Wald LIML intervals should be longer than the Wald IV ones. In fact, the AR intervals tend to be much longer in all cases, while the Wald LIML ones are only a little bit longer even when $a = 4$.
- Whereas the median length of the AR intervals always overstates the dispersion of $\hat{\beta}_{\text{LIML}}$, that of the Wald LIML intervals always understates it (but not by much when $a = 16$). In contrast, the median length of the Wald IV intervals provides an excellent guide to the dispersion of $\hat{\beta}_{\text{IV}}$ for $a = 8$ and $a = 16$. For $a = 4$, it provides a slight underestimate.
- The lengths of the Wald intervals vary much less than those of the AR intervals. For example, when $a = 4$, the upper-tail quantile of the lengths of the Wald IV intervals is always less than 1.6 times the dispersion of $\hat{\beta}_{\text{IV}}$.

These results suggest that one would never want to use an AR confidence set when the instruments are reasonably strong. Even when they exist, AR intervals are much less informative than Wald ones. They do not have correct coverage conditional on being bounded and non-empty. Moreover, they do not provide reliable information about the dispersion of $\hat{\beta}_{\text{LIML}}$; they can be much too long or much too short.

In contrast, when $a = 16$, even the Wald IV interval works quite well. Its length provides a good guide to the dispersion of $\hat{\beta}_{\text{IV}}$, and its coverage is reasonably good. Coverage varies from 91.7% to 95.1% and always exceeds 94% for $|\rho| < 0.5$. The Wald LIML interval works even better, with coverage between 94.4% and 94.7% when $a = 16$ for all values of ρ. It is natural to ask whether one can improve upon these asymptotic Wald intervals by using the bootstrap. That turns out to be the case, and it is the topic of the next section.

5. Bootstrap Confidence Sets

There are numerous ways to bootstrap tests and confidence sets for β in the linear simultaneous equations model given by (1) and (2). Some of these are discussed in Davidson and MacKinnon (2008, 2010). However, the earlier paper does not discuss confidence sets at all, and the later one presents simulation results only for tests. In this section, we therefore provide some evidence on the performance of bootstrap confidence sets. Our objective is not to provide a detailed study of the many bootstrap methods that can be used to make inferences about β. It is simply to demonstrate that there exist bootstrap methods based on Wald (that is, t) statistics which provide excellent coverage, at least when the instruments are not too weak, and do not suffer from some of the undesirable features of AR confidence sets. However, the bootstrap must be used with care, as we show that there also exist bootstrap methods which can sometimes be less reliable than asymptotic ones.

The oldest, and conceptually the simplest, bootstrap DGP for the linear simultaneous equations model is the pairs bootstrap, which was proposed by Freedman (1984). The idea is simply to resample the rows of the matrix $[\boldsymbol{y}_1 \;\; \boldsymbol{y}_2 \;\; \boldsymbol{Z} \;\; \boldsymbol{W}_2]$. Each such bootstrap sample, indexed by $j = 1, \ldots, B$, is then used to compute a bootstrap test statistic

$$t_j^* = \frac{\hat{\beta}_j^* - \hat{\beta}}{s(\hat{\beta}_j^*)},$$

where $\hat{\beta}$ could be either the IV or LIML estimate of β, $\hat{\beta}_j^*$ is the corresponding estimate from the j^{th} bootstrap sample, and $s(\hat{\beta}_j^*)$ is the standard error of $\hat{\beta}_j^*$, which may or may not be robust to heteroskedasticity of unknown form. Using $\hat{\beta}$, $s(\hat{\beta})$, and the B values of t_j^*, one then constructs an equal-tail percentile t confidence interval (also called a studentized bootstrap confidence interval) in the usual way; see, among many others, Davison and Hinkley (1997) or Davidson and MacKinnon (2004, Chapter 5).

Figure 5 shows the coverage of asymptotic and pairs bootstrap confidence intervals based on t statistics for $\hat{\beta}_{\text{IV}}$ and $\hat{\beta}_{\text{LIML}}$. Since the DGP given by (9) and (10) has disturbances that are independent and identically distributed, we do not use heteroskedasticity-robust standard errors, but it would generally be advisable to do so in practice. To reduce computational costs, the bootstrap experiments use only 100,000 replications, instead of 500,000. For the same reason, ρ is varied from 0.00 to 0.99 by increments of 0.03 instead of 0.01. The bootstrap results are based on $B = 399$ bootstrap samples, which is a much smaller number than one would want to use in practice, but it is sufficient for a simulation experiment. Results are shown for $a = 2.8284$, $a = 4$, $a = 5.6569$, $a = 8$, $a = 11.3137$, and $a = 16$. Each of these is larger than its predecessor by a factor of $\sqrt{2}$. The instruments vary from extremely weak to very strong.

It is evident from the upper left-hand panel that asymptotic confidence intervals based on LIML t statistics always undercover. However, this undercoverage is very

modest for $a = 11.3137$ and $a = 16$, and it is not severe even for $a = 8$. Using the pairs bootstrap (lower left-hand panel) improves the results substantially. In most cases, the bootstrap intervals overcover. This overcoverage is most severe for $a = 4$ and $a = 5.6569$, but even in those cases it is fairly modest.

The performance of confidence intervals based on IV t statistics is much worse than that of ones based on LIML t statistics; see the two right-hand panels of Figure 5. The asymptotic intervals overcover when ρ is small and undercover, often severely, when ρ is large. In contrast, the pairs bootstrap intervals always undercover. Using the pairs bootstrap improves matters greatly for large values of ρ, but it actually makes them worse for small ones. For the two largest values of a, however, the undercoverage of the pairs bootstrap intervals is always quite modest.

The pairs bootstrap is by no means the only one for linear simultaneous equations models. In Davidson and MacKinnon (2008), we proposed the restricted efficient, or RE, bootstrap. As we will demonstrate shortly, confidence intervals based on the RE bootstrap perform very much better than ones based on the pairs bootstrap. However, they are more complicated and expensive to compute.

The RE bootstrap has two key features. The bootstrap DGP is conditional on a particular value of β (hence "restricted"), and it uses an efficient estimate of $\boldsymbol{\pi}$ (hence "efficient"). For any specified value β_0, we can run regression (1) to obtain parameter estimates $\tilde{\boldsymbol{\gamma}}$ and residuals $\tilde{\boldsymbol{u}}_1$. The latter may be rescaled by multiplying them by a factor of $(n/(n-k))^{1/2}$. We then run the regression

$$\boldsymbol{y}_2 = \boldsymbol{W}\boldsymbol{\pi} + \delta\tilde{\boldsymbol{u}}_1 + \text{residuals}. \tag{18}$$

This yields parameter estimates $\tilde{\boldsymbol{\pi}}$ and adjusted residuals $\tilde{\boldsymbol{u}}_2 \equiv \boldsymbol{y}_2 - \boldsymbol{W}\tilde{\boldsymbol{\pi}}$. The latter should be rescaled by multiplying them by a factor of $(n/(n-l))^{1/2}$. It can be shown that $\tilde{\boldsymbol{\pi}}$ is asymptotically equivalent to the estimate one would obtain by using FIML or 3SLS. This estimate was used by Kleibergen (2002) in a different context.

