Bootstrap Inference in a Linear Equation Estimated by Instrumental Variables

Russell Davidson

GREQAM
Centre de la Vieille Charité
2 rue de la Charité
13236 Marseille cedex 02, France

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

email: **Russell.Davidson@mcgill.ca**

and

James G. MacKinnon

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

email: **jgm@econ.queensu.ca**

Abstract

We study several tests for the coefficient of the single right-hand-side endogenous variable in a linear equation estimated by instrumental variables. We show that writing all the test statistics — Student's t, Anderson-Rubin, Kleibergen's K, and likelihood ratio (LR) — as functions of six random quantities leads to a number of interesting results about the properties of the tests under weak-instrument asymptotics. We then propose several new procedures for bootstrapping the three non-exact test statistics and also a new conditional bootstrap version of the LR test. These use more efficient estimates of the parameters of the reduced-form equation than existing procedures. When the best of these new procedures is used, both the K and conditional bootstrap LR tests have excellent performance under the null. However, power considerations suggest that the latter is probably the method of choice.

JEL codes: C10, C12, C15, C30

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. We are grateful to five referees and to seminar participants at the University of New South Wales, the University of Sydney, the University of California Santa Barbara, and the University of Michigan for comments on earlier versions.

Revised, August 2007

1. Introduction

This paper is concerned with tests for the value of the coefficient of the single right-hand-side endogenous variable in a linear structural equation estimated by instrumental variables. We consider the Wald (or t) test, Kleibergen's (2002) K test, which can be thought of as an LM test, and the likelihood ratio (LR) test, as well as its conditional variant (Moreira, 2003), which we refer to as CLR. Both asymptotic and bootstrap versions of these tests are studied, and their relationships to the Anderson-Rubin (AR) test (Anderson and Rubin, 1949) are explored. The analysis allows for instruments that may be either strong or weak.

The major theoretical contributions of the paper depend on a simple way of writing all the test statistics of interest as functions of six random quantities. This makes it easy to understand the properties of all the tests under both weak and strong instruments. Our theoretical results are supported by extensive simulations.

Our results also make it inexpensive to simulate and bootstrap all the test statistics under the assumption of normally distributed disturbances. Although, in practice, it is generally preferable to use bootstrap methods based on resampling residuals, our experiments suggest that results from the parametric bootstrap under normality provide a very good guide to the performance of methods that resample residuals.

The paper's main practical contribution is to propose new procedures for bootstrapping the test statistics we study. These use more efficient estimates of the parameters of the reduced-form equation than existing procedures, and what seems to be the best procedure also employs a form of bias correction. Using this procedure instead of more conventional ones greatly improves the performance under the null of all the tests. The improvement is generally greatest for the Wald test and least for the K test, because the latter already works very well in most cases. Using the new procedures also severely reduces the apparent power of the Wald test when the instruments are weak, making its power properties much more like those of the other tests.

The practical conclusions we come to are consistent with those of Andrews, Moreira, and Stock (2006). Two tests seem to be particularly reliable under the null when the instruments are weak. One is the K test when it is bootstrapped using one of our new methods. The other is a conditional bootstrap version of the CLR test that uses one of our new bootstrap methods. Power considerations suggest that the latter test is probably the best procedure overall.

In the next section, we discuss the four test statistics and show that they are all functions of six random quantities. Then, in Section 3, we show how all the statistics can be simulated very efficiently under the assumption of normally distributed disturbances. In Section 4, we consider the asymptotic properties of the statistics under both strong and weak instruments. In Section 5, we discuss some new and old ways of bootstrapping the statistics and show how, in some cases, the properties of bootstrap tests differ greatly from those of asymptotic tests. Finally, in Section 6, we present extensive simulation evidence on the performance of asymptotic and bootstrap tests based on all of the test statistics.

2. The Four Test Statistics

The model treated in this paper consists of just two equations,

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

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

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 instruments such that $\mathcal{S}(\boldsymbol{Z}) \subset \mathcal{S}(\boldsymbol{W})$, where the notation $\mathcal{S}(\boldsymbol{A})$ means the linear span of the columns of the matrix $\boldsymbol{A}$. The disturbances are assumed to be serially uncorrelated and, for many of the analytical results, normally distributed. We assume that $l > k$, so that the model is either exactly identified or, more commonly, overidentified.

The parameters of this model are the scalar β, the k–vector $\boldsymbol{\gamma}$, the l–vector $\boldsymbol{\pi}$, and the 2×2 contemporaneous covariance matrix of the disturbances $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$:

$$\boldsymbol{\Sigma} \equiv \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}. \tag{3}$$

Equation (1) is the structural equation we are interested in, and equation (2) is a reduced-form equation for the second endogenous variable $\boldsymbol{y}_2$. We wish to test the hypothesis that $\beta = 0$. There is no loss of generality in considering only this null hypothesis, since we could test the hypothesis that $\beta = \beta_0$ for any nonzero β_0 by replacing the left-hand side of (1) by $\boldsymbol{y}_1 - \beta_0 \boldsymbol{y}_2$.

Since we are not directly interested in the parameters contained in the l–vector $\boldsymbol{\pi}$, we may without loss of generality suppose that $\boldsymbol{W} = [\boldsymbol{Z} \;\; \boldsymbol{W}_1]$, with $\boldsymbol{Z}^\top \boldsymbol{W}_1 = \boldsymbol{O}$. Notice that $\boldsymbol{W}_1$ can easily be constructed by projecting the columns of $\boldsymbol{W}$ that do not belong to $\mathcal{S}(\boldsymbol{Z})$ off $\boldsymbol{Z}$.

We consider four test statistics: an asymptotic t statistic on which we may base a Wald test, the Anderson-Rubin (AR) statistic, Kleibergen's K statistic, and a likelihood ratio (LR) statistic. The 2SLS (or IV) estimate $\hat{\beta}$ from (1), with instruments the columns of $\boldsymbol{W}$, satisfies the estimating equation

$$\boldsymbol{y}_2^\top \boldsymbol{P}_1 (\boldsymbol{y}_1 - \hat{\beta}\boldsymbol{y}_2) = 0, \tag{4}$$

where $\boldsymbol{P}_1 \equiv \boldsymbol{P}_{\boldsymbol{W}_1}$ is the matrix that projects on to $\mathcal{S}(\boldsymbol{W}_1)$. This follows because $\boldsymbol{Z}^\top \boldsymbol{W}_1 = \boldsymbol{O}$. It is not hard to see that the asymptotic t statistic for a test of the hypothesis that $\beta = 0$ is

$$t = \frac{n^{1/2} \boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_1}{\|\boldsymbol{P}_1 \boldsymbol{y}_2\| \left\| \boldsymbol{M}_{\boldsymbol{Z}} \left(\boldsymbol{y}_1 - \dfrac{\boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_1}{\boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2} \boldsymbol{y}_2 \right) \right\|}. \tag{5}$$

It can be seen that the right-hand side of (5) is homogeneous of degree zero with respect to $\boldsymbol{y}_1$ and also with respect to $\boldsymbol{y}_2$. Consequently, the distribution of the

statistic is invariant to the scales of each of the endogenous variables. In addition, the expression is unchanged if $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are replaced by the projections $\boldsymbol{M_Z y}_1$ and $\boldsymbol{M_Z y}_2$, since $\boldsymbol{P}_1\boldsymbol{M_Z} = \boldsymbol{M_Z}\boldsymbol{P}_1 = \boldsymbol{P}_1$, given the orthogonality of $\boldsymbol{W}_1$ and $\boldsymbol{Z}$. It follows that the statistic (5) depends on the data only through the six quantities

$$\boldsymbol{y}_1^\top\boldsymbol{P}_1\boldsymbol{y}_1, \quad \boldsymbol{y}_1^\top\boldsymbol{P}_1\boldsymbol{y}_2, \quad \boldsymbol{y}_2^\top\boldsymbol{P}_1\boldsymbol{y}_2, \quad \boldsymbol{y}_1^\top\boldsymbol{M_W y}_1, \quad \boldsymbol{y}_1^\top\boldsymbol{M_W y}_2, \text{ and } \boldsymbol{y}_2^\top\boldsymbol{M_W y}_2; \tag{6}$$

notice that $\boldsymbol{y}_i^\top\boldsymbol{M_Z y}_j = \boldsymbol{y}_i^\top(\boldsymbol{M_W} + \boldsymbol{P}_1)\boldsymbol{y}_j$, for $i, j = 1, 2$.

It has been known for some time — see Mariano and Sawa (1972) — that the 2SLS and LIML estimators of β depend only on these six quantities. We can think of them as sufficient statistics. They can easily be calculated by means of four OLS regressions on just two sets of regressors. By regressing $\boldsymbol{y}_i$ on $\boldsymbol{Z}$ and $\boldsymbol{W}$ for $i = 1, 2$, we obtain four sets of residuals. Using the fact that $\boldsymbol{P}_1\boldsymbol{y}_i = (\boldsymbol{M_Z} - \boldsymbol{M_W})\boldsymbol{y}_i$, all six quantities can be obtained as sums of squared residuals, differences of sums of squared residuals, inner products of residual vectors, or inner products of differences of residual vectors.

Another way to test a hypothesis about β is to use the famous test statistic of Anderson and Rubin (1949). The Anderson-Rubin statistic for the hypothesis that $\beta = \beta_0$ can be written as

$$\mathrm{AR}(\beta_0) = \frac{n-l}{l-k}\frac{(\boldsymbol{y}_1 - \beta_0\boldsymbol{y}_2)^\top\boldsymbol{P}_1(\boldsymbol{y}_1 - \beta_0\boldsymbol{y}_2)}{(\boldsymbol{y}_1 - \beta_0\boldsymbol{y}_2)^\top\boldsymbol{M_W}(\boldsymbol{y}_1 - \beta_0\boldsymbol{y}_2)}. \tag{7}$$

Notice that, when $\beta_0 = 0$, the AR statistic depends on the data only through the first and fourth of the six quantities (6). Under the normality assumption, this statistic is exactly distributed as $F(l-k, n-l)$ under the null hypothesis. However, because it has $l-k$ degrees of freedom, it has lower power than statistics with only one degree of freedom when $l - k > 1$.

Kleibergen (2002) therefore proposed a modification of the Anderson-Rubin statistic which has only one degree of freedom. His statistic for testing $\beta = 0$, which can also be interpreted as an LM statistic, is

$$K = (n-l)\,\frac{\boldsymbol{y}_1^\top\boldsymbol{P}_{\boldsymbol{M_Z W}\tilde{\boldsymbol{\pi}}}\boldsymbol{y}_1}{\boldsymbol{y}_1^\top\boldsymbol{M_W y}_1}, \tag{8}$$

which is asymptotically distributed as $\chi^2(1)$ under the null hypothesis that $\beta = 0$. The matrix $\boldsymbol{P}_{\boldsymbol{M_Z W}\tilde{\boldsymbol{\pi}}}$ projects orthogonally on to the one-dimensional subspace generated by the vector $\boldsymbol{M_Z W}\tilde{\boldsymbol{\pi}}$, where $\tilde{\boldsymbol{\pi}}$ is a vector of efficient estimates of the reduced-form parameters. Under our assumptions, the vector $\boldsymbol{M_Z W}\tilde{\boldsymbol{\pi}}$ is equal to the vector $\boldsymbol{W}_1\tilde{\boldsymbol{\pi}}_1$, where $\tilde{\boldsymbol{\pi}}_1$ is the vector of OLS estimates from the artificial regression

$$\boldsymbol{M_Z y}_2 = \boldsymbol{W}_1\boldsymbol{\pi}_1 + \delta\boldsymbol{M_Z y}_1 + \text{residuals}. \tag{9}$$

The estimator $\tilde{\boldsymbol{\pi}}_1$ will be discussed in Section 5 in the context of bootstrapping.

A somewhat lengthy calculation shows that the K statistic (8) is given explicitly by

$$K = \frac{(n-l)(\boldsymbol{y}_1^\top\boldsymbol{P}_1\boldsymbol{y}_2\,\boldsymbol{y}_1^\top\boldsymbol{M_W y}_1 - \boldsymbol{y}_1^\top\boldsymbol{P}_1\boldsymbol{y}_1\,\boldsymbol{y}_1^\top\boldsymbol{M_W y}_2)^2}{\begin{array}{c}\boldsymbol{y}_2^\top\boldsymbol{P}_1\boldsymbol{y}_2\,(\boldsymbol{y}_1^\top\boldsymbol{M_W y}_1)^3 + \boldsymbol{y}_1^\top\boldsymbol{P}_1\boldsymbol{y}_1\,\boldsymbol{y}_1^\top\boldsymbol{M_W y}_1\,(\boldsymbol{y}_1^\top\boldsymbol{M_W y}_2)^2 \\ -2\boldsymbol{y}_1^\top\boldsymbol{P}_1\boldsymbol{y}_2\,\boldsymbol{y}_1^\top\boldsymbol{M_W y}_2\,(\boldsymbol{y}_1^\top\boldsymbol{M_W y}_1)^2\end{array}}. \tag{10}$$

From this, it can be seen that the K statistic, like the t statistic and the AR statistic, depends on the data only through the six quantities (6) and is invariant to the scales of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$.

It is well known that, except for an additive constant, the concentrated loglikelihood function for the model specified by (1), (2), and (3) can be written as

$$-\frac{n}{2}\log\left(1+\frac{l-k}{n-l}\,\mathrm{AR}(\beta)\right), \tag{11}$$

where $\mathrm{AR}(\beta)$ is the Anderson-Rubin statistic (7) evaluated at β. It can then be shown, using results in Anderson and Rubin (1949), that the likelihood ratio statistic for testing the hypothesis that $\beta = 0$ can be written as

$$\mathrm{LR} = n\log(1+SS/n) - n\log\left(1+\frac{SS+TT}{2n}-\frac{1}{2n}\sqrt{(SS-TT)^2+4ST^2}\right), \tag{12}$$

where

$$\begin{aligned}
SS &\equiv n\,\frac{\boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_1}{\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_1}, \\
ST &\equiv \frac{n}{\Delta^{1/2}}\left(\boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_2 - \frac{\boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_1\ \boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2}{\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_1}\right), \\
TT &\equiv \frac{n}{\Delta}\left(\boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2\ \boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_1 - 2\boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_2\ \boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2 + \frac{\boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_1\ (\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2)^2}{\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_1}\right),
\end{aligned} \tag{13}$$

and

$$\Delta \equiv \boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_1\ \boldsymbol{y}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2 - (\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2)^2. \tag{14}$$

The notation is chosen so as to be reminiscent of that used by Moreira (2003) in his discussion of a conditional LR test. His development is different from ours in that he assumes for most of his analysis that the contemporaneous disturbance correlation matrix $\boldsymbol{\Sigma}$ is known. Moreira also introduces a simplified statistic, LR_0, which is obtained by Taylor expanding the logarithms in (12) and discarding terms of order smaller than unity as $n \to \infty$. This procedure yields

$$\mathrm{LR}_0 = \tfrac{1}{2}\left(SS - TT + \sqrt{(SS-TT)^2+4ST^2}\right). \tag{15}$$

We see that both LR and LR_0 are invariant to the scales of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ and depend only on the six quantities (6).

Some tedious algebra shows that the Kleibergen statistic (10) can also be expressed in terms of the quantities ST and TT, as follows:

$$K = \frac{n-l}{n}\frac{ST^2}{TT}. \tag{16}$$

Moreira (2003) demonstrates this relation, although without the degrees-of-freedom adjustment. Finally, it is worth noting that, except for the initial deterministic factors, SS is equal to the Anderson-Rubin statistic AR(0).

3. Simulating the Test Statistics

Now that we have expressions for the test statistics of interest in terms of the six quantities (6), we can explore the properties of these statistics and how to simulate them efficiently. Our results will also be used in the next two sections when we discuss asymptotic and bootstrap tests.

In view of the scale invariance that we have established for all the statistics, the contemporaneous covariance matrix of the disturbances $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ can without loss of generality be set equal to

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \tag{17}$$

with both variances equal to unity. Thus we can represent the disturbances in terms of two independent n–vectors, say $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, of independent standard normal elements, as follows:

$$\boldsymbol{u}_1 = \boldsymbol{v}_1, \quad \boldsymbol{u}_2 = \rho \boldsymbol{v}_1 + r \boldsymbol{v}_2, \tag{18}$$

where $r \equiv (1 - \rho^2)^{1/2}$. We now show that we can write all the test statistics as functions of $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, the exogenous variables, and just three parameters.

With the specification (18), we see from (2) that

$$\begin{aligned} \boldsymbol{y}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2 &= (\rho \boldsymbol{v}_1 + r \boldsymbol{v}_2)^\top \boldsymbol{M}_{\boldsymbol{W}} (\rho \boldsymbol{v}_1 + r \boldsymbol{v}_2) \\ &= \rho^2 \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_1 + r^2 \boldsymbol{v}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_2 + 2\rho r \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_2, \end{aligned} \tag{19}$$

and

$$\begin{aligned} \boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2 = \boldsymbol{\pi}_1^\top \boldsymbol{W}_1^\top \boldsymbol{W}_1 \boldsymbol{\pi}_1 + 2\boldsymbol{\pi}_1^\top \boldsymbol{W}_1^\top (\rho \boldsymbol{v}_1 + r \boldsymbol{v}_2) \\ + \rho^2 \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1 + r^2 \boldsymbol{v}_2^\top \boldsymbol{P}_1 \boldsymbol{v}_2 + 2\rho r \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_2. \end{aligned} \tag{20}$$

Now let $\boldsymbol{W}_1 \boldsymbol{\pi}_1 = a \boldsymbol{w}_1$, with $\|\boldsymbol{w}_1\| = 1$. The square of the parameter a is the so-called scalar concentration parameter; see Phillips (1983, p. 470) and Stock, Wright, and Yogo (2002). Further, let $\boldsymbol{w}_1^\top \boldsymbol{v}_i = x_i$, for $i = 1, 2$. Clearly, x_1 and x_2 are independent standard normal variables. Then

$$\boldsymbol{\pi}_1^\top \boldsymbol{W}_1^\top \boldsymbol{W}_1 \boldsymbol{\pi}_1 = a^2 \text{ and } \boldsymbol{\pi}_1^\top \boldsymbol{W}_1^\top \boldsymbol{v}_i = a x_i, \quad i = 1, 2. \tag{21}$$

Thus (20) becomes

$$\boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2 = a^2 + 2a(\rho x_1 + r x_2) + \rho^2 \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1 + r^2 \boldsymbol{v}_2^\top \boldsymbol{P}_1 \boldsymbol{v}_2 + 2\rho r \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_2. \tag{22}$$

From (1), we find that

$$\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_1 = \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_1 + 2\beta(\rho \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_1 + r \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_2) + \beta^2 \boldsymbol{y}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2. \tag{23}$$

Similarly,

$$\begin{aligned} \boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_1 &= \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1 + 2\beta \boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{v}_1 + \beta^2 \boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2 \\ &= \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1 + 2\beta(a x_1 + \rho \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1 + r \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_2) + \beta^2 \boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2. \end{aligned} \tag{24}$$

Further, from both (1) and (2),

$$\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2 = \rho \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_1 + r \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_2 + \beta \boldsymbol{y}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2, \text{ and} \tag{25}$$

$$\boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_2 = a x_1 + \rho \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1 + r \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_2 + \beta \boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2. \tag{26}$$

The relations (19), (22), (23), (24), (25), and (26) show that the six quantities given in (6) can be generated in terms of eight random variables and three parameters. The eight random variables are x_1 and x_2, along with six quadratic forms of the same sort as those in (6),

$$\boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1, \quad \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_2, \quad \boldsymbol{v}_2^\top \boldsymbol{P}_1 \boldsymbol{v}_2, \quad \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_1, \quad \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_2, \text{ and } \boldsymbol{v}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_2, \tag{27}$$

and the three parameters are a, ρ, and β. Under the null hypothesis, of course, $\beta = 0$. Since $\boldsymbol{P}_1 \boldsymbol{M}_{\boldsymbol{W}} = \mathbf{O}$, the first three variables of (27) are independent of the last three.

If we knew the distributions of the eight random variables on which all the statistics depend, we could simulate them directly. We now characterize these distributions. The symmetric matrix

$$\begin{bmatrix} \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1 & \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_2 \\ \boldsymbol{v}_2^\top \boldsymbol{P}_1 \boldsymbol{v}_1 & \boldsymbol{v}_2^\top \boldsymbol{P}_1 \boldsymbol{v}_2 \end{bmatrix} \tag{28}$$

follows the Wishart distribution $\mathrm{W}(\mathbf{I}_2, l-k)$, and the matrix

$$\begin{bmatrix} \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_1 & \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_2 \\ \boldsymbol{v}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_1 & \boldsymbol{v}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_2 \end{bmatrix}$$

follows the distribution $\mathrm{W}(\mathbf{I}_2, n-l)$. It follows from the analysis of the Wishart distribution in Anderson (1984, Section 7.2) that $\boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_1$ is equal to a random variable t_{11}^M which follows the chi-squared distribution with $n-l$ degrees of freedom, $\boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_2$ is the square root of t_{11}^M multiplied by a standard normal variable z_M independent of it, and $\boldsymbol{v}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_2$ is z_M^2 plus a chi-squared variable t_{22}^M with $n-l-1$ degrees of freedom, independent of z_M and t_{11}^M.

The elements of the matrix (28) can, of course, be characterized in the same way. However, since the elements of the matrix are not independent of x_1 and x_2, it is preferable to define $\boldsymbol{v}_2^\top \boldsymbol{P}_1 \boldsymbol{v}_2$ as x_2^2 plus t_{22}^P, $\boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_2$ as $x_1 x_2$ plus the square root of t_{22}^P times z_P, and $\boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1$ as $x_1^2 + z_P^2 + t_{11}^P$. Here t_{11}^P and t_{22}^P are both chi-squared, with $l-k-2$ and $l-k-1$ degrees of freedom, respectively, and z_P is standard normal. All these variables are mutually independent, and they are also independent of x_1 and x_2. Of course, if $l-k \leq 2$, chi-squared variables with zero or negative degrees of freedom are to be set to zero, and if $l-k=0$, then $z_P = 0$.

An alternative way to simulate the test statistics, which does not require the normality assumption, is to make use of a much simplified model. This model may help to provide an intuitive understanding of the results in the next two sections. The simplified model is

$$\boldsymbol{y}_1 = \beta \boldsymbol{y}_2 + \boldsymbol{u}_1, \tag{29}$$

$$\boldsymbol{y}_2 = a \boldsymbol{w}_1 + \boldsymbol{u}_2, \tag{30}$$

where the disturbances are generated according to (18). Here the n–vector $\boldsymbol{w}_1 \in \mathcal{S}(\boldsymbol{W})$ with $\|\boldsymbol{w}_1\| = 1$, where $\boldsymbol{W}$, as before, is an $n \times l$ matrix of instruments. By normalizing $\boldsymbol{w}_1$ in this way, we are implicitly using weak-instrument asymptotics; see Staiger and Stock (1997). Clearly, we may choose $a \geq 0$. The DGPs of this simple model, which are completely characterized by the parameters β, ρ, and a, can generate the six quantities (6) so as to have the same distributions as those generated by any DGP of the more complete model specified by (1), (2), and (3).

If the disturbances are not Gaussian, the distributions of the statistics depend not only on the parameters a, ρ, and β but also on the vector $\boldsymbol{w}_1$ and the linear span of the instruments. We may suspect, however, that this dependence is weak, and limited simulation evidence (not reported) strongly suggests that this is indeed the case. The distribution of the disturbances seems to have a much greater effect on the distributions of the test statistics than the features of $\boldsymbol{W}$.

4. Asymptotic Theory

To fix ideas, we begin with a short discussion of the conventional asymptotic theory of the tests discussed in Section 2. By "conventional", we mean that the instruments are assumed to be strong, in a sense made explicit below. Under this assumption, the tests are all classical. In particular, Kleibergen (2002) shows that the K statistic is a version of the Lagrange Multiplier test.

The reduced-form equation (30) of the simplified model of the previous section is written in terms of an instrumental variable $\boldsymbol{w}_1$ such that $\|\boldsymbol{w}_1\| = 1$. Conventional asymptotics would set $\|\boldsymbol{w}_1\|^2 = O_p(n)$ and let the parameter a be independent of the sample size. Our setup is better suited to the weak-instrument asymptotics of Staiger and Stock (1997). For conventional asymptotics, we may suppose that $a = n^{1/2}\alpha$, for α constant as the sample size $n \to \infty$.

Under the null, $\beta = 0$. Under local alternatives, we let $\beta = n^{-1/2}b$, for b constant as $n \to \infty$. Conventional asymptotics applied to (19), (22), (23), (24), (25), and (26) then give

$$\begin{aligned} \boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_1 &\stackrel{a}{=} (x_1 + \alpha b)^2 + t_{11}^P, & n^{-1}\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_1 &\stackrel{a}{=} 1, \\ n^{-1/2}\boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_2 &\stackrel{a}{=} \alpha x_1 + \alpha^2 b, & n^{-1}\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2 &\stackrel{a}{=} \rho, \\ n^{-1}\boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2 &\stackrel{a}{=} \alpha^2, & n^{-1}\boldsymbol{y}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2 &\stackrel{a}{=} 1. \end{aligned} \tag{31}$$

Using these results, it is easy to check that, for $\beta = 0$, the statistics t^2, K, and LR, given by (5) squared, (10), and (12), respectively, are all equal to $(x_1 + \alpha b)^2$ asymptotically. They have a common asymptotic distribution of χ^2 with one degree of freedom and noncentrality parameter $\alpha^2 b^2 = a^2\beta^2$. We can also see that the Anderson-Rubin statistic AR(0), as given by (7), is asymptotically equal to $(x_1 + \alpha b)^2 + z_P^2 + t_{11}^P$, with $l - k$ degrees of freedom and the same noncentrality parameter. Thus AR(0) is asymptotically equal to the same noncentral $\chi^2(1)$ random variable as the other three statistics, plus an independent central $\chi^2(l - k - 1)$ random variable.

We now turn to the more interesting case of weak-instrument asymptotics, for which a is kept constant as $n \to \infty$. The three right-hand results of (31) are unchanged, but the left-hand ones have to be replaced by the following equations, which involve no asymptotic approximation, but hold even in finite samples:

$$\begin{aligned}
\boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_1 &= x_1^2 + z_P^2 + t_{11}^P, \\
\boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_2 &= a x_1 + \rho\big(x_1^2 + z_P^2 + t_{11}^P\big) + r\Big(x_1 x_2 + z_P \sqrt{t_{22}^P}\Big), \text{ and} \\
\boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2 &= a^2 + 2a(\rho x_1 + r x_2) + \rho^2\big(x_1^2 + z_P^2 + t_{11}^P\big) \\
&\qquad + r^2(x_2^2 + t_{22}^P) + 2\rho r\Big(x_1 x_2 + z_P \sqrt{t_{22}^P}\Big).
\end{aligned} \tag{32}$$

Since the Anderson-Rubin statistic AR(0) is exactly pivotal for the model we are studying, its distribution under the null that $\beta = 0$ depends neither on a nor on ρ. Since the quantity SS in (13) is equal to AR(0) except for degrees-of-freedom factors, it too is exactly pivotal. Its asymptotic distribution under weak-instrument asymptotics is that of $\boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_1$. Thus, as we see from the first line of (32),

$$SS \stackrel{a}{=} x_1^2 + z_P^2 + t_{11}^P, \tag{33}$$

which follows the central $\chi^2(l-k)$ distribution.

Although Kleibergen's K statistic is not exactly pivotal, it is asymptotically pivotal under both weak-instrument and strong-instrument asymptotics. From (10), and using (31) and (32), we can see, after some algebra, that, under weak-instrument asymptotics and under the null,

$$\begin{aligned}
K &\stackrel{a}{=} \frac{(\boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_2 - \rho \boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_1)^2}{\boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2 - 2\rho \boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_2 + \rho^2 \boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_1} \\
&= \frac{\big(a x_1 + r(x_1 x_2 + z_P \sqrt{t_{22}^P})\big)^2}{a^2 + 2arx_2 + r^2(x_2^2 + t_{22}^P)} = \frac{\big(x_1(a + r x_2) + r z_P \sqrt{t_{22}^P}\big)^2}{(a + r x_2)^2 + r^2 t_{22}^P}.
\end{aligned} \tag{34}$$

Although the last expression above depends on a and ρ, it is in fact just a chi-squared variable with one degree of freedom. To see this, argue conditionally on all random variables except x_1 and z_P, recalling that all the random variables in the expression are mutually independent. The numerator is the square of a linear combination of the standard normal variables x_1 and z_P, and the denominator is the conditional variance of this linear combination. Thus the conditional asymptotic distribution of K is χ_1^2, and so also its unconditional distribution. As Kleibergen (2002) remarks, this implies that K is asymptotically pivotal in all configurations of the instruments, including that in which $a = 0$ and the instruments are completely invalid.

For the LR statistic, we can write down expressions asymptotically equal to the quantities ST and TT in (13). First, from (14), we have

$$\Delta / n^2 \stackrel{a}{=} 1 - \rho^2.$$

It is then straightforward to check that

$$ST \stackrel{a}{=} \frac{1}{(1-\rho^2)^{1/2}}\Big(ax_1 + r\big(x_1x_2 + z_p\sqrt{t_{22}^P}\,\big)\Big),$$

and

$$TT \stackrel{a}{=} \frac{1}{1-\rho^2}\big(a^2 + 2arx_2 + r^2(x_2^2 + t_{22}^P)\big). \tag{35}$$

Comparison with (34) then shows that, in accordance with (16), $K \stackrel{a}{=} ST^2/TT$.

It is clear from (35) and (33) that SS and TT are asymptotically independent, since the former depends only on the random variables x_1, z_P, and t_{11}^P, while the latter depends only on x_2 and t_{22}^P. The discussion based on (34) shows that, conditional on TT, ST is distributed as $\sqrt{TT}$ times a standard normal variable, and that K is asymptotically distributed as χ_1^2.

Even though SS and ST are not conditionally independent, the variables ST and $SS - ST^2/TT$ are so asymptotically. This follows because, conditionally on x_2 and t_{22}^P, the normally distributed variable $x_1(a + rx_2) + rz_P\sqrt{t_{22}^P}$ is a linear combination of the standard normal variables x_1 and z_P that partially constitute the asymptotically chi-squared variable SS. These properties led Moreira (2003) to suggest that the distribution of the statistics LR and LR_0, which are deterministic functions of SS, ST, and TT, conditional on a given value of TT, say tt, can be estimated by a simulation experiment in which ST and $SS - ST^2/TT$ are generated as independent variables distributed respectively as $\mathrm{N}(0, tt)$ and χ^2_{l-k-1}. The variable SS is then generated by combining these two variables, replacing TT by tt. Such an experiment has a bootstrap interpretation that we develop in the next section.

It may be helpful to make explicit the link between the quantities SS, ST, and TT, defined in (13), and the vectors $\boldsymbol{S}$ and $\boldsymbol{T}$ used in Andrews, Moreira, and Stock (2006). For the simplified model given by (29) and (30), these vectors can be expressed as

$$\boldsymbol{S} = \boldsymbol{W}^\top \boldsymbol{v}_1 \text{ and } \boldsymbol{T} = \frac{1}{r}\boldsymbol{W}^\top(\boldsymbol{y}_2 - \rho\boldsymbol{y}_1).$$

It is straightforward to check that, with these definitions, $\boldsymbol{S}^\top\boldsymbol{S}$, $\boldsymbol{S}^\top\boldsymbol{T}$, and $\boldsymbol{T}^\top\boldsymbol{T}$ are what the expressions SS, ST, and TT would become if the three quadratic forms in the right panel of (31) were replaced by their asymptotic limits.

It should also be noted that all of the weak-instrument asymptotic results continue to hold with non-Gaussian disturbances under a very few additional regularity conditions. Firstly, the disturbances must have moments of order at least 2. Secondly, the instrument matrix $\boldsymbol{W}$ must be such as to allow us to apply a law of large numbers to obtain the right-hand panel of (31) and to apply a central limit theorem to show that the random variables in (32) are asymptotically standard normal or chi-squared, as required. For this, it is enough to be able to apply a central limit theorem to the vectors $n^{-1/2}\boldsymbol{W}_1^\top\boldsymbol{v}_i$, $i = 1, 2$.

The results (32) are independent of b. This shows that, for local alternatives of the sort used in conventional asymptotic theory, no test statistic that depends only on

the six quantities (6) can have asymptotic power greater than asymptotic size under weak-instrument asymptotics. However, if instead we consider fixed alternatives, with parameter β independent of n, then the expressions do depend on β.

For notational ease, denote the three right-hand sides in (32) by Y_{11}, Y_{12}, and Y_{22}, respectively. Then it can be seen that the weak-instrument results become

$$\begin{aligned} \boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_1 &= Y_{11} + 2\beta Y_{12} + \beta^2 Y_{22}, & n^{-1} \boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_1 &\stackrel{a}{=} 1 + 2\beta\rho + \beta^2, \\ \boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_2 &= Y_{12} + \beta Y_{22}, & n^{-1} \boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2 &\stackrel{a}{=} \rho + \beta, \\ \boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2 &= Y_{22}, & n^{-1} \boldsymbol{y}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2 &\stackrel{a}{=} 1. \end{aligned} \tag{36}$$

Notice that, if we specialize the above results, letting β be $O(n^{-1/2})$ and letting a be $O(n^{1/2})$, then we obtain the conventional strong-instrument results (31).

We have not written down the weak-instrument asymptotic expression for the Wald t statistic given in (5), because it is complicated and not very illuminating. Suffice it to say that it depends nontrivially on the parameters a and ρ, as does its distribution. Consequently, the statistic t is not asymptotically pivotal. Indeed, in the terminology of Dufour (1997), it is not even boundedly pivotal, by which we mean that rejection probabilities of tests based on it cannot be bounded away from one. We will see this explicitly in a moment.

The estimating equations (4) imply that the IV estimate of β is $\hat{\beta} = \boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_2 / \boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2$. Under weak-instrument asymptotics, we see from (36) that

$$\hat{\beta} \stackrel{a}{=} \beta + Y_{12}/Y_{22}. \tag{37}$$

Since $\mathrm{E}(Y_{12}/Y_{22}) \neq 0$, it follows that $\hat{\beta}$ is biased and inconsistent. The square of the t statistic (5) can be seen to be asymptotically equal to

$$\frac{Y_{22}(Y_{12} + \beta Y_{22})^2}{Y_{22}^2 - 2\rho Y_{12} Y_{22} + Y_{12}^2} \tag{38}$$

under weak-instrument asymptotics. Observe that this expression is of the order of β^2 as $\beta \to \infty$. Thus, for fixed a and ρ, the distributions of t^2 for $\beta = 0$ and $\beta \neq 0$ can be arbitrarily far apart.