Generating a bootstrap sample using the RE bootstrap is quite simple. We form two vectors of bootstrap disturbances, $\boldsymbol{u}_1^*$ and $\boldsymbol{u}_2^*$, with elements u_{i1}^* and u_{i2}^* for $i = 1, \ldots, n$, resampled from the pairs of rescaled residuals. We then set

$$\begin{aligned} \boldsymbol{y}_2^* &= \boldsymbol{W}\tilde{\boldsymbol{\pi}} + \boldsymbol{u}_2^*, \text{ and} \\ \boldsymbol{y}_1^* &= \beta_0\boldsymbol{y}_2^* + \boldsymbol{Z}\tilde{\boldsymbol{\gamma}} + \boldsymbol{u}_1^*. \end{aligned} \tag{19}$$

If we generate B bootstrap samples, we can compute an equal-tail bootstrap P value for the hypothesis that $\beta = \beta_0$. It is simply

$$\hat{p}^*(\beta_0) = \frac{2}{B}\min\left(\sum_{j=1}^{B} \mathrm{I}(\tau_j^* < \hat{\tau}),\ \mathrm{I}(\tau_j^* \geq \hat{\tau})\right), \tag{20}$$

where $\mathrm{I}(\cdot)$ is the indicator function, $\hat{\tau} = (\hat{\beta} - \beta_0)/s(\hat{\beta})$, and $\tau_j^* = (\hat{\beta}_j^* - \beta_0)/s(\hat{\beta}_j^*)$. Here $\hat{\beta}$ may denote either $\hat{\beta}_{\mathrm{IV}}$ or $\hat{\beta}_{\mathrm{LIML}}$, and $\hat{\beta}_j^*$ is the corresponding estimate from

the j^{th} bootstrap sample. It is important to calculate the standard errors $s(\hat{\beta})$ and $s(\hat{\beta}_j^*)$ in the same way. By using the equal-tail P value (20), we do not impose symmetry on the distribution of τ.

The wild restricted efficient, or WRE, bootstrap (Davidson and MacKinnon, 2010) is very similar to the RE bootstrap, except that the i^{th} pair of rescaled residuals remains associated with the i^{th} observation. To generate the bootstrap disturbances, one simply multiplies each pair of rescaled residuals by a random variable v_i^* with mean zero and variance one. See Davidson and Flachaire (2008) for more about the wild bootstrap. Unless heteroskedasticity is clearly absent, it would probably be wise to use heteroskedasticity-consistent standard errors and the WRE bootstrap. In samples of reasonable size (more than a few hundred observations) with heteroskedastic disturbances, this combination should work just about as well as ordinary standard errors and the RE bootstrap when the disturbances are actually homoskedastic.

Using the RE bootstrap to obtain a confidence set is a bit complicated. Consider the upper limit, $\hat{\beta}_u$. Start with an initial estimate, say $\hat{\beta}_u^1$ (one obvious candidate is the upper limit of the asymptotic confidence interval) and compute $\hat{p}^*(\hat{\beta}_u^1)$ using equation (20). If $\hat{p}^*(\hat{\beta}_u^1) > \alpha$, then $\hat{\beta}_u^1$ is too small; if $\hat{p}^*(\hat{\beta}_u^1) < \alpha$, then it is too large. Try another candidate, say $\hat{\beta}_u^2$, which must be larger than $\hat{\beta}_u^1$ in the former case and smaller in the latter case. Calculate $\hat{p}^*(\hat{\beta}_u^2)$ and repeat if necessary. The way in which $\hat{\beta}_u^2$ is chosen may have a significant impact on computational cost, but it should have no effect on the properties of the RE bootstrap confidence set.

If, after m tries, we have found $\hat{\beta}_u^{m-1}$ and $\hat{\beta}_u^m$ such that $\hat{p}^*(\hat{\beta}_u^{m-1}) - \alpha$ and $\hat{p}^*(\hat{\beta}_u^m) - \alpha$ have opposite signs, then $\hat{\beta}_u$ must lie between them. At this point, various numerical methods can be used to find it. Since $\hat{p}^*(\beta_0)$ is not differentiable, we must use a method that does not need derivatives. In Davidson and MacKinnon (2010), we used golden section search, but in this paper we use bisection, which is easier to program and somewhat faster. Note that exactly the same set of random numbers must be used for all the bootstrap samples. Otherwise, the value of $\hat{p}^*(\beta_0)$ would be different each time we evaluated it.

The procedure for finding the lower limit, $\hat{\beta}_l$, is essentially the same as the one for finding the upper limit, with obvious changes in sign at various points.

In the above description of the algorithm, we have implicitly assumed that, if β_0 is sufficiently large or sufficiently small, $\hat{p}^*(\beta_0)$ must be less than α. However, that is not always true. The confidence set has no upper bound if $p^*(\beta_0) > \alpha$ as β_0 tends to plus infinity, and it has no lower bound if $p^*(\beta_0) > \alpha$ as β_0 tends to minus infinity. In practice, we may reasonably conclude that the confidence set is unbounded from above (below) if $p^*(\beta_0) > \alpha$ for a very large positive (negative) value of β_0.

Like AR confidence sets, unbounded RE bootstrap confidence sets may contain holes. It is therefore important to check for unboundedness even if the procedure described above has apparently located both $\hat{\beta}_u$ and $\hat{\beta}_l$. If there are values of β_0 greater than $\hat{\beta}_u$ or less than $\hat{\beta}_l$ for which $p^*(\beta_0) > \alpha$, it is easy enough to locate the other end of the hole. However, we do not recommend using unbounded confidence sets to make

inferences. The fact that a confidence set is unbounded strongly suggests that the instruments are so weak as to make reliable inference impossible.

The fact that RE bootstrap confidence sets may be unbounded (and in fact often are unbounded for small values of a; see below) is actually a good feature. The important result of Dufour (1997) implies that any confidence set which has approximately correct coverage when the instruments are weak must be unbounded with positive probability. The fact that RE bootstrap confidence sets can be unbounded makes it possible for them to have extremely good coverage.

Figure 6 shows coverage for RE bootstrap confidence sets based on LIML (left panel) and IV (right panel) estimates for the same simulations (including the same random numbers to generate the data) as the two lower panels of Figure 5. Note the scales of the vertical axes! For $a \geq 8$ in the LIML case and $a \geq 11.3137$ in the IV case, it is impossible to see any evidence that coverage is not precisely 95%. For the two largest values of a, there is also no statistical evidence that coverage differs from 95%, either unconditionally or conditional on ρ, for either the IV or LIML intervals.

For the RE bootstrap LIML confidence sets, there is very slight undercoverage when $a = 5.6569$, noticeable undercoverage when $a = 4$, and more serious undercoverage when $a = 2.8284$. The undercoverage is always more severe for small values of ρ than for large ones. For the RE bootstrap IV confidence sets, there can be either overcoverage (for $a = 2.8284$ and, except for small values of ρ, for $a = 4$) or undercoverage (for $a = 5.6569$, except for large values of ρ, and for $a = 8$, except for small values of ρ). But the magnitudes of both overcoverage and undercoverage are very small indeed. To see how extraordinarily well the RE bootstrap confidence sets perform, compare Figure 6 with Figure 5.

Figure 7 shows the proportion of RE bootstrap confidence sets that are actually bounded intervals. For comparison, the proportion of AR confidence sets that are bounded is also shown. No results are presented for $a = 16$ and $a = 11.3137$, because all the RE bootstrap sets were bounded intervals in those cases. For the LIML case, the fraction of bounded intervals is negligibly different from 100% when $a = 8$ and never less than 98.5% when $a = 5.6569$. For smaller values of a, it is comparable to the proportion of bounded AR confidence sets, although it is much more sensitive to the value of ρ. For the IV case, the fraction of bounded intervals is smaller and more dependent on the value of ρ than for the other two cases.

Whenever some RE bootstrap confidence sets are unbounded, it is inevitable that others should be bounded but extremely long. Even for the larger values of a, where this does not happen, the RE bootstrap intervals tend to be somewhat longer than the asymptotic ones (which is not surprising, since their coverage is much better). However, they tend to be much shorter and much less variable in length than AR intervals. For example, when $a = 16$, the 0.99, 0.50, and 0.01 quantiles of the RE bootstrap IV intervals for $\rho = 0.51$ are 0.333, 0.254, and 0.201, respectively. There is just a small amount of dependence on ρ, so results for $\rho = 0.51$ are quite typical, and we do not need a figure. The same quantiles of the RE bootstrap LIML intervals are

virtually identical, at 0.330, 0.253, and 0.200. In contrast, the quantiles of the AR intervals are 0.540, 0.379, and 0.112. The dispersions of the IV and LIML estimates (see the discussion of Figure 4) are 0.243 and 0.250, respectively, so the lengths of the RE bootstrap intervals, in striking contrast to those of the AR intervals, provide reasonably reliable guides to the dispersion of the estimates.

6. Conclusion

In this paper, we have studied the properties of confidence sets based on inverting the Anderson-Rubin test, and we have proposed a bootstrap procedure that appears to have very much better properties. The fundamental problem with AR confidence sets is not that they can be unbounded when the instruments are weak. That feature is shared by the RE bootstrap confidence sets that we propose. Moreover, as Dufour (1997) showed, it is a necessary feature of any confidence set which has approximately correct coverage when the instruments are weak.

Instead, the fundamental problem with AR confidence sets is that they are obtained by inverting a test which has more than one degree of freedom, specifically, $l - k$ of them. As a result, the length (and even the existence) of an AR confidence set, when it is not unbounded, depends on the value of the Sargan statistic. When the instruments are reasonably strong, the AR confidence set will be empty whenever the Sargan statistic is sufficiently large. Bounded AR confidence intervals will be misleadingly short if the Sargan statistic is large, and they will be misleadingly long if the Sargan statistic is small. The coverage of bounded AR intervals is not correct (they have to overcover to make up for the empty sets that undercover), and their coverage conditional on the value of the Sargan statistic is much worse. We therefore do not recommend the use of AR confidence sets in any circumstances.

In Section 5, we proposed a procedure for constructing confidence intervals by inverting bootstrap tests. Unlike most such procedures, this one employs restricted and efficient estimates of the bootstrap DGP. It can be used with any asymptotically valid test, but we applied it only to t tests based on IV and LIML estimates. For sufficiently strong instruments, where the resulting confidence sets are always bounded intervals, this procedure appears to work perfectly. For weaker ones, where some of the confidence sets are unbounded, its coverage is not perfect, but it still appears to be extremely good.

Appendix

In this appendix, we study the properties of confidence sets that are constructed by inverting F tests in the classical normal linear model

$$\boldsymbol{y} = \boldsymbol{x}\beta + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\mathbf{I}), \tag{21}$$

where $\boldsymbol{y}$ and $\boldsymbol{x}$ are $n \times 1$ vectors, $\boldsymbol{X}_2$ is an $n \times k_2$ matrix, $\boldsymbol{Z}$ is an $n \times k_3$ matrix. We wish to construct a confidence set for β by inverting the F test for the joint hypothesis

$$H(\beta_0): \qquad \beta = \beta_0; \quad \boldsymbol{\beta}_2 = \boldsymbol{0},$$

assuming of course that the true $\boldsymbol{\beta}_2$ is indeed zero. The null model can be written as

$$\boldsymbol{y} - \boldsymbol{x}\beta_0 = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u},$$

and the alternative as

$$\boldsymbol{y} - \boldsymbol{x}\beta_0 = \boldsymbol{Y}\boldsymbol{\delta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}, \tag{22}$$

where $\boldsymbol{Y} \equiv [\boldsymbol{x} \;\; \boldsymbol{X}_2]$. Clearly, (21) and (22) are just different parametrizations of the same model.

The F statistic for a test of $H(\beta_0)$ at nominal level α is

$$F(\beta_0) = \frac{\|\boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{Y}}(\boldsymbol{y} - \boldsymbol{x}\beta_0)\|^2/(k_2+1)}{\|\boldsymbol{M}_{[\boldsymbol{Y}\;\boldsymbol{Z}]}\boldsymbol{y}\|^2/(n-k)},$$

where $k = k_2 + k_3 + 1$. Any value of β_0 for which $F(\beta_0) \leq q$, where q is the $1-\alpha$ quantile of the $F_{k_2+1,n-k}$ distribution, belongs to the confidence set formed by inverting the F test. The inequality $F(\beta_0) \leq q$ can be expressed as a quadratic inequality in β_0, as follows:

$$(\boldsymbol{x}^\top \boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{Y}} \boldsymbol{x})\beta_0^2 - 2(\boldsymbol{x}^\top \boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{Y}} \boldsymbol{y})\beta_0 + \boldsymbol{y}^\top(\boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{Y}} - c\boldsymbol{M}_{[\boldsymbol{Y}\;\boldsymbol{Z}]})\boldsymbol{y} \leq 0, \tag{23}$$

where $c \equiv (k_2+1)q/(n-k)$. The discriminant of the quadratic is

$$\Delta \equiv 4\big((\boldsymbol{x}^\top \boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{Y}} \boldsymbol{y})^2 - \boldsymbol{x}^\top \boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{Y}} \boldsymbol{x}\; \boldsymbol{y}^\top(\boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{Y}} - c\boldsymbol{M}_{[\boldsymbol{Y}\;\boldsymbol{Z}]})\boldsymbol{y}\big), \tag{24}$$

and $\Delta < 0$ if and only if

$$\boldsymbol{y}^\top(\boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{Y}} - c\boldsymbol{M}_{[\boldsymbol{Y}\;\boldsymbol{Z}]})\boldsymbol{y} > \frac{(\boldsymbol{x}^\top \boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{Y}} \boldsymbol{y})^2}{\boldsymbol{x}^\top \boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{Y}} \boldsymbol{x}}. \tag{25}$$

The right-hand side of this inequality is the squared norm of the projection of $\boldsymbol{y}$ on to the direction of $\boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{Y}}\boldsymbol{x}$. But

$$\boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{Y}}\boldsymbol{x} = \boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{Y}}\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{x} = \boldsymbol{M}_\boldsymbol{Z}\boldsymbol{x},$$

and so the right-hand side of (25) is simply $\boldsymbol{y}^\top \boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{x}} \boldsymbol{y}$.