For $\beta = 0$, however, the distribution of t^2, for a and ρ sufficiently close to 0 and 1, respectively, is also arbitrarily far from that with fixed $a \neq 0$ and $\rho \neq 1$. It is this fact that leads to the failure of t^2 to be boundedly pivotal. Let a and $r = (1 - \rho^2)^{1/2}$ be treated as small quantities, and then expand the denominator of expression (38) through the second order in small quantities. Note that, to this order, $\rho = 1 - r^2/2$. Then we see that, to the desired order,

$$\begin{aligned} Y_{12} &= \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1 + (a x_1 + r \boldsymbol{v}_1^\top \boldsymbol{P}_2 \boldsymbol{v}_2) - \tfrac{1}{2} r^2 \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1, \text{ and} \\ Y_{22} &= \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1 + 2(a x_1 + r \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_2) + (a^2 + 2arx_2 + r^2 \boldsymbol{v}_2^\top \boldsymbol{P}_1 \boldsymbol{v}_2 - \tfrac{1}{2} r^2 \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1). \end{aligned}$$

Consequently, to the order at which we are working,

$$Y_{22}^2 - 2\rho Y_{12}Y_{22} + Y_{12}^2 = r^2(\boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1)^2.$$

To leading order, the numerator of (38) for $\beta = 0$ is just $(\boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1)^3$, and so, in the neighborhood of $a = 0$ and $r = 0$, we have from (38) that

$$t^2 \stackrel{a}{=} \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1 / r^2. \tag{39}$$

The numerator here is just a chi-squared variable, but the denominator can be arbitrarily close to zero. Thus the distribution of t^2 can be moved arbitrarily far away from any finite distribution by letting a tend to zero and ρ tend to one.

The points in the parameter space at which $a = 0$ and $\rho = \pm 1$, which implies that $r = 0$, are points at which β is completely unidentified. To see this, consider the DGP from model (29) and (30) that corresponds to these parameter values. The DGP can be written as

$$\boldsymbol{y}_2 = \boldsymbol{v}_1, \quad \boldsymbol{y}_1 = (1 \pm \beta)\boldsymbol{y}_2. \tag{40}$$

It follows from (37) that $\hat{\beta} = 1 \pm \beta$. This is not surprising, since the second equation in (40) fits perfectly. This fact then accounts for the t statistic tending to infinity.

All the other tests have power that does not tend to 1 when $\beta \to \infty$ under weak-instrument asymptotics. For K, some algebra shows that

$$K \stackrel{a}{=} \frac{\big(Y_{12} - \rho Y_{11} + \beta(Y_{22} - Y_{11}) + \beta^2(\rho Y_{22} - Y_{12})\big)^2}{D(1 + 2\beta\rho + \beta^2)}, \tag{41}$$

where

$$D \equiv Y_{22} - 2\rho Y_{12} + \rho^2 Y_{11} + 2\beta\big(\rho Y_{22} - (1+\rho^2)Y_{12} + \rho Y_{11}\big) + \beta^2(Y_{11} - 2\rho Y_{12} + \rho^2 Y_{22}). \tag{42}$$

As $\beta \to \infty$, the complicated expression (41) tends to the much simpler limit of

$$\frac{(\rho Y_{22} - Y_{12})^2}{Y_{11} - 2\rho Y_{12} + \rho^2 Y_{22}}. \tag{43}$$

Thus, unlike t^2, the K statistic does not become unbounded as $\beta \to \infty$. Consequently, under weak-instrument asymptotics, the test based on K is inconsistent for any nonzero β, in the sense that the rejection probability does not tend to 1 however large the sample size.

A similar result holds for the AR test. It is easy to see that

$$SS \stackrel{a}{=} \frac{Y_{11} + 2\beta Y_{12} + \beta^2 Y_{22}}{1 + 2\beta\rho + \beta^2} \underset{\beta \to \infty}{\longrightarrow} Y_{22}, \tag{44}$$

which does not depend on β. Thus, since AR is proportional to SS, we see that the asymptotic distribution of AR does not depend on β under weak-instrument asymptotics.

Similar results also hold for the LR and LR_0 statistics. By an analysis like the one that produced (44), we have that

$$ST \stackrel{a}{=} \frac{Y_{12} - \rho Y_{11} + \beta(Y_{22} - Y_{11}) + \beta^2(\rho Y_{22} - Y_{12})}{r(1 + 2\beta\rho + \beta^2)} \underset{\beta\to\infty}{\longrightarrow} \frac{\rho Y_{22} - Y_{12}}{r}, \text{ and}$$

$$TT \stackrel{a}{=} \frac{D}{r^2(1 + 2\beta\rho + \beta^2)} \underset{\beta\to\infty}{\longrightarrow} \frac{Y_{11} - 2\rho Y_{12} + \rho^2 Y_{22}}{r^2},$$

where D is given by (42). From (15), therefore,

$$\mathrm{LR}_0 \underset{\beta\to\infty}{\longrightarrow} \frac{1}{2r^2}\Big(Y_{22}(1 - 2\rho^2) + 2\rho Y_{12} - Y_{11} + \\ \big(Y_{11}^2 - 2Y_{11}Y_{22}(1 - 2\rho^2) - 4\rho Y_{11}Y_{12} + Y_{22}^2 + 4Y_{12}^2 - 4\rho Y_{12}Y_{22}\big)^{1/2}\Big).$$

The inconsistency of the LR_0 test follows from the fact that this random variable has a bounded distribution. This is true for the LR test as well, but we will spare readers the details.

We saw above that, when $a = 0$ and $\rho = 1$, the parameter β is unidentified. We expect, therefore, that a test statistic for the hypothesis that $\beta = 0$ would have the same distribution whatever the value of β. This turns out to be the case for the K statistic. If one computes the limit of expression (41) for $a = 0$, $r \to 0$, the limiting expression is just $(\boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_2)^2 / \boldsymbol{v}_2^\top \boldsymbol{P}_1 \boldsymbol{v}_2$, independently of the value of β. Presumably a more complicated calculation would show that the same is true for LR and LR_0.

The result that the AR, K, and LR tests are inconsistent under weak-instrument asymptotics appears to contradict some of the principal results of Andrews, Moreira, and Stock (2006). The reason for this apparent contradiction is that we have made a different, and in our view more reasonable, assumption about the covariance matrix of the disturbances. We assume that the matrix $\boldsymbol{\Sigma}$, defined in (3) as the covariance matrix of the disturbances in the structural equation (1) and the reduced form equation (2), remains constant as β varies. In contrast, Andrews, Moreira, and Stock (2006) assumes that the covariance matrix of the reduced form disturbances does so.

In terms of our parametrization, the covariance matrix of the disturbances in the two reduced form equations is

$$\begin{bmatrix} \sigma_1^2 + 2\rho\beta\sigma_1\sigma_2 + \beta^2\sigma_2^2 & \rho\sigma_1\sigma_2 + \beta\sigma_2^2 \\ \rho\sigma_1\sigma_2 + \beta\sigma_2^2 & \sigma_2^2 \end{bmatrix}. \tag{45}$$

This expression depends on β in a nontrivial way. In order for it to remain constant as β changes, both ρ and σ_1 must be allowed to vary. Thus the assumption that (45) is fixed, which was made by Andrews, Moreira, and Stock (2006), implies that (3) cannot remain constant as $|\beta| \to \infty$.

A little algebra shows that, as $\beta \to \pm\infty$ with the covariance matrix (45) held fixed, the parameters β and ρ of the observationally equivalent DGP of the model given by

(29) and (30) along with (18) tend to ± 1 and ∓ 1 respectively. Thus the omnipresent denominator $1 + 2\beta\rho + \beta^2$ tends to zero in either of these limits. But it is clear from (36) that this means that the estimate of σ_1^2 from the full model (1) and (2) tends to zero.

Based on the above remark, our view is that it is more reasonable that (3) should remain constant than that (45) should. The parameter ρ is a much more interesting parameter than the correlation in (45). Even when $\rho = 0$, in which case the OLS estimator of β is consistent, the correlation between the two reduced form disturbances tends to ± 1 as $\beta \to \pm\infty$. Thus we believe that the latter correlation is not a sensible quantity to hold fixed. In any case, it is rather disturbing that something as seemingly innocuous as the parametrization of the covariance matrix of the disturbances can have profound consequences for the analysis of power when the instruments are weak.

5. Bootstrapping the Test Statistics

There are several ways to bootstrap the non-exact test statistics that we have been discussing (Wald, K, and LR). In this section, we discuss five different parametric bootstrap procedures, three of which are new. We obtain a number of interesting theoretical results. We also discuss the pairs bootstrap, show how to convert the new procedures into semiparametric bootstraps, and propose a new, semiparametric, conditional bootstrap LR test. In the next section, we will see that two of the new procedures, and one of them in particular, perform extremely well.

For all the statistics, we perform B bootstrap simulations and calculate the bootstrap P value as

$$\hat{p}^*(\hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} I(\tau_j^* > \hat{\tau}), \tag{46}$$

where $\hat{\tau}$ denotes the actual test statistic, and τ_j^* denotes the statistic calculated using the j^{th} bootstrap sample.

Dufour (1997) makes it clear that bootstrapping is not in general a cure for the difficulties associated with the Wald statistic t. However, since the Wald statistic is still frequently used in practice when there is no danger of weak instruments, it is interesting to look at the performance of the bootstrapped Wald test when instruments are strong. When they are weak, we confirm Dufour's result about the ineffectiveness of bootstrapping. In the context of our new bootstrap methods, this manifests itself in an almost complete loss of power, for reasons that we analyze.

Since the K statistic is asymptotically pivotal under weak-instrument asymptotics, it should respond well to bootstrapping, at least under the null. The LR statistic is not asymptotically pivotal, but, as shown by Moreira (2003), a conditional LR test gives the asymptotically pivotal statistic we call CLR. As we explain below, the implementation of this conditional likelihood ratio test is in fact a form of bootstrapping. Thus it is computationally quite intensive to bootstrap the conditional LR test, since doing so involves a sort of double bootstrap. As an alternative, we propose below a new conditional bootstrap version of the CLR test.

Since we have assumed up to now that the disturbances of our model are Gaussian, it is appropriate to use a parametric bootstrap in which the disturbances are normally distributed. In practice, however, investigators will often be reluctant to make this assumption. At the end of this section, we therefore discuss semiparametric versions of our new bootstrap techniques that resample the residuals. We also discuss the pairs bootstrap that was proposed by Freedman (1984) and has been used by Moreira, Porter, and Suarez (2005).

For any bootstrapping procedure, the first task, and usually the most important one, is to choose a suitable bootstrap DGP; see Davidson and MacKinnon (2006a). An obvious but important point is that the bootstrap DGP must be able to handle both of the endogenous variables, that is, $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. A straightforward, conventional approach is to estimate the parameters β, $\boldsymbol{\gamma}$, $\boldsymbol{\pi}$, σ_1, σ_2, and ρ of the model specified by (1), (2), and (3) and then to generate simulated data using these equations with the estimated parameters.

However, the conventional approach estimates more parameters than it needs to. The bootstrap DGP should take advantage of the fact that the simple model specified by (29) and (30) can generate statistics with the same distributions as those generated by the full model. Equation (29) becomes especially simple when the null hypothesis is imposed: It says simply that $\boldsymbol{y}_1 = \boldsymbol{u}_1$. If this approach is used, then only the parameters a and ρ need to be estimated. In order to estimate a, we may substitute an estimate of $\boldsymbol{\pi}$ into the definition (21) with an appropriate scaling factor to take account of the fact that a is defined for DGPs with unit disturbance variances.

We investigate five different ways of estimating the parameters ρ and a. To estimate ρ, we just need residuals from equations (29) and (30), or, in the general case, (1) and (2). To estimate a, we need estimates of the vector $\boldsymbol{\pi}_1$ from the reduced-form equation (30), or from (2), along with residuals from that equation. If $\ddot{\boldsymbol{u}}_1$ and $\ddot{\boldsymbol{u}}_2$ denote the two residual vectors, and $\ddot{\boldsymbol{\pi}}_1$ denotes the estimate of $\boldsymbol{\pi}_1$, then our estimates are

$$\ddot{\rho} = \frac{\ddot{\boldsymbol{u}}_1^\top \ddot{\boldsymbol{u}}_2}{(\ddot{\boldsymbol{u}}_1^\top \ddot{\boldsymbol{u}}_1 \ddot{\boldsymbol{u}}_2^\top \ddot{\boldsymbol{u}}_2)^{1/2}}, \text{ and} \tag{47}$$

$$\ddot{a} = \sqrt{n\, \ddot{\boldsymbol{\pi}}_1^\top \boldsymbol{W}_1^\top \boldsymbol{W}_1 \ddot{\boldsymbol{\pi}}_1 / \ddot{\boldsymbol{u}}_2^\top \ddot{\boldsymbol{u}}_2}. \tag{48}$$

Existing methods, and the new ones that we propose, use various estimates of $\boldsymbol{\pi}_1$ and various residual vectors.

The simplest way to estimate ρ and a is probably to use the restricted residuals

$$\tilde{\boldsymbol{u}}_1 = \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y}_1 = \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_1 + \boldsymbol{P}_1 \boldsymbol{y}_1,$$

which, in the case of the simple model, are just equal to $\boldsymbol{y}_1$, along with the OLS estimates $\hat{\boldsymbol{\pi}}_1$ and OLS residuals $\hat{\boldsymbol{u}}_2$ from the FWL regression

$$\boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y}_2 = \boldsymbol{W}_1 \boldsymbol{\pi}_1 + \boldsymbol{u}_2. \tag{49}$$

We call this widely-used method the RI bootstrap, for "Restricted, Inefficient". It can be expected to work better than the pairs bootstrap, and better than other parametric procedures that do not impose the null hypothesis.

As the name implies, the problem with the RI bootstrap is that $\hat{\boldsymbol{\pi}}_1$ is not an efficient estimator. That is why Kleibergen (2002) did not use $\hat{\boldsymbol{\pi}}_1$ in constructing the K statistic. Instead, the estimates $\tilde{\boldsymbol{\pi}}_1$ from equation (9) were used. It can be shown that these estimates are asymptotically equivalent to the ones that would be obtained by using 3SLS or FIML on the system consisting of equations (1) and (2). The estimated vector of disturbances from equation (9) is not the vector of OLS residuals but rather the vector $\tilde{\boldsymbol{u}}_2 = \boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{y}_2 - \boldsymbol{W}_1\tilde{\boldsymbol{\pi}}$.

Instead of equation (9), it may be more convenient to run the regression

$$\boldsymbol{y}_2 = \boldsymbol{W}_1\boldsymbol{\pi}_1 + \boldsymbol{Z}\boldsymbol{\pi}_2 + \delta\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{y}_1 + \text{ residuals.} \tag{50}$$

This is just the reduced form equation augmented by the residuals from restricted estimation of the structural equation. Because of the orthogonality between $\boldsymbol{Z}$ and $\boldsymbol{W}_1$, the vector $\tilde{\boldsymbol{u}}_2$ is equal to the vector of OLS residuals from regression (50) plus $\hat{\delta}\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{y}_1$. We call the bootstrap that uses $\tilde{\boldsymbol{u}}_1$, $\tilde{\boldsymbol{\pi}}_1$, and $\tilde{\boldsymbol{u}}_2$ the RE bootstrap, for "Restricted, Efficient".

Two other bootstrap methods do not impose the restriction that $\beta = 0$ when estimating ρ and a. For the purposes of testing, it is a bad idea not to impose this restriction, as we argued in Davidson and MacKinnon (1999). However, it is quite inconvenient to impose restrictions when constructing bootstrap confidence intervals, and, since confidence intervals are implicitly obtained by inverting tests, it is of interest to see how much harm is done by not imposing the restriction.

The UI bootstrap, for "Unrestricted, Inefficient", uses the unrestricted residuals $\hat{\boldsymbol{u}}_1$ from IV estimation of (1), along with the estimates $\hat{\boldsymbol{\pi}}_1$ and residuals $\hat{\boldsymbol{u}}_2$ from OLS estimation of (2). The UE bootstrap, for "Unrestricted, Efficient", also uses $\hat{\boldsymbol{u}}_1$, but the other quantities come from the artificial regression

$$\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{y}_2 = \boldsymbol{W}_1\boldsymbol{\pi}_1 + \delta\hat{\boldsymbol{u}}_1 + \text{ residuals,} \tag{51}$$

which is similar to regression (9). Of course, a regression analogous to (50) could be used instead of (51). A fifth bootstrap method will be proposed after we have obtained some results on which it depends.