If we subtract $\boldsymbol{y}^\top \boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{x}} \boldsymbol{y}$ from both sides of (25), the first term inside the parentheses on the left-hand side becomes $\boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{Y}} - \boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{x}}$. Since

$$\boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{Y}} = \boldsymbol{P}_{\boldsymbol{M}_\boldsymbol{Z}\boldsymbol{x}} + \boldsymbol{P}_{\boldsymbol{M}_{[\boldsymbol{x}\;\boldsymbol{Z}]}\boldsymbol{X}_2}, \tag{26}$$

the inequality (25) can then be rewritten as

$$\boldsymbol{y}^\top(\boldsymbol{P}_{\boldsymbol{M}_{[\boldsymbol{x}\;\boldsymbol{Z}]}\boldsymbol{X}_2} - c\boldsymbol{M}_{[\boldsymbol{Y}\;\boldsymbol{Z}]})\boldsymbol{y} > 0,$$

which can be rearranged as

$$\frac{\boldsymbol{y}^\top \boldsymbol{P}_{\boldsymbol{M}_{[\boldsymbol{x}\ \boldsymbol{Z}]}\boldsymbol{X}_2}\boldsymbol{y}/k_2}{\boldsymbol{y}^\top \boldsymbol{M}_{[\boldsymbol{Y}\ \boldsymbol{Z}]}\boldsymbol{y}/(n-k)} > \Big(1+\frac{1}{k_2}\Big)q. \tag{27}$$

The left-hand side of this inequality is distributed as $F_{k_2,n-k}$, and so the probability that $\Delta < 0$ can be readily calculated. The numerical value depends on the nominal coverage $1-\alpha$, the sample size n, and the numbers k and k_2 of regressors in the model (21).

The probability that $\Delta < 0$ is the probability of obtaining an empty confidence set, because the coefficient of β_0^2 in (23) is always positive, so that, if the corresponding quadratic equation has no real roots, the quadratic function is everywhere positive, and the inequality is satisfied nowhere. This probability is, of course, less than α.

Suppose without loss of generality that the true value of β is zero and the true value of σ is one. Then the confidence set covers zero if and only if it is non-empty, that is, $\Delta > 0$, and the two real roots of the quadratic have opposite signs. The product of the roots is the ratio of the last term on the left-hand side of (23) to the coefficient of β_0^2. Since the latter is always positive, the roots have opposite signs if and only if

$$\boldsymbol{y}^\top(\boldsymbol{P}_{\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{Y}} - c\boldsymbol{M}_{[\boldsymbol{Y}\ \boldsymbol{Z}]})\boldsymbol{y} \le 0, \tag{28}$$

since this inequality implies that $\Delta > 0$; compare (25). The inequality (28) can be rewritten as

$$\frac{\boldsymbol{y}^\top \boldsymbol{P}_{\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{Y}}\boldsymbol{y}/(k_2+1)}{\boldsymbol{y}^\top \boldsymbol{M}_{[\boldsymbol{Y}\ \boldsymbol{Z}]}\boldsymbol{y}/(n-k)} \le q, \tag{29}$$

and the probability that the inequality is satisfied is of course just $1-\alpha$, since the left-hand side of (29) is distributed as $F_{k_2+1,n-k}$.

Consider next the statistic for the F test of the part of $H(\beta_0)$ that has nothing to do with β_0, namely that $\boldsymbol{\beta}_2 = \boldsymbol{0}$. This statistic is

$$F_2 \equiv \frac{\boldsymbol{y}^\top \boldsymbol{P}_{\boldsymbol{M}_{[\boldsymbol{x}\ \boldsymbol{Z}]}\boldsymbol{X}_2}\boldsymbol{y}/k_2}{\boldsymbol{y}^\top \boldsymbol{M}_{[\boldsymbol{Y}\ \boldsymbol{Z}]}\boldsymbol{y}/(n-k)}. \tag{30}$$

From (26), the left-hand side of (29) can be rewritten as

$$\frac{k_2}{k_2+1}F_2 + \frac{\boldsymbol{y}^\top \boldsymbol{P}_{\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{x}}\boldsymbol{y}/(k_2+1)}{\boldsymbol{y}^\top \boldsymbol{M}_{[\boldsymbol{Y}\ \boldsymbol{Z}]}\boldsymbol{y}/(n-k)},$$

and so, if we write $s^2 = \boldsymbol{y}^\top \boldsymbol{M}_{[\boldsymbol{Y}\ \boldsymbol{Z}]}\boldsymbol{y}/(n-k)$, the coverage event can be expressed as

$$k_2F_2 + \frac{\boldsymbol{y}^\top \boldsymbol{P}_{\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{x}}\boldsymbol{y}}{s^2} \le (k_2+1)q,$$

or, equivalently,

$$\boldsymbol{y}^\top \boldsymbol{P}_{\boldsymbol{M_Z x}} \boldsymbol{y} \leq s^2\big((k_2+1)q - k_2 F_2\big). \tag{31}$$

The two sides of this inequality are independent, and the left-hand side is distributed as $\chi^2(1)$. Therefore, conditional on F_2 and s^2, coverage is given by the CDF of $\chi^2(1)$ evaluated at the right-hand side of (31).

Observe from (27) and (30) that the event $\Delta < 0$ can be written as $k_2 F_2 > (k_2+1)q$, which means that (31) cannot be satisfied if $\Delta < 0$. Moreover, the length of the confidence interval, when it exists, is the distance between the two roots of the quadratic in (23), that is, $2\sqrt{\Delta}/\boldsymbol{x}^\top \boldsymbol{P}_{\boldsymbol{M_Z Y}} \boldsymbol{x}$. It can be seen from (24) and (30) that

$$\Delta = 4 \boldsymbol{x}^\top \boldsymbol{P}_{\boldsymbol{M_Z Y}} \boldsymbol{x}\; s^2\big((k_2+1)q - k_2 F_2\big),$$

and so the length of the interval, when it exists, is

$$\frac{2s\big((k_2+1)q - k_2 F_2\big)^{1/2}}{(\boldsymbol{x}^\top \boldsymbol{P}_{\boldsymbol{M_Z Y}} \boldsymbol{x})^{1/2}}. \tag{32}$$

This may be compared with expression (17) for the bounded AR interval. We see from (31) and (32) that conditional coverage is given by the CDF of $\chi^2(1)$ evaluated at 4 times the squared length of the confidence interval multiplied by $\boldsymbol{x}^\top \boldsymbol{P}_{\boldsymbol{M_Z Y}} \boldsymbol{x}$.

It is evident from expression (32) that, if $\hat{\boldsymbol{\beta}}_2$ differed substantially from a zero vector, and F_2 were consequently a large number, $(k_2+1)q - k_2 F_2$ would be negative, and there would not exist a bounded interval. That could happen either by chance or because $\boldsymbol{\beta}_2 \neq \boldsymbol{0}$. It seems very unsatisfactory that the length, and even the existence, of a confidence interval for β should depend on the value of $\boldsymbol{\beta}_2$. That is one of the reasons why the interval discussed in this appendix would never be used in practice. But the AR confidence set suffers from exactly the same defects. The analog of $\boldsymbol{\beta}_2$ not being quite zero is the overidentifying restrictions not being quite satisfied.

References

Anderson, T. W., and H. Rubin (1949), "Estimation of the parameters of a single equation in a complete system of stochastic equations," *Annals of Mathematical Statistics*, 20, 46–63.

Davidson, R., and E. Flachaire (2008), "The wild bootstrap, tamed at last," *Journal of Econometrics*, 146, 162–169.

Davidson, R. and J. G. MacKinnon (2008), "Bootstrap inference in a linear equation estimated by instrumental variables," *Econometrics Journal*, 11, 443–477.

Davidson, R. and J. G. MacKinnon (2010), "Wild bootstrap tests for IV regression," *Journal of Business and Economic Statistics*, 28, 128–144.

Davison, A. C., and D. V. Hinkley (1997), *Bootstrap Methods and Their Application*, Cambridge, Cambridge University Press.

Dufour, J.-M. (1997), "Some impossibility theorems in econometrics with applications to structural and dynamic models," *Econometrica*, 65, 1365–1388.

Dufour, J.-M., and M. Taamouti (2005), "Projection-based statistical inference in linear structural models with possibly weak instruments," *Econometrica*, 73, 1351–1365.

Forchini, G., and G. Hillier (2003), "Conditional inference for possibly unidentified structural equations," *Econometric Theory*, 19, 707–743.

Freedman, D. A. (1984), "On bootstrapping stationary two-stage least-squares estimates in stationary linear models," *Annals of Statistics*, 12, 827–842.

Kleibergen, F. (2002), "Pivotal statistics for testing structural parameters in instrumental variables regression," *Econometrica*, 70, 1781–1803.

Phillips, P. C. B. (1983), "Exact small sample theory in the simultaneous equations model," Chp. 8 in Z. Griliches and M. D. Intriligator (eds.), *Handbook of Econometrics*, Vol. 1, pp. 449–516. Amsterdam: North Holland.

Sargan, J. D. (1958), "The estimation of economic relationships using instrumental variables," *Econometrica*, 26, 393–415.

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

Stock, J. H., J. H. Wright, and M. Yogo (2002), "A survey of weak instruments and weak identification in generalized method of moments," *Journal of Business and Economic Statistics*, 20, 518–29.

Stock, J. H., and M. Yogo (2005), "Testing for weak instruments in linear IV regression," in D. W. K. Andrews and J. H. Stock (eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, 80–108. Cambridge: Cambridge University Press.

Zivot, E., R. Startz, and C. R. Nelson (1998), "Valid confidence intervals and inference in the presence of weak instruments," *International Economic Review*, 39, 1119–1144.