It is possible to write the estimates of a and ρ used by all four of these bootstrap schemes as functions solely of the six quantities (6). This makes it possible to program the bootstrap very efficiently. Because many of the functions are quite complicated, we will spare readers most of the details. However, we need the following results for the RE bootstrap:

$$\tilde{\rho} = \frac{\boldsymbol{y}_1^\top\boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{y}_2 + \dfrac{\boldsymbol{y}_1^\top\boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{y}_2}{\boldsymbol{y}_1^\top\boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{y}_1}\boldsymbol{y}_1^\top\boldsymbol{P}_1\boldsymbol{y}_1}{\left(\left(\boldsymbol{y}_1^\top\boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{y}_1 + \boldsymbol{y}_1^\top\boldsymbol{P}_1\boldsymbol{y}_1\right)\left(\boldsymbol{y}_2^\top\boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{y}_2 + \left(\dfrac{\boldsymbol{y}_1^\top\boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{y}_2}{\boldsymbol{y}_1^\top\boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{y}_1}\right)^2\boldsymbol{y}_1^\top\boldsymbol{P}_1\boldsymbol{y}_1\right)\right)^{1/2}}, \tag{52}$$

and

$$\tilde{a}^2 = \frac{\boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2 - 2\boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_2 \dfrac{\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2}{\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_1} + \boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_1 \left(\dfrac{\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2}{\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_1} \right)^2}{\boldsymbol{y}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2 + \boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_1 \left(\dfrac{\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2}{\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_1} \right)^2}. \tag{53}$$

Davidson and MacKinnon (1999) show that the size distortion of bootstrap tests may be reduced by use of a bootstrap DGP that is asymptotically independent of the statistic that is bootstrapped. In general, this is true only for bootstrap DGPs that are based on efficient estimators. Thus it makes sense to use the efficient estimator $\tilde{\boldsymbol{\pi}}_1$ rather than the inefficient estimator $\hat{\boldsymbol{\pi}}_1$ in order to estimate a, and, via the reduced-form residuals, ρ. Either restricted or unrestricted residuals from (1) can be used as the extra regressor in estimating $\boldsymbol{\pi}_1$ without interfering with the desired asymptotic independence, but general considerations of efficiency suggest that restricted residuals are the better choice. Thus we would expect that, when conventional asymptotics yield a good approximation, the best choice for bootstrap DGP is RE.

Under weak-instrument asymptotics, things are rather different. We use the results of (32) and the right-hand results of (31) to see that, with data generated by the model (29) and (30) under the null hypothesis,

$$\tilde{\sigma}_1^2 \stackrel{a}{=} 1, \quad \tilde{\sigma}_2^2 \stackrel{a}{=} 1, \quad \text{and} \quad \tilde{\rho}\tilde{\sigma}_1\tilde{\sigma}_2 \stackrel{a}{=} \rho.$$

Thus the RE bootstrap estimator $\tilde{\rho}$, as defined by (52), is a consistent estimator, as is the estimator used by the RI bootstrap. It can be checked that this result does *not* hold for any of the estimators that use unrestricted residuals from equation (29), since they depend on the inconsistent IV estimate of β; recall (37).

The weak-instrument asymptotic version of (53) under the null can be seen to be

$$\tilde{a}^2 \stackrel{a}{=} a^2 + 2arx_2 + r^2(x_2^2 + t_{22}^P). \tag{54}$$

Unless $r = 0$, then, $\tilde{a}^2$ is inconsistent. It is also biased, the bias being equal to $r^2(l-k)$. It seems plausible, therefore, that the bias-corrected estimator

$$\tilde{a}^2_{\text{BC}} \equiv \max\bigl(0, \tilde{a}^2 - (l-k)(1-\tilde{\rho}^2)\bigr) \tag{55}$$

may be better for the purposes of defining the bootstrap DGP. Thus we consider a fifth bootstrap method, REC, for "Restricted, Efficient, Corrected." It differs from RE in that it uses $\tilde{a}_{\text{BC}}$ instead of $\tilde{a}$. This has the effect of reducing the R^2 of the reduced-form equation in the bootstrap DGP.

For the purposes of an analysis of power, it is necessary to look at the properties of the estimates $\tilde{\rho}$ and $\tilde{a}^2$ under the alternative, that is, for nonzero β. From (36), we see that

$$\tilde{\sigma}_1^2 \stackrel{a}{=} 1 + 2\beta\rho + \beta^2, \quad \tilde{\sigma}_2^2 \stackrel{a}{=} 1, \quad \text{and} \quad \tilde{\rho}\tilde{\sigma}_1\tilde{\sigma}_2 \stackrel{a}{=} \rho + \beta,$$

from which we find that

$$\tilde{\rho} \stackrel{a}{=} \frac{\rho + \beta}{(1 + 2\beta\rho + \beta^2)^{1/2}}.$$

As $\beta \to \infty$, then, we see that $\tilde{\rho} \to 1$, for all values of a and ρ. For the rate of convergence, it is better to reason in terms of the parameter $r \equiv (1 - \rho^2)^{1/2}$. We have

$$\tilde{r}^2 = 1 - \tilde{\rho}^2 \stackrel{a}{=} 1 - \frac{(\rho + \beta)^2}{1 + 2\beta\rho + \beta^2} = \frac{r^2}{1 + 2\beta\rho + \beta^2}.$$

Thus $\tilde{r} = O_p(\beta^{-1})$ as $\beta \to \infty$.

The calculation for $\tilde{a}^2$ is a little more involved. From (53) and (36), we find that

$$\begin{aligned}\tilde{a}^2 &\stackrel{a}{=} Y_{22} - \frac{2(\rho + \beta)}{1 + 2\beta\rho + \beta^2}(Y_{12} + \beta Y_{22}) + \left(\frac{\rho + \beta}{1 + 2\beta\rho + \beta^2}\right)^2 (Y_{11} + 2\beta Y_{12} + \beta^2 Y_{22}) \\ &\stackrel{a}{=} \frac{1}{(1 + 2\beta\rho + \beta^2)^2}\big((1 + \beta\rho)Y_{22} - 2(\rho + \beta)(1 + \beta\rho)Y_{12} + (\rho + \beta)^2 Y_{11}\big).\end{aligned}$$

Clearly, this expression is of the order of β^{-2} in probability as $\beta \to \infty$, so that $\tilde{a} \to 0$, again for all a and ρ. In fact, it is clear that $\tilde{a} = O_p(\beta^{-1})$ as $\beta \to \infty$, from which we conclude that $\tilde{a}$ and $\tilde{r}$ tend to zero at the same rate as $\beta \to \infty$, as in the calculation that led to (39).

These results can be understood intuitively by considering (29) and (30). Estimation of ρ uses residuals which for that model are just the vector $\boldsymbol{y}_1 = \beta\boldsymbol{y}_2 + \boldsymbol{v}_1$. For large β, this residual vector is almost collinear with $\boldsymbol{y}_2$, and so also with the residual vector $\tilde{\boldsymbol{u}}_2$. The estimated correlation coefficient therefore tends to 1. Similarly, when $\boldsymbol{y}_1$ is introduced as an extra regressor for the estimation of a, it is highly collinear with the dependent variable and explains almost all of it, leaving no apparent explanatory power for the weak instruments.

For large β, then, the RE bootstrap DGP is characterized by parameters a and ρ close to 0 and 1, respectively. As we saw near the end of the last section, at this point in the parameter space, β is unidentified, and the Wald statistic has an unbounded distribution. These facts need not be worrisome for the bootstrapping of statistics that are asymptotically pivotal with weak instruments, but they mean that the bootstrap version of the Wald test, like the Kleibergen and LR tests, is inconsistent, having a probability of rejecting the null hypothesis that does not tend to one as $\beta \to \infty$.

To see this, we make use of expression (39) to see that the distribution of the Wald statistic t^2, for a and r small and of the same order and $\beta = 0$, is of order r^{-2}. For large β, therefore, the distribution of the bootstrap Wald statistic, under the null, is of order $\tilde{r}^{-2}$, which we have just seen is the same order as β^2. But the distribution of the Wald statistic itself for large β is also of order β^2, unlike the K and LR statistics. Although the distribution of the actual statistic t^2 for large β and that of the bootstrap statistic $(t^*)^2$ are not the same, and are unbounded, the distributions of t^2/β^2 and $(t^*)^2/\beta^2$ are of order unity in probability, and, having support on the whole real line,

they overlap. Thus the probability of rejection of the null by the bootstrap test does not tend to 1 however large β may be.

This conclusion, which is borne out by the simulation experiments of the next section, merits some discussion. In Horowitz and Savin (2000), it is pointed out that, unless one is working with pivotal statistics, it is not in general possible to define an empirically relevant definition of the power of a test that does not have true level equal to its nominal level. They conclude that the best measure in practice is the rejection probability of a well-constructed bootstrap test.

In Davidson and MacKinnon (2006b), we point out that, even for well-constructed bootstrap tests, ambiguity remains in general. Only when the bootstrap DGP is asymptotically independent of the asymptotically pivotal statistic being bootstrapped can level adjustment be performed unambiguously on the basis of the DGP in the null hypothesis of which the parameters are the probability limits of the estimators used to define the bootstrap DGP. This result, as proved, applies only to the parametric bootstrap, and, more importantly here, to cases in which these estimators have nonrandom probability limits. But, as we have seen, that is not the case here. It seems therefore that there is no theoretically satisfying measure of the power of tests for which the bootstrap DGP is asymptotically nonrandom. It is therefore pointless to try to refine our earlier result for the Wald test, whereby we learn merely that its bootstrap version is inconsistent.

Because the Wald statistic is not boundedly pivotal, a test based on it has size equal to one, as shown by Dufour (1997). Dufour also draws the conclusion that no Wald-type confidence set based on a statistic that is not at least boundedly pivotal can be valid, whether or not the confidence set is constructed by bootstrapping. If, however, instead of using a conventional bootstrap confidence set, we invert the bootstrap Wald test to obtain a confidence set that contains all parameter values which are not rejected by a bootstrap Wald test, we may well obtain confidence sets with a level of less than one, since, on account of the inconsistency of the bootstrap test, unbounded confidence sets can arise with positive probability.

We mentioned earlier that the conditional LR, or CLR, test of Moreira (2003) has a bootstrap interpretation. We may consider the variable TT defined in (13) as a random variable on which a bootstrap distribution is conditioned. In fact, as can be seen from (35) and (54), TT is equivalent under weak-instrument asymptotics to $\tilde{a}^2/r^2$. Conditional on TT, Moreira shows that the statistics LR and LR_0 are asymptotically pivotal. Thus, rather than estimating both a and ρ and using the estimates to generate bootstrap versions of the six sufficient statistics (6), we can evaluate TT and then generate s simulated versions of the two (conditionally) sufficient statistics SS and ST based on their asymptotic conditional distributions, as discussed earlier. From this, we can obtain conditional empirical distributions for either LR or LR_0 which may be used to compute P values in the usual way.

This procedure is not quite a real bootstrap, although it is almost as computationally intensive as a fully parametric bootstrap based on simulating the six quantities. Moreover, the CLR test as originally proposed involves an approximation which may not be a good one in small samples. The "bootstrap" conditional distributions of SS

and ST are not known exactly. Instead, they are approximated on the basis of the distributions when the contemporaneous covariance matrix is known.

Recently, Moreira, Porter, and Suarez (2005) have proposed a "conditional bootstrap" CLR test which uses the pairs bootstrap to generate the two sufficient statistics, but still conditions on TT. In the case of the LR_0 statistic, for each of B bootstrap samples, they compute the statistic

$$\mathrm{LR}^*_{0j} = \frac{1}{2}\Big(SS^*_j - TT + \sqrt{(SS^*_j + TT)^2 - 4TT\big(SS^*_j - (ST^*_j)^2/TT^*_j\big)}\,\Big). \tag{56}$$

The quantities SS^*_j, ST^*_j, and TT^*_j here are computed from the j^{th} bootstrap dataset, but TT is computed from the actual data. The "statistic" that we will call CLRb then has the form of a bootstrap P value. It is simply the fraction of the bootstrap samples for which LR^*_{0j} exceeds LR_0.

In principle, the CLRb test can be based on any bootstrap DGP. The pairs bootstrap is not a good choice, however, because its bootstrap DGP does not satisfy the null hypothesis. This makes it quite tricky to compute SS^*_j, ST^*_j, and TT^*_j, as we discuss at the end of this section. Moreover, as our simulation results show, the pairs bootstrap tends to perform much less well than our new RE and REC bootstraps for all the test statistics. It therefore seems attractive to consider CLRb tests based on the semiparametric versions of the RE and REC bootstraps that are described below. We study these in the next section, and they turn out to work very well indeed.

Some remarks concerning bootstrap validity are in order at this point. If the statistic that is bootstrapped is asymptotically pivotal, then the bootstrap is valid asymptotically in the sense that the difference between the bootstrap distribution and the distribution under the true DGP, provided the latter satisfies the null hypothesis, converges to zero as the sample size tends to infinity; see, among many others, Davidson and MacKinnon (2006a). The bootstrap provides higher-order refinements if, in addition, the bootstrap DGP consistently estimates the true DGP under the null; see Beran (1988). Yet another level of refinement can be attained if the statistic bootstrapped is asymptotically independent of the bootstrap DGP; see Davidson and MacKinnon (1999). All of these requirements are satisfied by any of the statistics considered here under strong-instrument asymptotics when either the RE or REC bootstrap is used.

Besides the AR statistic, only the K statistic and CLR test P value are asymptotically pivotal with weak instruments, and so it is only for them that we can conclude without further ado that the bootstrap is valid with weak-instrument asymptotics. However, even if the statistic bootstrapped is not asymptotically pivotal, the bootstrap may still be valid if the bootstrap DGP consistently estimates the true DGP under the null. But, with weak instruments, this is true of no conceivable bootstrap method, since there is no consistent estimate of the parameter a. Consequently, however large the sample size may be, the bootstrap distributions of the Wald statistic and the LR statistic are different from their true distributions.

Although our discussion has focused on parametric bootstrap procedures, we do not necessarily recommend that they should be used in practice, since the assumption of

Gaussian disturbances may often be uncomfortably strong. Any parametric bootstrap procedures that are valid with Gaussian disturbances, under either weak- or strong-instrument asymptotics, remain valid with non-Gaussian disturbances provided only that laws of large numbers and central limit theorems can be applied, as discussed in Section 4. This follows because, under those conditions, asymptotically pivotal statistics remain so, and the parameter estimators are consistent or not regardless of whether the disturbances are Gaussian.

If the disturbances are very far from being normally distributed, we may reasonably expect that a semiparametric resampling bootstrap will work better than a parametric bootstrap. All of the parametric bootstrap procedures that we have discussed have semiparametric analogs which do not require that the disturbances should be normally distributed. We discuss only the RE and REC bootstraps, partly because they are new, partly because they will be seen in the next section to work very well, and partly because it will be obvious how to construct semiparametric analogs of the other procedures.

For the semiparametric RE bootstrap, we first estimate equation (1) under the null hypothesis to obtain restricted residuals $\boldsymbol{M_Z y}_1$. We then run regression (9) or, equivalently, regression (50). The residual vector $\tilde{\boldsymbol{u}}_2$ that we want to resample from is the vector of residuals from the regression plus $\hat{\delta}\boldsymbol{M_Z y}_1$. The bootstrap DGP is then

$$\begin{aligned} \boldsymbol{y}_1^* &= \boldsymbol{u}_1^* \\ \boldsymbol{y}_2^* &= \boldsymbol{W}_1\tilde{\boldsymbol{\pi}}_1 + \boldsymbol{u}_2^*, \end{aligned} \qquad [\boldsymbol{u}_1^* \;\; \boldsymbol{u}_2^*] \sim \text{EDF}[\tilde{\boldsymbol{u}}_1 \;\; \tilde{\boldsymbol{u}}_2]. \tag{57}$$

Thus the bootstrap disturbances are resampled from the joint EDF of the two residual vectors. This preserves the sample correlation between them.

For the REC bootstrap, we need to use a different set of fitted values in the reduced-form equation. To do so, we first compute $\tilde{a}^2$ using either the formula (53) or, more conveniently,

$$\tilde{a}^2 = \frac{\tilde{\boldsymbol{\pi}}_1^\top \boldsymbol{W}_1^\top \boldsymbol{W}_1 \tilde{\boldsymbol{\pi}}_1}{\tilde{\boldsymbol{u}}_2^\top \tilde{\boldsymbol{u}}_2 / n},$$

which is the square of (48) evaluated at the appropriate values of $\boldsymbol{\pi}_1$ and $\boldsymbol{u}_2$. Then we calculate $\tilde{a}^2_{\text{BC}}$ from (55). The bootstrap DGP is almost the same as (57), except that the fitted values $\boldsymbol{W}_1\tilde{\boldsymbol{\pi}}_1$ are replaced by $\tilde{a}_{\text{BC}}/\tilde{a}$ times $\boldsymbol{W}_1\tilde{\boldsymbol{\pi}}_1$. This reduces the length of the vector of fitted values somewhat. The fitted values actually shrink to zero in the extreme case in which $\tilde{a}_{\text{BC}} = 0$.

The quantity $\tilde{a}^2$ is very closely related to the "test statistic" for weak instruments recently proposed by Stock and Yogo (2005); recall that $\boldsymbol{Z}^\top \boldsymbol{W}_1 = \boldsymbol{O}$. When it is large, the instruments are almost certainly not weak, and even asymptotic inference should be reasonably reliable. When it is very small, however, many tests are likely to overreject severely, and those that do not are likely to be seriously lacking in power. There is evidence on these points in the next section.

An alternative to the parametric and semiparametric bootstrap methods that we have discussed in this section is the pairs bootstrap, which was proposed by Freedman

(1984). The simplest way to implement the pairs bootstrap is just to resample from the rows of the matrix

$$[\boldsymbol{y}_1 \;\; \boldsymbol{y}_2 \;\; \boldsymbol{Z} \;\; \boldsymbol{W}].$$

Moreira, Porter, and Suarez (2005) describe an alternative, semiparametric, resampling procedure, but it yields exactly the same results as the ordinary pairs bootstrap when applied to any of the tests that we study. One potential advantage of the pairs bootstrap is that it is valid in the presence of heteroskedasticity of unknown form. However, it is quite easy to create wild bootstrap versions of all the semiparametric bootstrap procedures that we have discussed, which are also valid for this case; see Davidson and MacKinnon (2007).

Because the pairs bootstrap does not impose the null hypothesis, it is necessary to modify the bootstrap test statistics so that what they test is true in the bootstrap samples. For the t statistic, this simply means replacing $\beta_0 = 0$ by $\beta_0 = \hat{\beta}$ in the numerator of the statistic. For K and LR, however, it means computing the quantities SS^*, ST^*, and TT^* under the null hypothesis that $\beta = \hat{\beta}_{\text{LIML}}$, where $\hat{\beta}_{\text{LIML}}$ denotes the LIML estimate of β.[1] Thus the pairs bootstrap is relatively difficult to implement. Moreover, as we will see in the next section, it generally performs much less well than the RE and REC bootstraps.

6. Simulation Evidence

In this section, we report the results of a large number of simulation experiments with data generated by the simplified model (29) and (30). All experiments have 100,000 replications for each set of parameter values. In many of the experiments, we use fully parametric bootstrap DGPs, but we also investigate the pairs bootstrap and the semiparametric RE and REC bootstraps.

In most of the experiments, $n = 100$. Using larger values of n, but for the same value(s) of a, would have had only a modest effect on the results for most of the bootstrap tests; see Figures 9 and 10. Many of the asymptotic tests are quite sensitive to n, however; we provide some evidence on this point in Figure 3. For the base cases, we consider every combination of $a = 2$ and $a = 8$ with $\rho = 0.1$ and $\rho = 0.9$. The limiting R^2 of the reduced-form regression (30) is $a^2/(n + a^2)$. Thus, when $a = 2$, the instruments are very weak, and, when $a = 8$, they are moderately strong. When $\rho = 0.1$, there is not much correlation between the structural and reduced-form disturbances; when $\rho = 0.9$, there is a great deal of correlation.

All our results are presented graphically. Note that, within each figure, the vertical scale often changes, because otherwise it would be impossible to see many important differences between alternative tests and bootstrap methods. Readers should check the vertical scales carefully when comparing results across figures or in different panels of the same figure.

[1] It would also be possible to compute SS^*, ST^*, and TT^* under the null that $\beta = \hat{\beta}$. However, since the K and LR tests are based on LIML estimates rather than IV ones, this does not seem appropriate, and it works much less well.

Figures 1 through 4 concern the properties of asymptotic tests. Figure 1 shows the rejection frequencies for the Wald, K, LR, and CLR tests (the last of these uses LR rather than LR_0 and is based on 199 simulations) for the four base cases as functions of $l - k$. These are all increasing functions, so test performance generally deteriorates as the number of over-identifying restrictions, $l - k - 1$, increases. Of particular note are the extremely poor performance of the Wald test when $\rho = 0.9$ and the surprisingly poor performance of the LR test when $\rho = 0.1$. It is also of interest that the Wald test underrejects severely when ρ, a, and $l - k$ are all small. This is a case that is rarely investigated in simulation experiments.

Figure 2 shows rejection frequencies as functions of ρ for four values of a. In this and all subsequent figures, $l - k = 9$, so that there are eight overidentifying restrictions. As expected, performance improves dramatically as a increases. The Wald test is extremely sensitive to ρ. The others, especially K and CLR, are much less so. Only when ρ is small and a is large does the Wald test perform at all well.

In Figure 3, we consider values of n between 20 and 1280 that increase by factors of approximately $\sqrt{2}$. This figure makes it clear that the somewhat mediocre performance of K and CLR evident in the first two figures is a consequence of using $n = 100$. The performance of both these tests always improves dramatically as n increases. Recall that K and CLR are asymptotically pivotal under both weak-instrument and conventional asymptotics. Thus it is not surprising that they can safely be used as asymptotic tests when the sample size is large but the instruments are weak. The performance of LR also improves as n increases, but it continues to overreject, sometimes very severely, even for large values of n. The Wald test is the least sensitive to n, but its performance often deteriorates as n increases.

Figure 4 shows what happens as a varies. We consider values from $a = 1$ to $a = 64$ that increase by factors of $\sqrt{2}$. As expected, the performance of the Wald and LR tests improves dramatically as a increases. There is little effect on K and only a modest effect on CLR. The latter tests would instead benefit from a larger sample size, holding a constant.

Figures 5, 6, and 7 concern the properties of parametric bootstrap tests under the null hypothesis. In all cases, $B = 199$. Results are presented for the five different bootstrap DGPs that were discussed in Section 5 and for the pairs bootstrap. Procedures with an "R" employ restricted estimates of the structural equation, while procedures with a "U" employ unrestricted estimates. Procedures with an "E" employ efficient estimates of the reduced-form equation, while procedures with an "I" employ inefficient ones. The REC procedure bias-corrects the estimate of a.

Figure 5 shows rejection frequencies for the Wald test as a function of a for two values of ρ, in the top two panels, and as a function of ρ for two values of a, in the bottom two panels. Our new RE and REC bootstraps perform reasonably well, although they do lead to significant underrejection in some cases. The other four methods lead to very severe overrejection when ρ is not small and a is not large. Despite this, the UE and pairs bootstraps actually underreject very severely when both a and ρ are small. It is interesting that, in all four panels, the performance of the UE and pairs bootstraps is very similar. It is also apparent, most clearly in the lower left-hand panel, that RI

and RE yield similar results when ρ is small. This makes sense, because $\tilde{\boldsymbol{\pi}}$ cannot be much more efficient than $\hat{\boldsymbol{\pi}}$ when there is little correlation between the structural and reduced form equations.

Figure 6 shows rejection frequencies for the K test in the same format as Figure 5. All methods except the pairs bootstrap, which always underrejects, work very well, with REC being arguably the best of a remarkably good bunch. Both a and n must be quite large for the pairs bootstrap to perform as well as RE and REC do when $a = 2$ and $n = 100$.

Figure 7 deals with the LR test, for which the REC bootstrap is unquestionably the best method, overall, in every one of the four panels. Using REC leads to only very modest overrejection in the worst cases, when ρ and a are both small. Except in the upper right-hand panel, the pairs bootstrap performs quite well here, but it is the only method for which the rejection frequency does not seem to converge to .05 as a becomes large. For that to happen, n must be quite large.

Figure 8 deals with the CLR and CLRb tests. The former are based on LR and the latter on LR_0. With $n = 100$, it would have made almost no difference if they had both been based on either LR or LR_0. As we have remarked, bootstrapping CLR is almost as computationally intensive as performing a double bootstrap.[2] In contrast, calculating the CLRb test is no more expensive than bootstrapping any of the other tests. To avoid cluttering the figure, we present results for only four cases, namely, the CLR test bootstrapped using the RE and REC bootstraps and the CLRb test computed using the RE and REC bootstraps.[3] Other bootstrap methods performed less well overall.

In Figure 8, the CLRb tests always perform extraordinarily well, as does the REC bootstrap version of the CLR test. The two CLRb tests tend to overreject very slightly when a is small, less for the version based on REC than for the one based on RE. Since CLRb, when based on REC, is very much faster and easier to compute than CLR bootstrapped using REC and performs just as well, there appears to be no reason to consider the latter any further.

As we mentioned at the end of the last section, it is probably better in practice to use a semiparametric rather than a fully parametric bootstrap, because the normality assumption is likely to be false. We therefore undertook a number of experiments to compare parametric and semiparametric versions of the REC and RE bootstraps. In every case, the semiparametric and fully parametric bootstraps yield very similar

[2] A recent paper by Hillier (2006) derives the exact conditional distribution of the LR statistic under the assumption that the disturbance covariance matrix $\boldsymbol{\Sigma}$ is known. Critical values for the test can be obtained from this conditional distribution by numerical methods. Use of these critical values might considerably reduce the computational burden of bootstrapping the CLR test.

[3] When bootstrapping the CLR test, we used $B = 199$ together with $s = 299$ simulations. It is important that s and B not be the same, because, if they are, the actual and bootstrap statistics will be equal with probability approximately $1/(B+1)$. Different choices for s and B would have resulted in slightly different results.

results. Of course, this would quite possibly not be the case if the disturbances in the DGP were not normally distributed.

Figures 9 and 10 plot rejection frequencies for parametric and semiparametric bootstrap tests as functions of the sample size for four sets of parameter values. For the Wald, K, and LR tests, they show rejection frequencies under the null as a function of the sample size for the REC and RE bootstraps, respectively, both parametric and semiparametric. The similarity of the results from the parametric and semiparametric bootstraps is striking. Note that the same random numbers were used to generate the underlying data, but different ones were used for bootstrapping, because generating pseudo-random normal variates does not work the same way as resampling.

The information provided by Figures 9 and 10 for the CLR tests differs from that for the other three tests. They show rejection frequencies for the original CLR test bootstrapped using the semiparametric REC or RE bootstraps and for the REC or RE versions of the CLRb test.[4] Even when bootstrapped, the CLR tests perform poorly when n and a are small, while the simpler CLRb tests perform very much better. Strikingly, the performance of CLRb is remarkably similar to that of K, especially for the REC bootstrap.

Figure 9, which presents the results for the REC bootstrap, makes it clear that the decision to focus on the case $n = 100$ in most of our experiments is not entirely inconsequential. The K test works very well indeed for all but the smallest sample sizes, as does the CLRb test. The Wald test underrejects for the smaller sample sizes in three cases out of four, but its performance improves as n increases. When $a = 2$, the LR test overrejects moderately for small sample sizes and underrejects moderately for large ones.

Figure 10 is similar to Figure 9, but it presents results for the RE bootstrap. Once again, we see that the fully parametric and semiparametric bootstraps produce almost identical results with normally distributed disturbances. At least in some cases, the RE procedure is substantially inferior to the REC one. In particular, the LR test overrejects quite severely when $a = 2$ and $\rho = 0.1$, and the Wald test performs less well for cases where a is small and n is large. However, for the K and CLRb tests, performance is once again excellent for $n \geq 50$.

It emerges clearly from these last two figures that the rejection frequencies of the RE bootstrap Wald and LR tests do not seem to converge to the nominal level as $n \to \infty$ when $a = 2$, whereas those of the K and CLRb tests do so. This is in accord with our discussion of the previous section. We echo Moreira, Porter, and Suarez (2005), however, in noting that it is remarkable that the bootstrap Wald and LR tests perform as well as they do.

Figures 11, 12, and 13 concern power. Because there is no point comparing the powers of tests that do not perform reliably under the null, only the (semiparametric)

[4] The CLR test here is based on LR. If it had instead been based on LR_0, it would have overrejected quite a bit more for very small values of n, but results for $n \geq 50$ would have been essentially identical.

REC bootstrap is used. In these figures, we present results for the AR test (not bootstrapped, since it is exact), the Wald test, the K test, and the CLRb test.[5] Since it is no more expensive to compute CLRb than to bootstrap the LR test, results for LR (which is generally less reliable under the null) are not reported. To reduce the power loss associated with small values of B, we set $B = 499$ in these experiments, which made them relatively expensive to perform.

Figure 11 deals with the weak-instrument case in which $a = 2$. The three panels correspond to $\rho = 0.1$, $\rho = 0.5$, and $\rho = 0.9$. No test has good power properties. When $\rho = 0.1$, CLRb generally has the most power, but AR is close behind and actually seems to be a little bit more powerful when $|\beta|$ is large. Both these tests are often much more powerful than K, the only other test that is reliable under the null, and CLRb is never less powerful. The Wald test appears to be the most powerful test for certain values of β, but it performs poorly when $|\beta|$ is large.

When $\rho = 0.5$, all the tests have strange-looking power functions. They tend to have more power against negative values of β than against positive ones. There is a small region in which Wald dominates, but it is severely lacking in power when $|\beta|$ is large. Once again, CLRb and AR are generally quite close. AR seems to have a bit more power for $\beta < -1$, but CLRb dominates for most other values of β. The K test is severely lacking in power for $\beta < 0$, but much less so for $\beta > 0$.

When $\rho = 0.9$, the power functions look stranger still, with far less power against positive values of β than against negative ones. The power function for K has a curious dip for some negative values of β, and in this region K can be much less powerful than AR. The dip is also evident in the simulation results of Stock, Wright, and Yogo (2002), which are not directly comparable to ours, since they use the same parametrization as Andrews, Moreira, and Stock (2006) and assume that $\boldsymbol{\Sigma}$ is known. See Poskitt and Skeels (2005) and Kleibergen (2007) for an explanation of this dip. The minimum of the power function for the Wald test occurs well to the left of $\beta = 0$. The CLRb test is more powerful than AR except in a small region near $\beta = -1$. Oddly, however, K is just a little bit more powerful than CLRb for most positive values of β.

Figures 12 and 13 show what happens as the instruments become stronger. They deal with the cases in which $a = 4$ and $a = 8$, respectively. As instrument strength increases, all the tests perform very much better. Even K outperforms AR for many values of β, especially when $\rho = 0.9$. However, when $\rho = 0.5$ and $\rho = 0.9$, the power function for K once again has a curious dip for some negative values of β, and the power functions for Wald have their minima noticeably to the left of $\beta = 0$. The power functions for K when $\rho = 0.1$ are particularly strange, since power actually declines as the absolute value of β increases beyond a certain point. The CLRb test does not have any of these problems, and it is very reliable under the null.

Results for the RE bootstrap are not reported because, in most respects, they are quite similar to the ones in Figures 11, 12, and 13. In several cases, the Wald test

[5] The Wald and K tests were bootstrapped using the semiparametric REC bootstrap rather than the parametric one for comparability with CLRb.

appears to have somewhat more power with RE than with REC, mainly because it is less prone to underreject, or even prone to overreject, under the null. Similarly, the RE version of the CLRb test tends to have very slightly more power than the REC version because it is very slightly more prone to overreject.

7. Concluding Remarks

We have provided a detailed analysis of the properties of several tests for the coefficient of a single right-hand-side endogenous variable in a linear structural equation estimated by instrumental variables. First, we showed that the Student's t (or Wald) statistic, Kleibergen's K statistic, and the LR statistic can be written as functions of six random quantities. The Anderson-Rubin test is also a function of two of these six quantities. Using these results, we obtained explicit expressions for the asymptotic distributions of all the test statistics under both conventional and weak-instrument asymptotics.

Under weak-instrument asymptotics, we found that none of the test statistics can have any real asymptotic power against local alternatives. Even when the alternative is fixed, AR, K, and LR are not consistent tests under weak-instrument asymptotics. The t test has very different properties, however. It is unbounded as $\beta \to \infty$, so that it appears to be consistent. But it is also unbounded as certain parameters of the DGP tend to limiting values, so that it is not asymptotically pivotal, or even boundedly pivotal. Note that these results depend in an essential way on how the DGP is specified, in particular, the disturbance covariance matrix.

We then proposed some new procedures for bootstrapping the three test statistics. Our RE and REC procedures use more efficient estimates of the coefficients of the reduced-form equation than existing procedures and impose the restriction of the null hypothesis. In addition, the REC procedure corrects for the tendency of the reduced-form equation to fit too well. A semiparametric version of this procedure is quite easy to implement. In most cases, the REC bootstrap outperforms the RE bootstrap, which in turn outperforms previously proposed methods. The improvement can be quite dramatic.

Even the Wald test performs quite well when bootstrapped using these procedures, although it sometimes underrejects fairly severely. Interestingly, however, as we show analytically, the RE and REC bootstrap versions of the Wald test are not consistent against fixed alternatives under weak-instrument asymptotics, and they can be much less powerful than the other tests when the instruments are weak and $|\beta|$ is large.

We also proposed two new variants of the conditional bootstrap LR test, or CLRb, based on the RE and REC bootstraps. Like the K test when bootstrapped using either the RE or REC procedures, these new CLRb tests have excellent performance under the null, even when the sample size is small and the instruments are weak. Unlike the K test, their power functions have no strange features.

All of our theoretical analysis is conducted under the assumption that the disturbances are Gaussian, although some results do not in fact depend on this assumption. To

our knowledge, little work has been done on the properties of tests in the presence of weak instruments and non-Gaussian disturbances. We conjecture that the qualitative features of the tests considered in this paper, both asymptotic and bootstrap, do not greatly depend on the assumption of Gaussianity.

In the light of our results, it is tempting to conclude that, when the number of over-identifying restrictions is large, so that the AR test may suffer significant power loss, the best method to use is the CLRb test based on the REC bootstrap. It has better performance under the null than every test except the K test bootstrapped using the same method (with which it is pretty much tied), and it can sometimes have substantially more power than the K test. However, when the instruments are even moderately strong and the sample size is not small, all the tests perform quite well when bootstrapped using the new RE and REC bootstraps.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd edition, Wiley, New York.

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

Andrews, D. W. K., M. J. Moreira, and J. H. Stock (2006). "Optimal two-sided invariant similar tests for instrumental variables regression", *Econometrica*, 74, 715–752.

Beran, R. (1988). "Prepivoting test statistics: A bootstrap view of asymptotic refinements", *Journal of the American Statistical Association*, 83, 687-697.

Davidson, R. and J. G. MacKinnon (1999). "The size distortion of bootstrap tests", *Econometric Theory*, 15, 361–376.