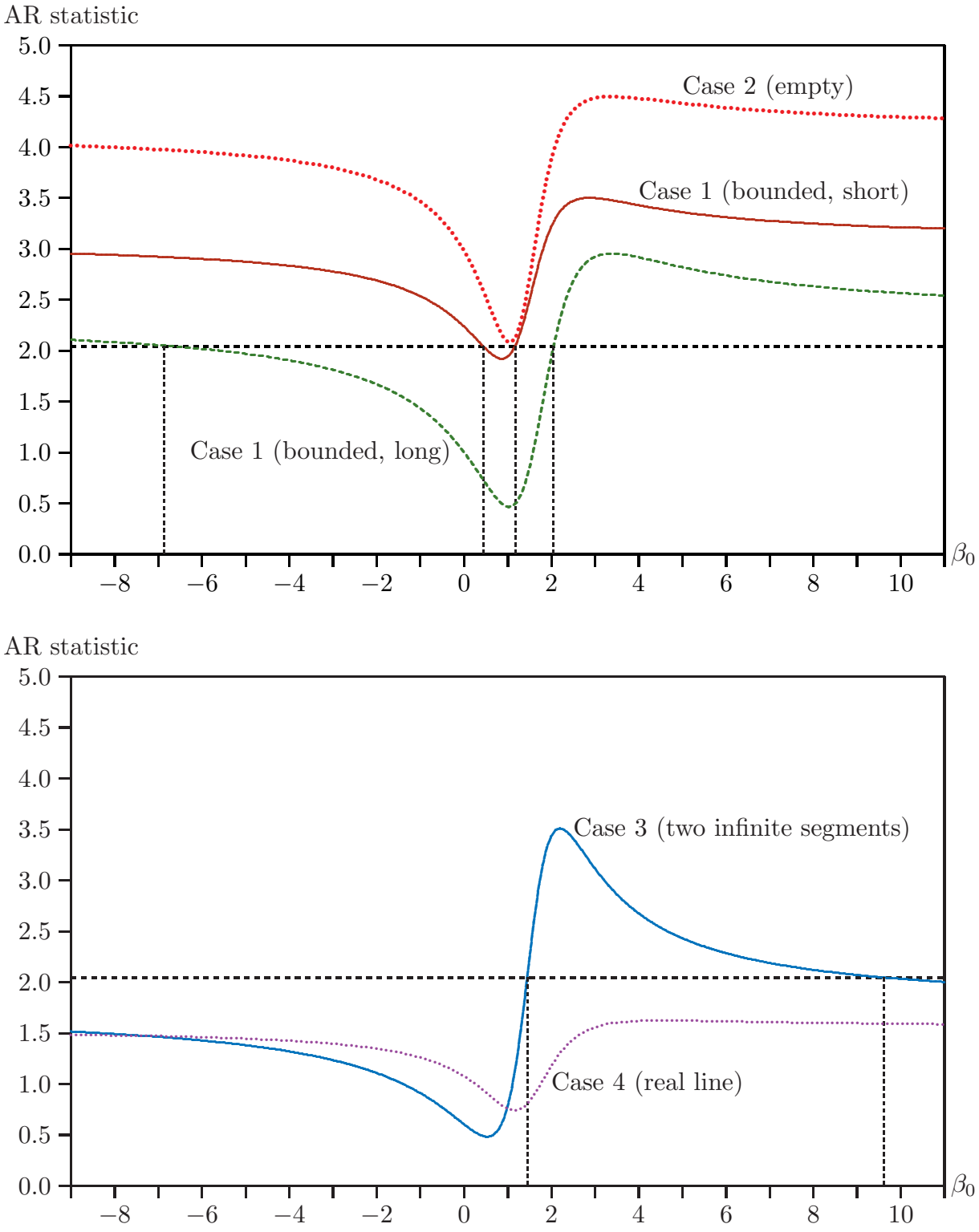

Figure 1. Five types of Anderson-Rubin confidence set

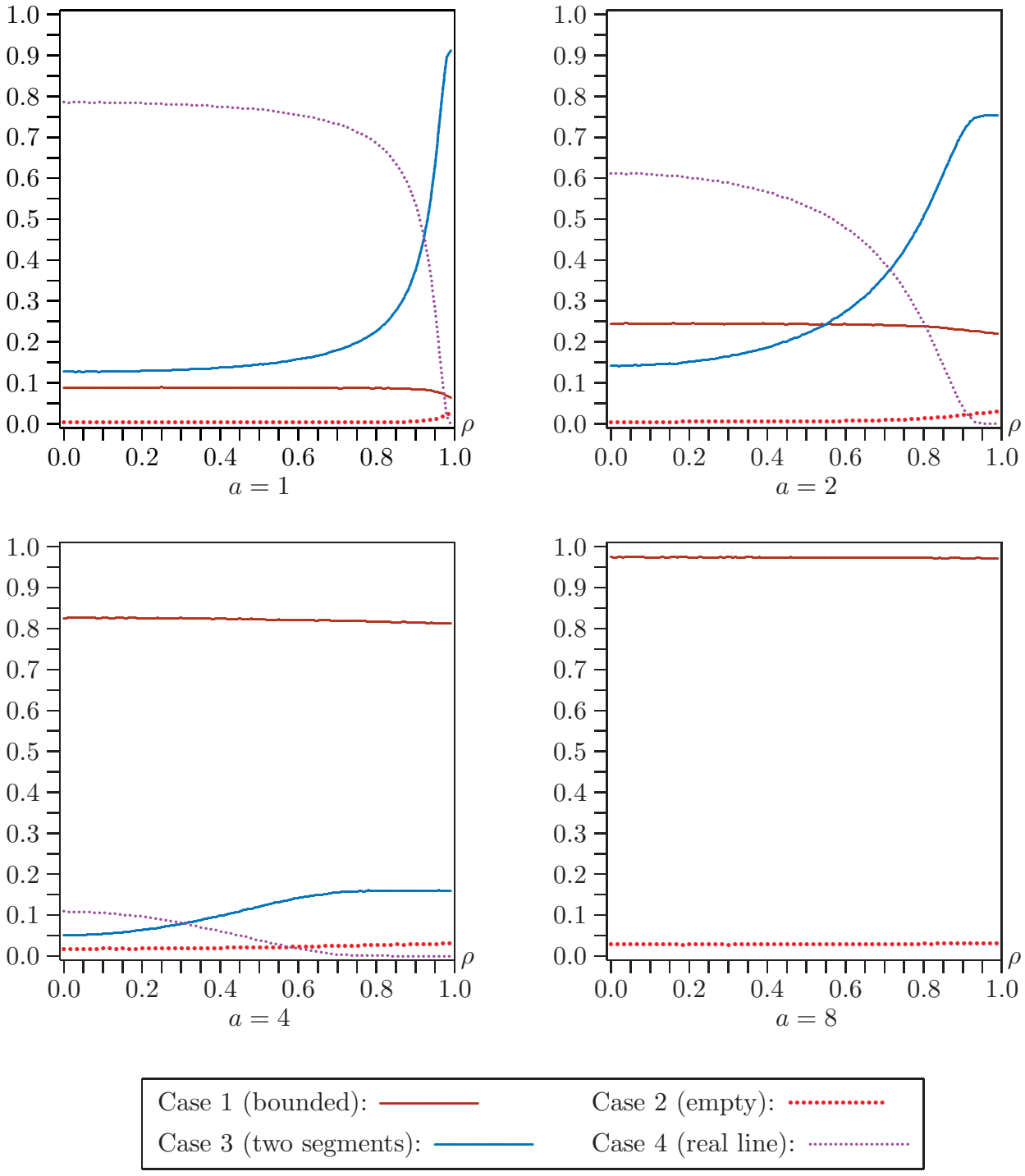

Figure 2. Frequencies of each type of confidence set as functions of ρ

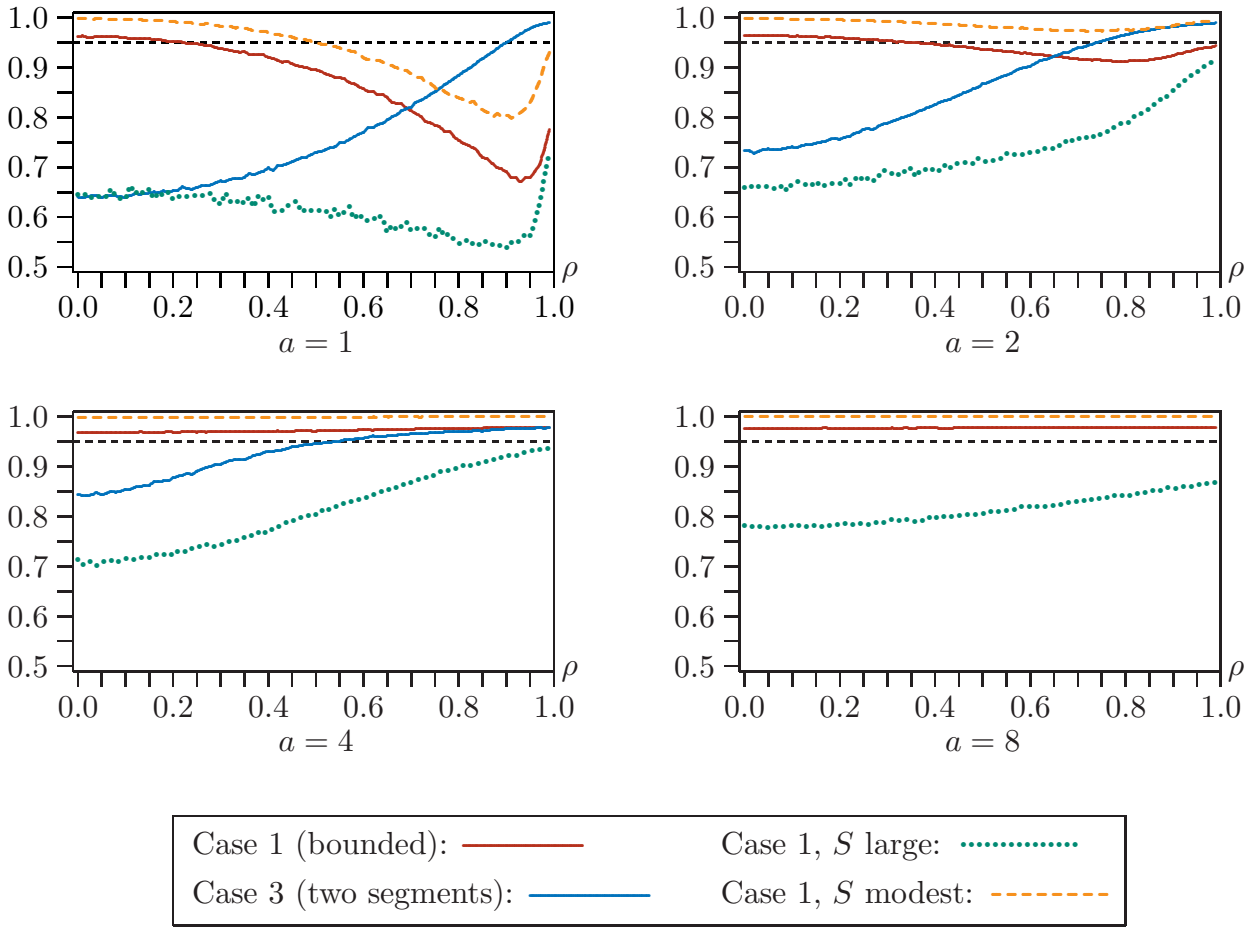