Davidson, R. and J. G. MacKinnon (2006a). "Bootstrap methods in econometrics", Chp. 23 in *Palgrave Handbook of Econometrics*, Volume 1, eds. T. C. Mills and K. Patterson, Palgrave-Macmillan, Basingstoke, 812–838.

Davidson, R. and J. G. MacKinnon (2006b). "The power of bootstrap and asymptotic tests", *Journal of Econometrics*, 133, 421–441.

Davidson, R. and J. G. MacKinnon (2007). "Wild bootstrap tests for IV regression," Queen's Economics Department Working Paper No. 1135.

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

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

Hillier, G. (2006). "Exact critical value and power functions for the conditional likelihood ratio and related tests in the IV regression model with known covariance", CeMMAP Working Paper No. CWP23/06.

Horowitz, J. L., and N. E. Savin (2000). "Empirically relevant critical values for hypothesis tests", *Journal of Econometrics*, 95, 375–389.

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

Kleibergen, F. (2007). "Generalizing weak instrument robust IV statistics towards multiple parameters, unrestricted covariance matrices and identification statistics," *Journal of Econometrics*, 139, 181–216.

Mariano, R. S., and T. Sawa (1972). "The exact finite-sample distribution of the limited-information maximum likelihood estimator in the case of two included endogenous variables", *Journal of the American Statistical Association*, 67, 159–163.

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

Moreira, M. J., J. R. Porter, and G. A. Suarez (2005). "Bootstrap and higher-order expansion validity when instruments may be weak", NBER Working Paper No. 302, revised.

Phillips, P. C. B. (1983). "Exact small sample theory in the simultaneous equations model", in *Handbook of Econometrics*, Vol I, eds. Z. Griliches and M. D. Intriligator, North Holland.

Poskitt, S. S., and C. L. Skeels (2005). "Small concentration asymptotics and instrumental variables inference", University of Melbourne, Department of Economics Working Paper 948.

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, 518–529.

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

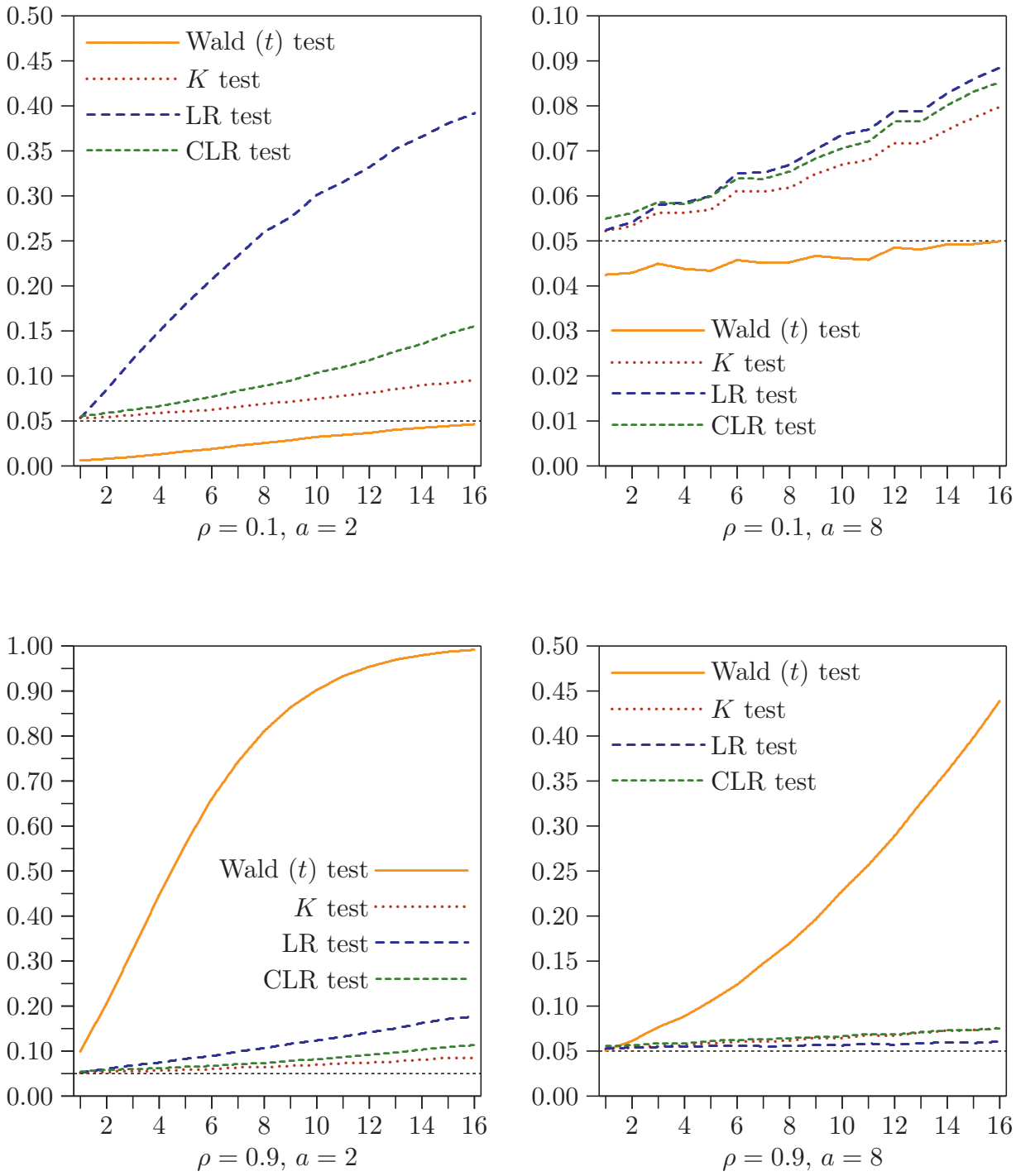

Figure 1. Rejection frequencies for asymptotic tests as functions of $l - k$, $n = 100$

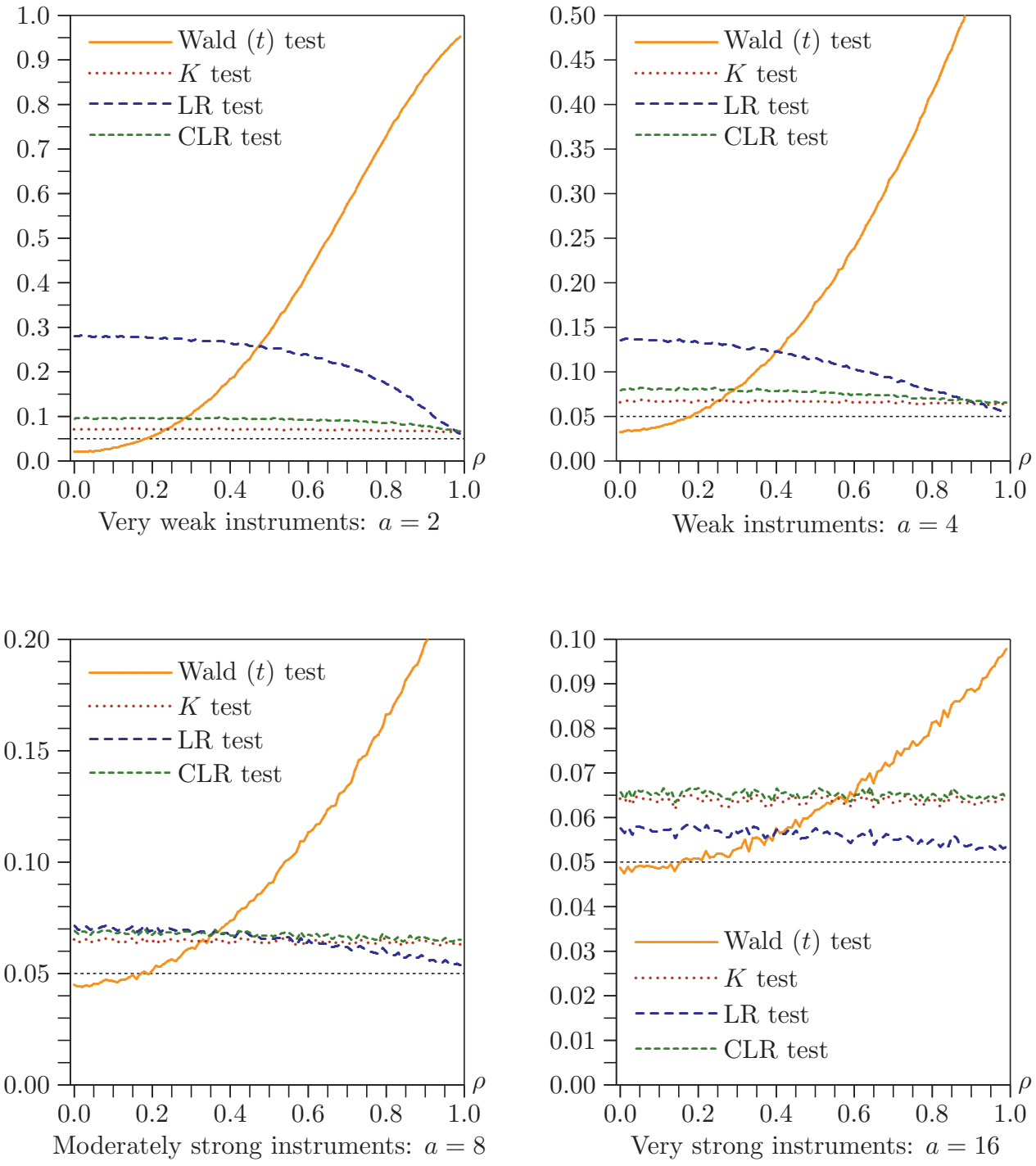

Figure 2. Rejection frequencies for asymptotic tests as functions of ρ for $l - k = 9$, $n = 100$

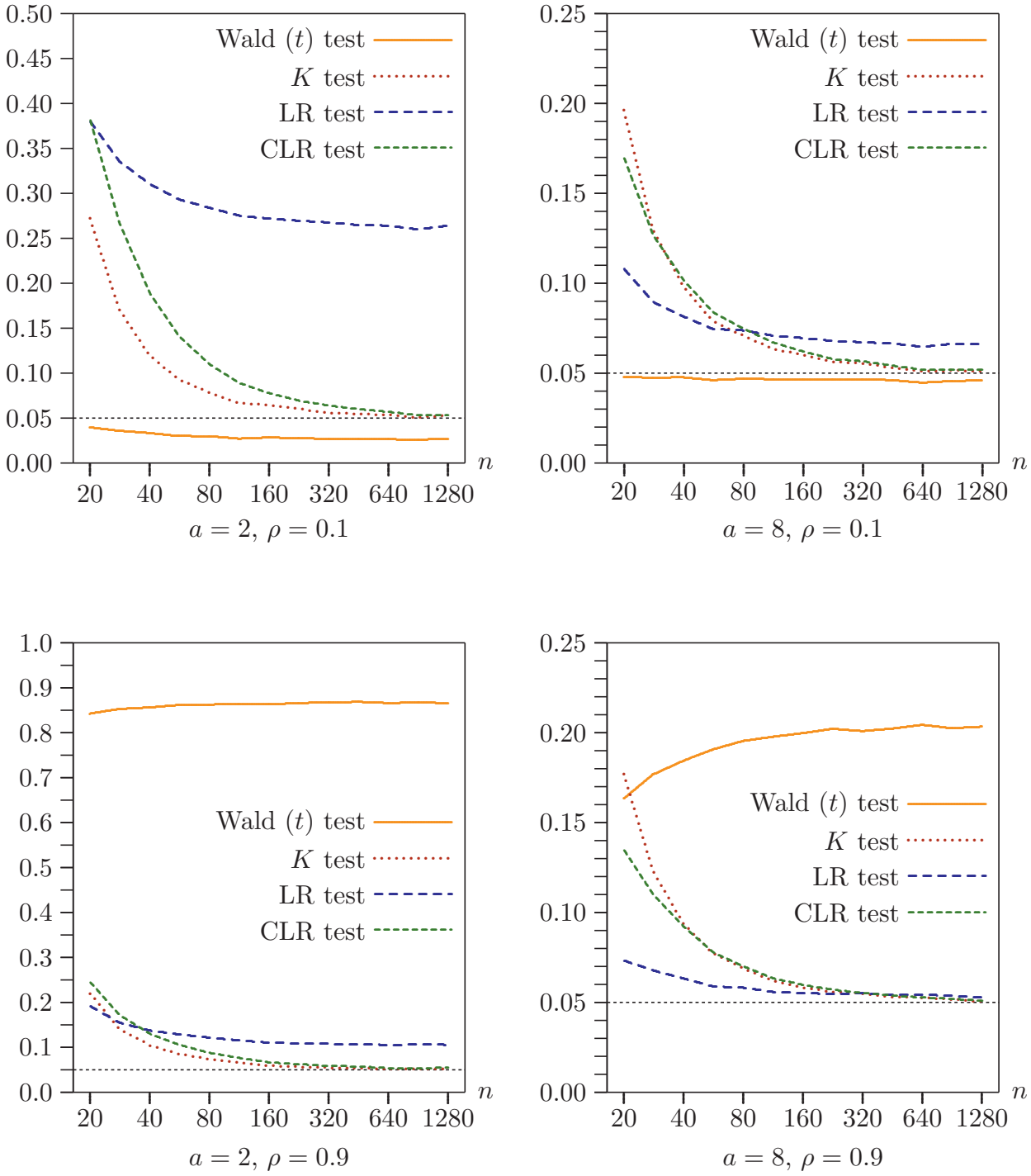

Figure 3. Rejection frequencies for asymptotic tests as a function of n for $l - k = 9$

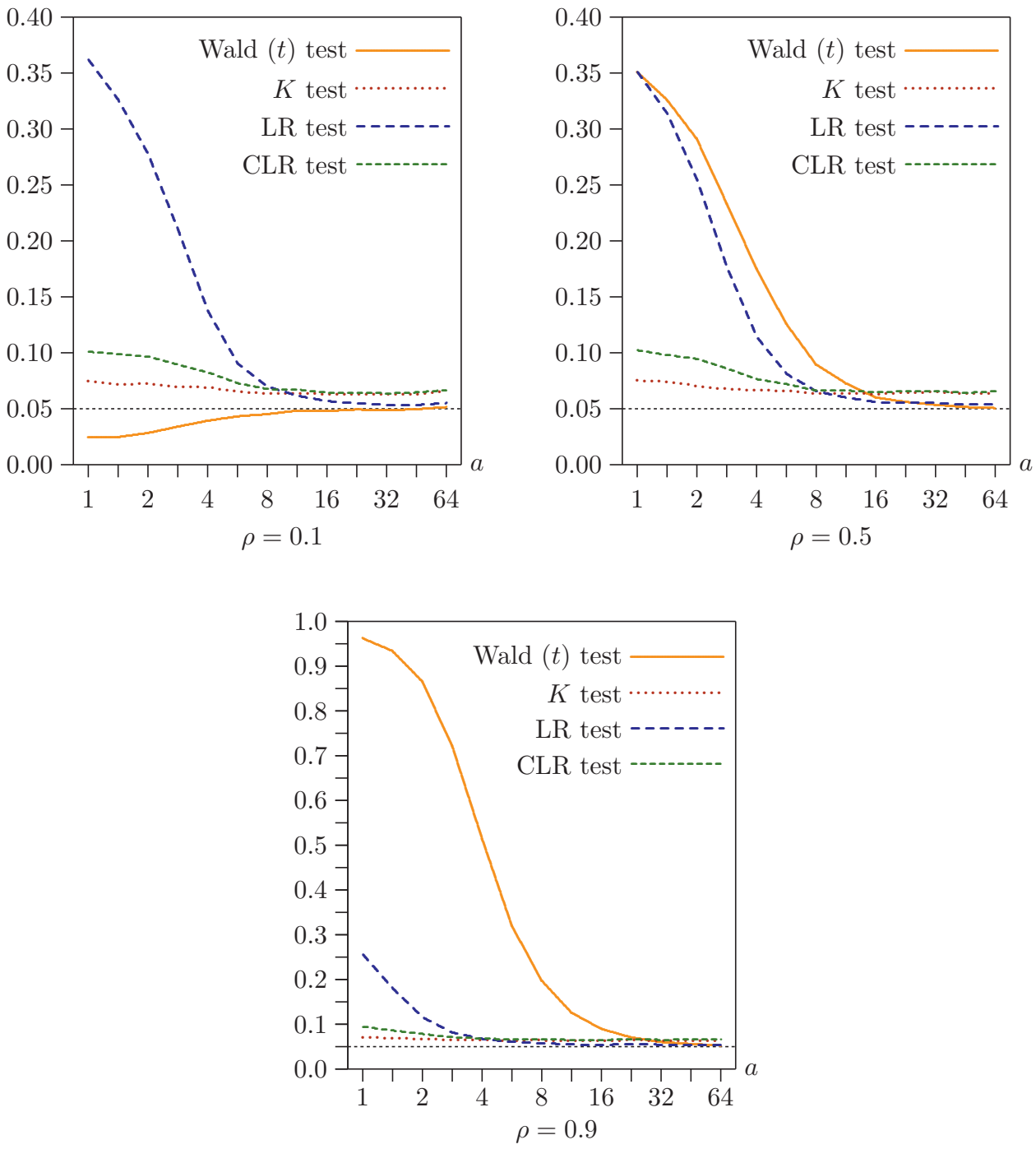

Figure 4. Rejection frequencies for asymptotic tests as functions of a, for $l - k = 9$, $n = 100$