Figure 3. Coverage of each type of confidence set as functions of ρ

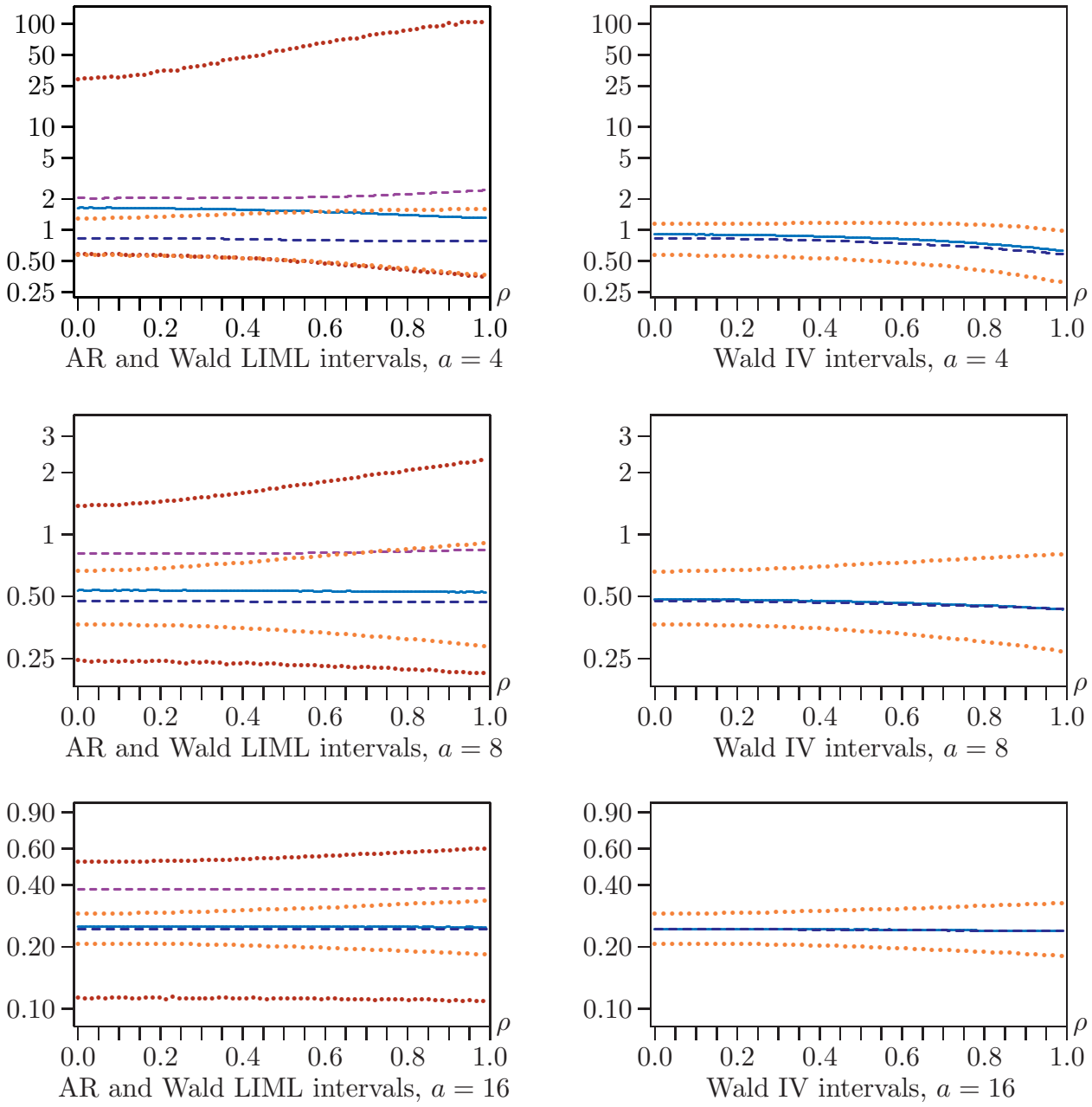

Figure 4. Dispersion of estimates and lengths of confidence intervals

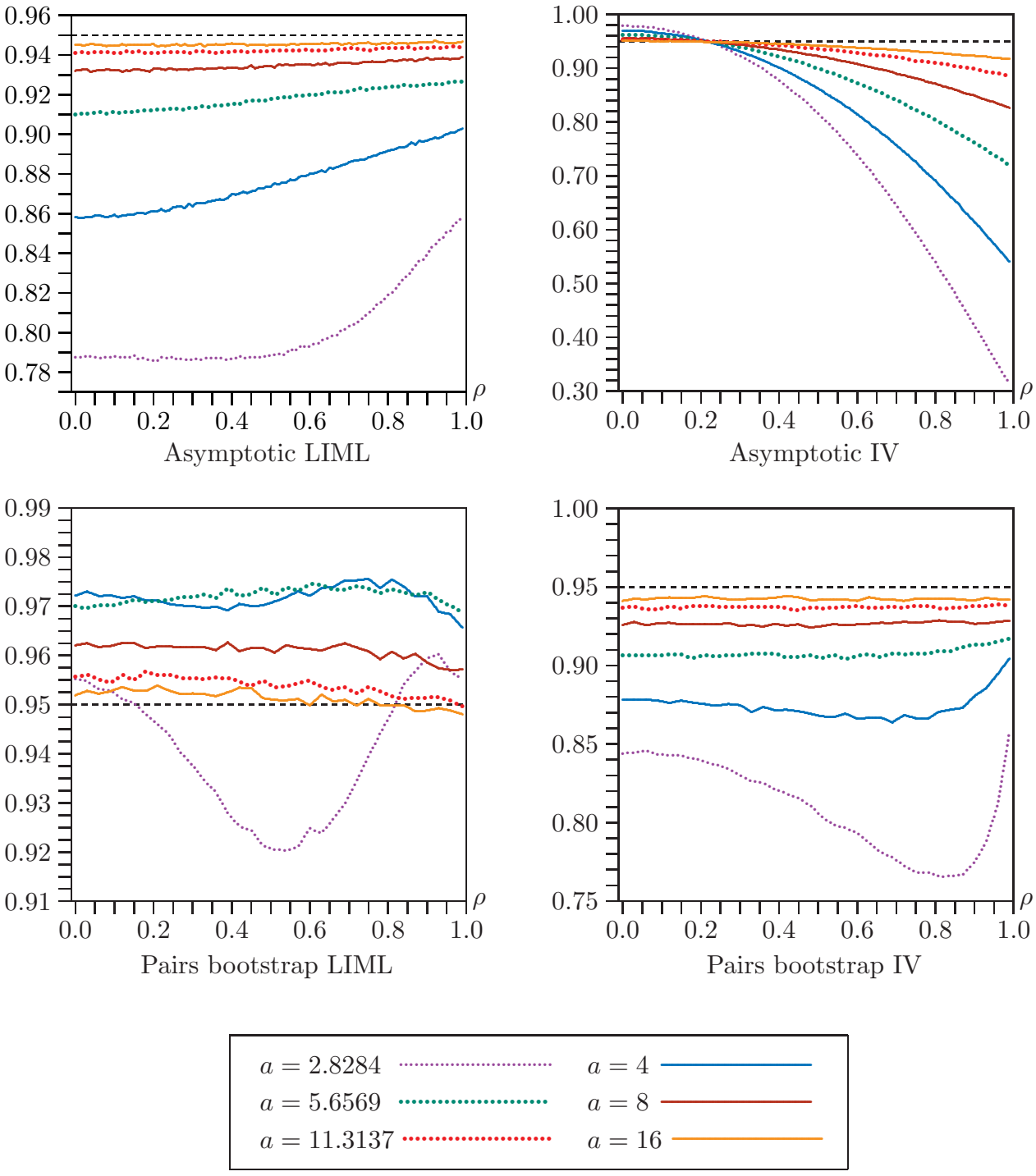

Figure 5. Coverage of asymptotic and pairs bootstrap Wald confidence sets as functions of ρ