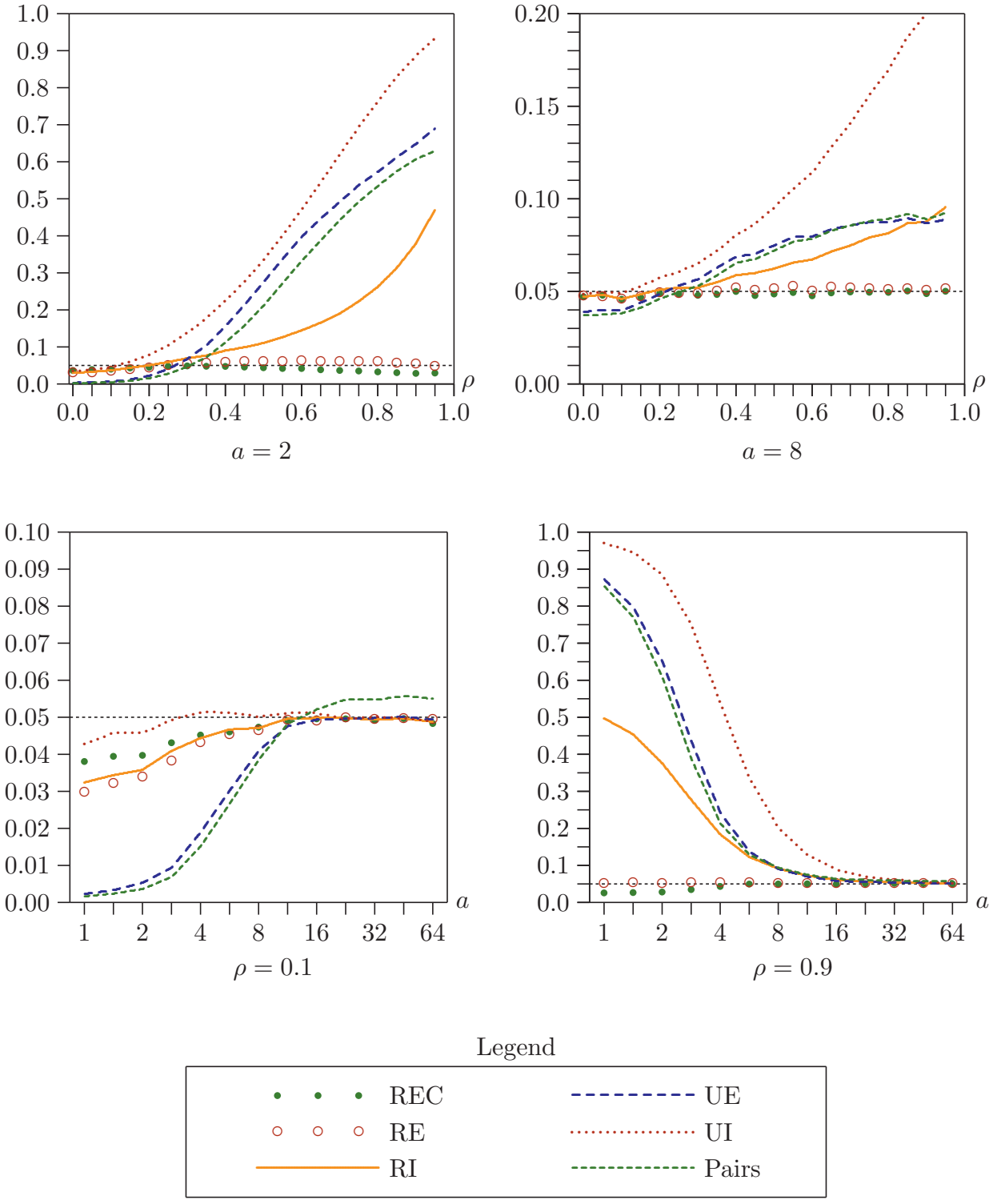

Figure 5. Rejection frequencies for bootstrap Wald (t) tests for $l - k = 9$, $n = 100$

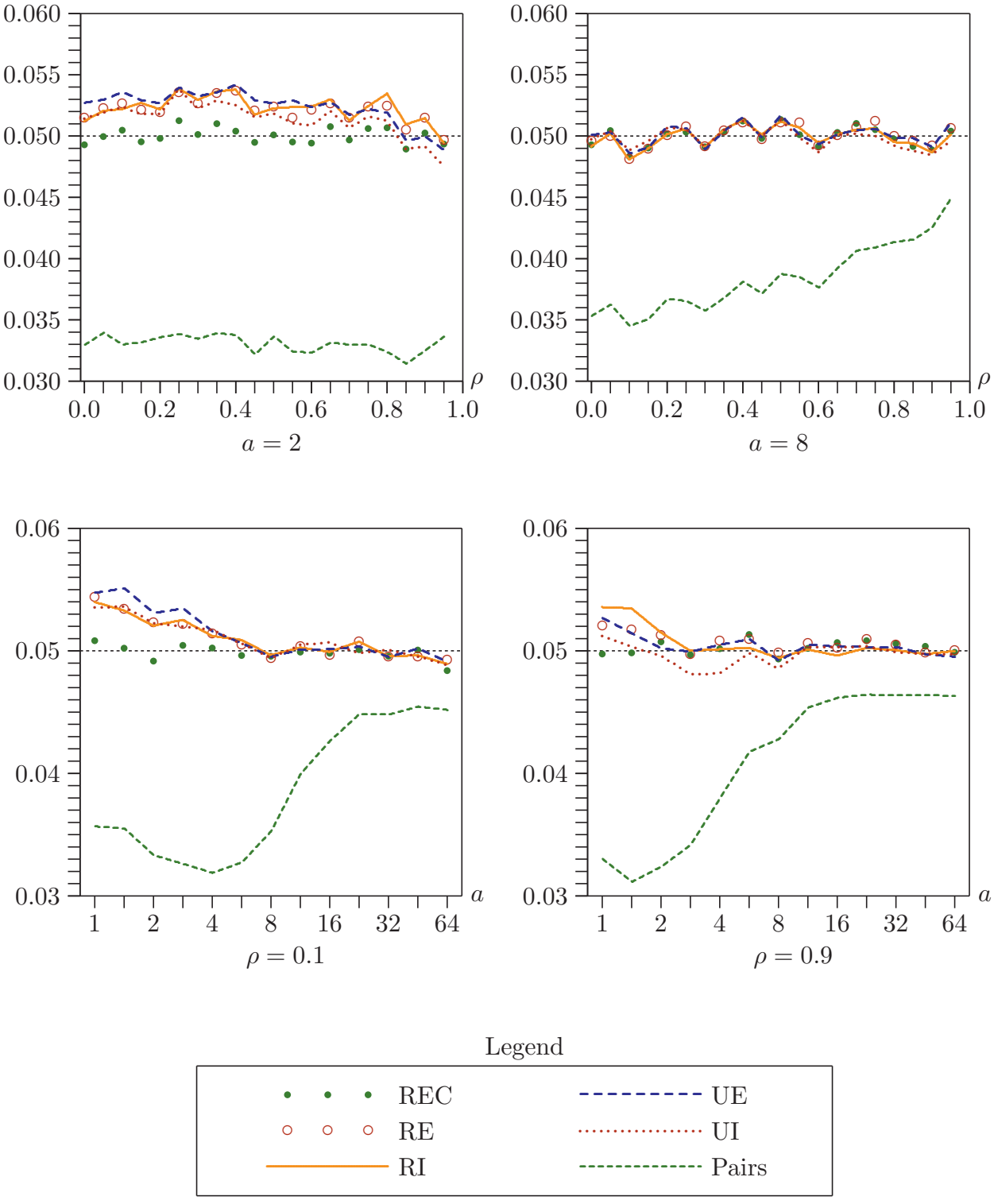

Figure 6. Rejection frequencies for bootstrap K tests for $l - k = 9$, $n = 100$

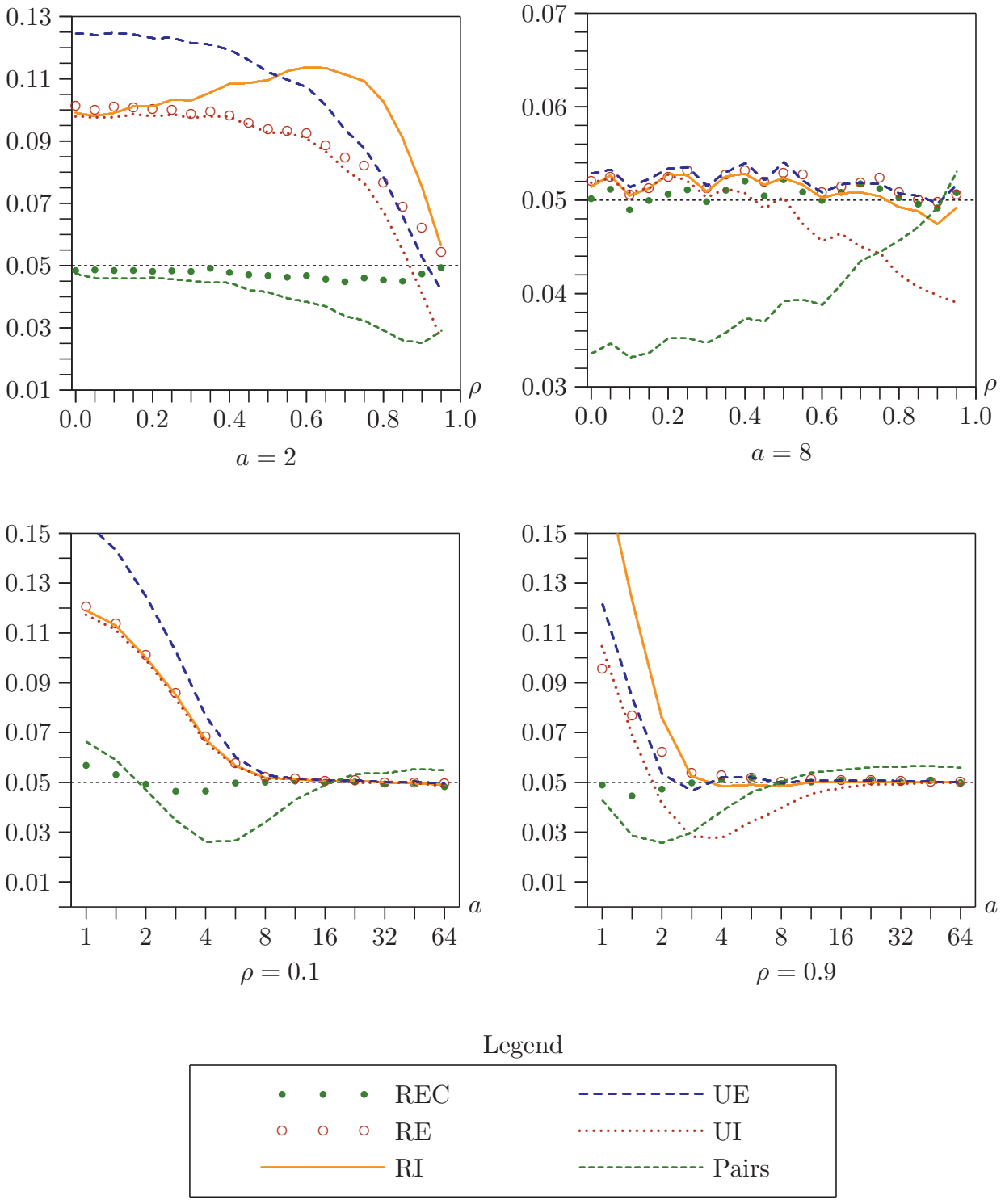

Figure 7. Rejection frequencies for bootstrap LR tests for $l - k = 9$, $n = 100$

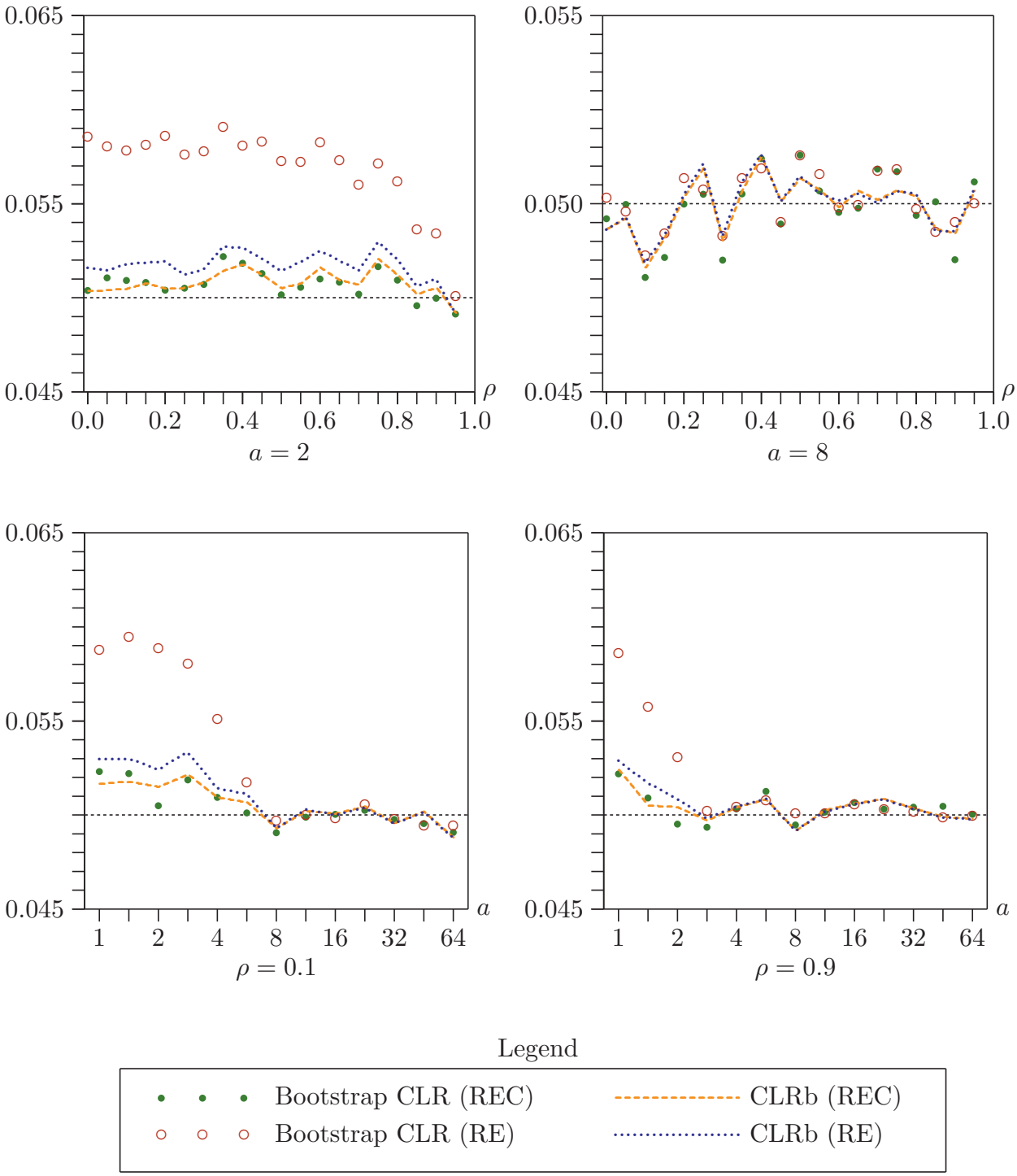

Figure 8. Rejection frequencies for bootstrap CLR and CLRb tests for $l - k = 9$, $n = 100$

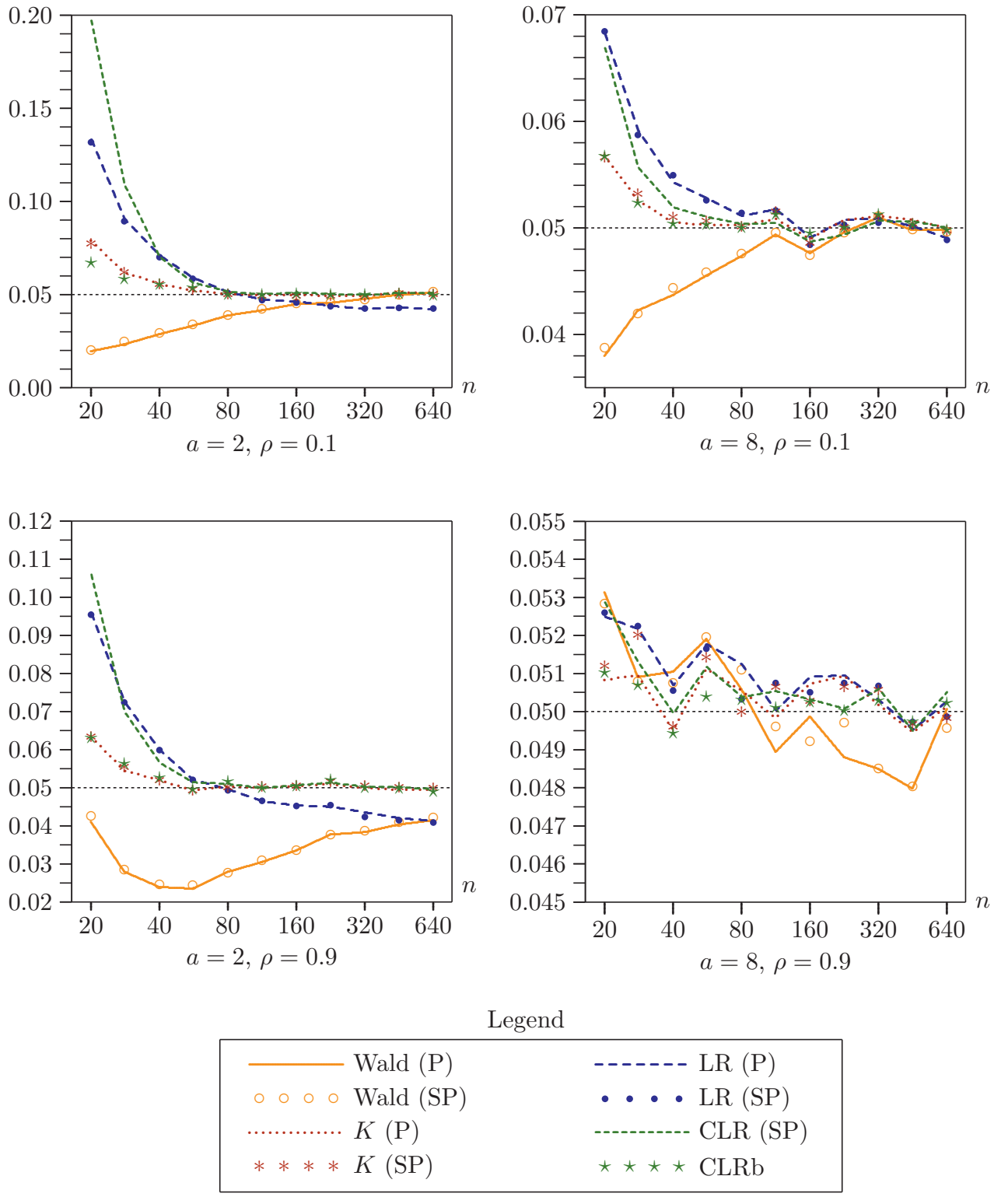

Figure 9. Rejection frequencies for REC bootstrap tests as a function of n for $l - k = 9$

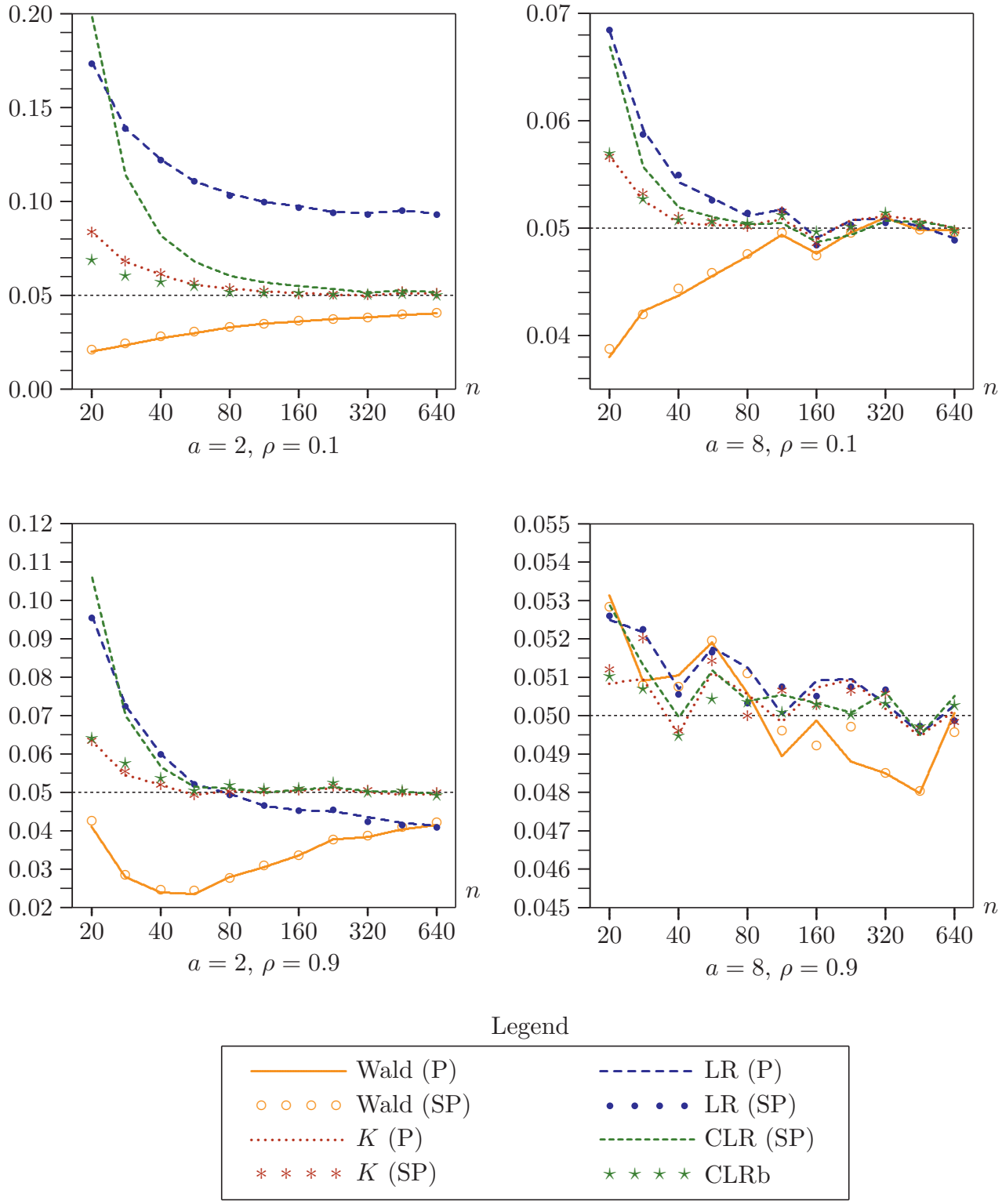

Figure 10. Rejection frequencies for RE bootstrap tests as a function of n for $l - k = 9$

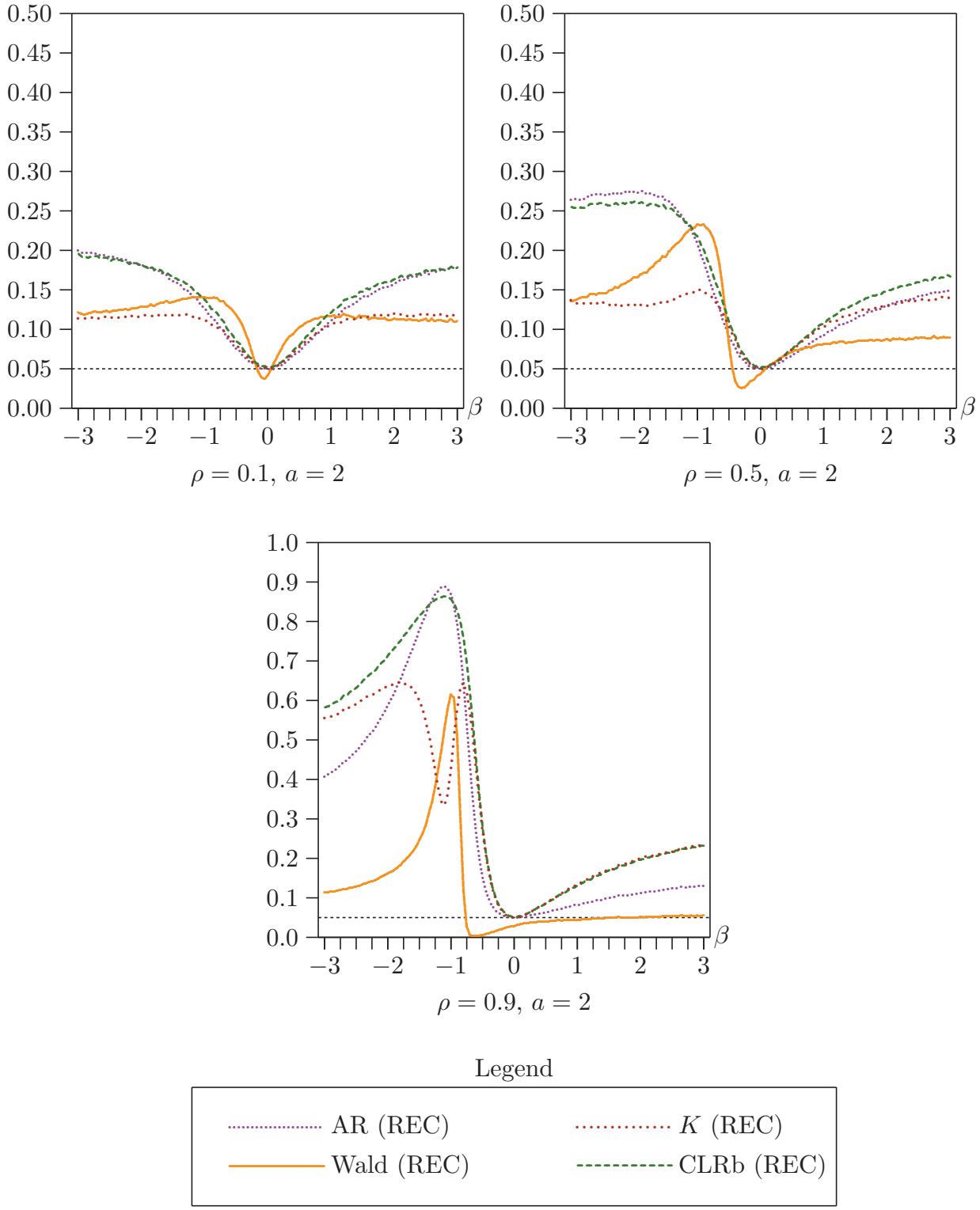

Figure 11. Power of tests when instruments are very weak, for $l - k = 9$, $n = 100$

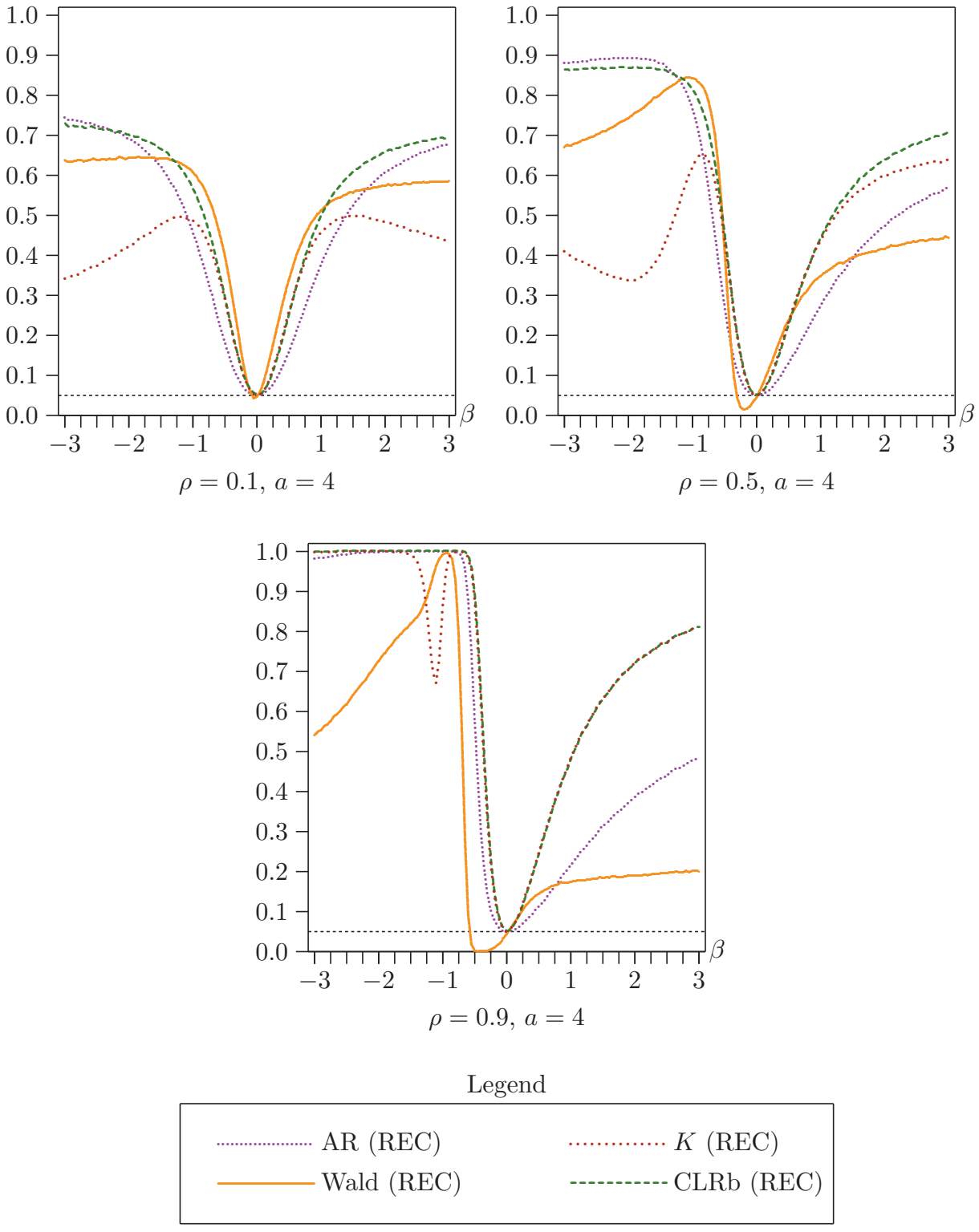

Figure 12. Power of tests when instruments are fairly weak, for $l - k = 9$, $n = 100$

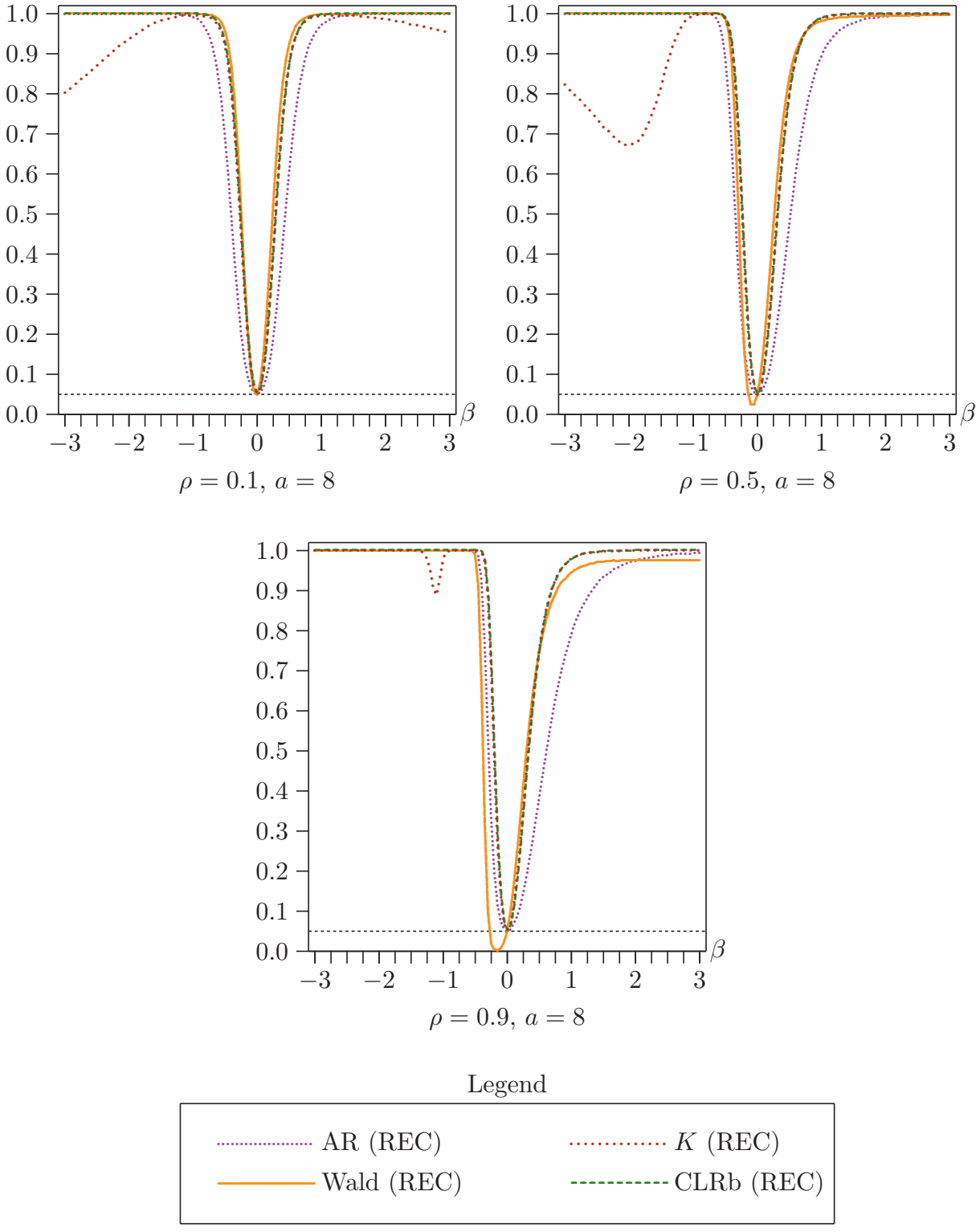

Figure 13. Power of tests when instruments are strong, for $l - k = 9$, $n = 100$