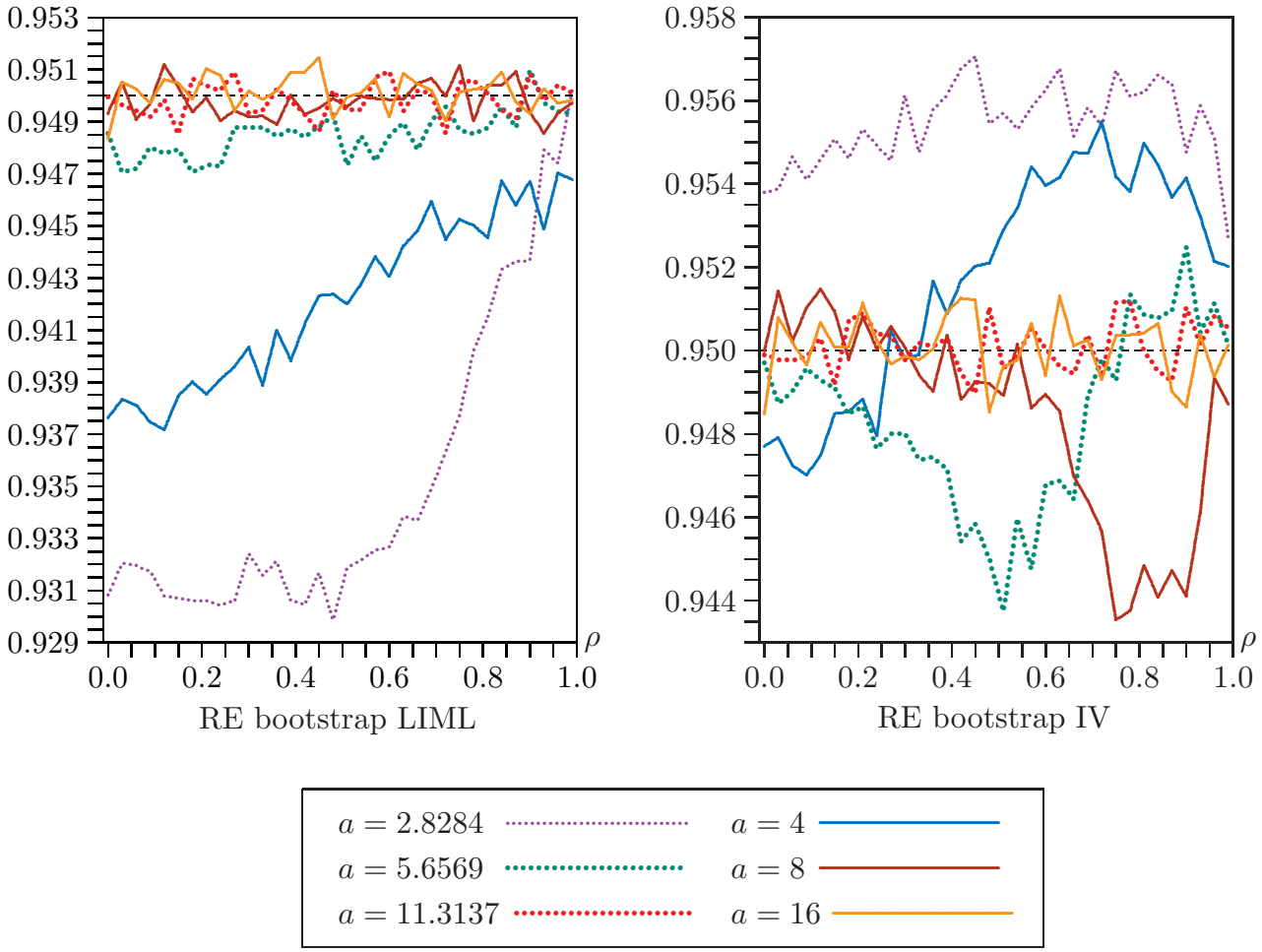

Figure 6. Coverage of restricted efficient bootstrap Wald confidence sets as functions of ρ

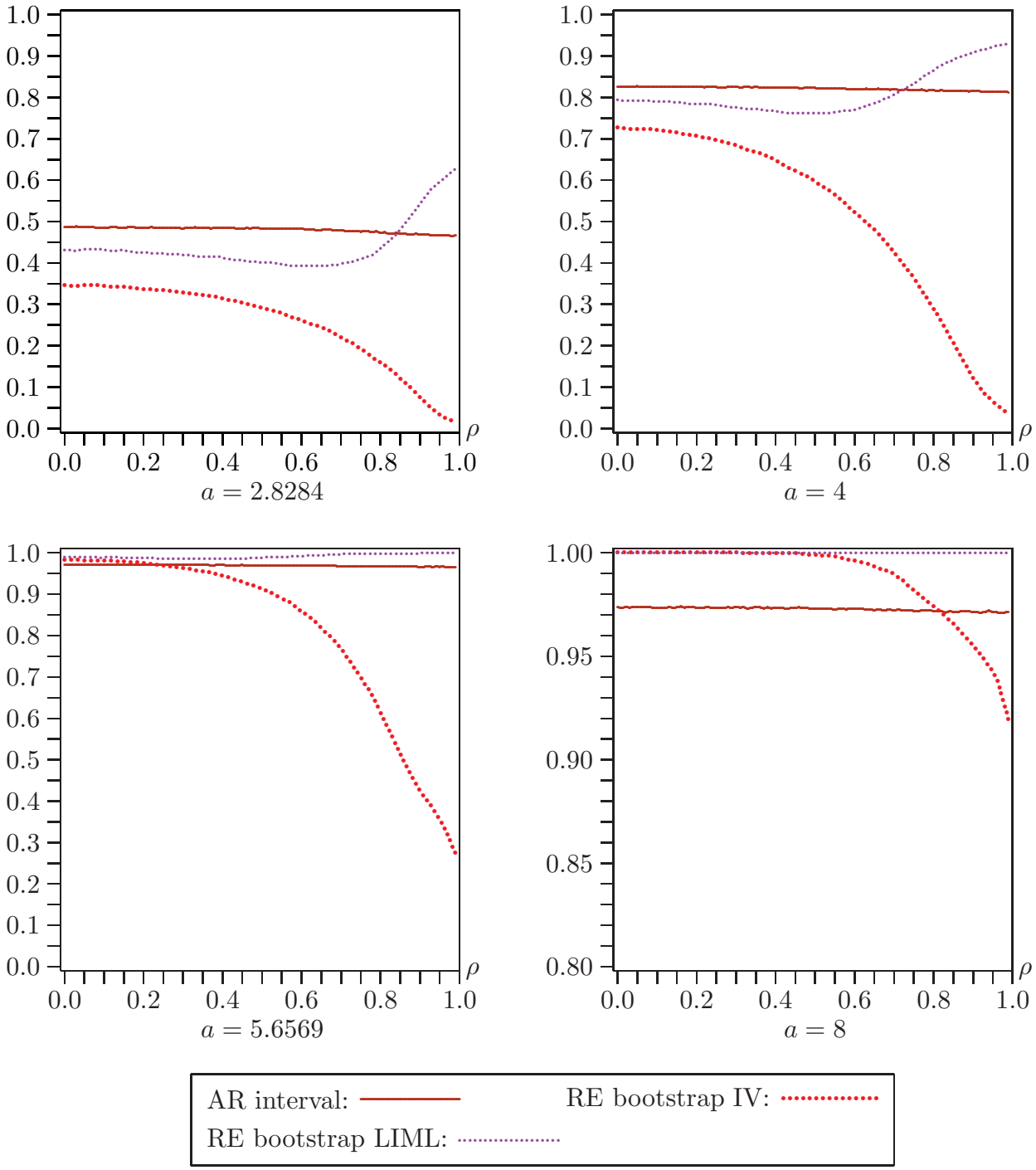

Figure 7. Proportions of bounded confidence intervals as functions of ρ