Queen's Economics Department Working Paper No. 1318

Bootstrap tests for overidentification in linear regression models

Russell Davidson
McGill University

James G. MacKinnon
Queen's University

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

4-2014

Bootstrap Tests for Overidentification in Linear Regression Models

Russell Davidson

Department of Economics and CIREQ
McGill University
Montréal, Québec, Canada
H3A 2T7

AMSE-GREQAM
Centre de la Vieille Charité
2 Rue de la Charité
13236 Marseille cedex 02, France

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

Little attention has been paid to the finite-sample properties of tests for overidentifying restrictions in linear regression models with a single endogenous regressor and weak instruments. We study several such tests in models estimated by instrumental variables (IV) and limited-information maximum likelihood (LIML). Under the assumption of Gaussian disturbances, we derive expressions for a variety of test statistics as functions of eight mutually independent random variables and two nuisance parameters. The distributions of the statistics are shown to have an ill-defined limit as the parameter that determines the strength of the instruments tends to zero and as the correlation between the disturbances of the structural and reduced-form equations tends to plus or minus one. Simulation experiments demonstrate that this makes it impossible to perform reliable inference near the point at which the limit is ill-defined. Several bootstrap procedures are proposed. They alleviate the problem and allow reliable inference when the instruments are not too weak. We also study the power properties of the bootstrap tests.

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.

March 2014

1. Introduction

In recent years, there has been a great deal of work on the finite-sample properties of estimators and tests for linear regression models with endogenous regressors when the instruments are weak. Much of this work has focused on the case in which there is just one endogenous variable on the right-hand side, and numerous procedures for testing hypotheses about the coefficient of this variable have been studied. See, among many others, Staiger and Stock (1997), Stock, Wright, and Yogo (2002), Kleibergen (2002), Moreira (2003, 2009), Andrews, Moreira, and Stock (2006), and Davidson and MacKinnon (2008, 2010). However, the closely related problem of testing overidentifying restrictions when the instruments are weak does not appear to have been studied to anything like the same extent.

In the next section, we discuss the famous test of Sargan (1958) and other asymptotic tests for overidentification in linear regression models estimated by instrumental variables (IV) or limited information maximum likelihood (LIML). We show that the test statistics are all functions of six quadratic forms defined in terms of the two endogenous variables of the model, the linear span of the instruments, and its orthogonal complement. In fact, they can be expressed as functions of a certain ratio of sums of squared residuals and are closely related to the test proposed by Anderson and Rubin (1949). In Section 3, we analyze the properties of these overidentification test statistics. We use a simplified model with only three parameters, which is nonetheless capable of generating statistics with exactly the same distributions as those generated by a more general model. In Section 4, we derive the limiting behavior of the statistics in the context of weak-instrument asymptotics as the instrument strength tends to zero, as the correlation between the disturbances in the structural and reduced-form equations tends to unity, and as the sample size tends to infinity.

In Section 5, we investigate by simulation the finite-sample behavior of the statistics we consider. We find that simulation evidence and theoretical analysis concur in strongly preferring a variant of a likelihood-ratio test to the more conventional forms of Sargan test. Section 6 discusses a number of bootstrap procedures that can be used in conjunction with any of the overidentification tests. Some of these procedures are purely parametric, while others make use of resampling. In Section 7, we look at the performance of bootstrap tests, finding that the best of them behave very well if the instruments are not too weak. However, as our theory suggests, they improve very little over tests based on asymptotic critical values in the neighborhood of the singularity that occurs where the instrument strength tends to zero and the correlation of the disturbances tends to one.

In Section 8, we analyze the power properties of the two main variants of bootstrap test. We obtain analytical results that generalize those of Section 3. Using those analytical results, we conduct extensive simulation experiments, mostly for cases that allow the bootstrap to yield reliable inference. We find that bootstrap tests based on IV estimation seem to have a slight power advantage over those based on LIML, at the cost of slightly greater size distortion under the null when the instruments are not too

weak. Section 9 presents a brief discussion of how both test statistics and bootstrap procedures can be modified to take account of heteroskedasticity and clustered data. Finally, some concluding remarks are made in Section 10.

2. Tests for Overidentification

Although the tests for overidentification that we deal with are applicable to linear regression models with any number of endogenous right-hand side variables, we restrict attention in this paper to a model with just one such variable. We do so partly for expositional convenience and partly because this special case is of particular interest and has been the subject of much research in recent years. The model consists of just two equations,

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

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

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 homoskedastic and serially uncorrelated. We assume that $l > k + 1$, so that the model is 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 u_{1i} and u_{2i}:

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

Equation (1) is the structural equation we are interested in, and equation (2) is a reduced-form equation for the second endogenous variable $\boldsymbol{y}_2$.

The model (1) and (2) implicitly involves one identifying restriction, which cannot be tested, and $q \equiv l - k - 1$ overidentifying restrictions. These restrictions say, in effect, that if we append q regressors all belonging to $\mathcal{S}(\boldsymbol{W})$ to equation (1) in such a way that the equation becomes just identified, then the coefficients of these q additional regressors are zero.

The most common way to test the overidentifying restrictions is to use a Sargan test (Sargan, 1958), which can be computed in various ways. The easiest is probably to estimate equation (1) by instrumental variables (IV), using the l columns of $\boldsymbol{W}$ as instruments, and then to regress the IV residuals $\hat{\boldsymbol{u}}_1$ on $\boldsymbol{W}$. The explained sum of squares from this regression divided by the IV estimate of σ_1^2 is the test statistic, and it is asymptotically distributed as $\chi^2(q)$.

The numerator of the Sargan statistic can be written as

$$(\boldsymbol{y}_1 - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}_{\text{IV}} - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2)^\top \boldsymbol{P}_{\boldsymbol{W}} (\boldsymbol{y}_1 - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}_{\text{IV}} - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2), \tag{4}$$

where $\hat{\beta}_{\text{IV}}$ and $\hat{\boldsymbol{\gamma}}_{\text{IV}}$ denote the IV estimates of β and $\boldsymbol{\gamma}$, respectively, and $\boldsymbol{P_W} \equiv \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top$ projects orthogonally into $\mathcal{S}(\boldsymbol{W})$. We define $\boldsymbol{P_Z}$ similarly, and let $\boldsymbol{M_W} \equiv \mathbf{I} - \boldsymbol{P_W}$ and $\boldsymbol{M_Z} \equiv \mathbf{I} - \boldsymbol{P_Z}$. Since $\boldsymbol{Z}$ is orthogonal to the IV residuals,

$$\boldsymbol{y}_1 - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}_{\text{IV}} - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2 = \boldsymbol{M_Z}(\boldsymbol{y}_1 - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}_{\text{IV}} - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2) = \boldsymbol{M_Z}(\boldsymbol{y}_1 - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2).$$

Then, since $\boldsymbol{P_W}\boldsymbol{M_Z} = \boldsymbol{M_Z}\boldsymbol{P_W} = \boldsymbol{P_W} - \boldsymbol{P_Z} = \boldsymbol{M_Z} - \boldsymbol{M_W}$, the numerator of the Sargan statistic can also be written as

$$(\boldsymbol{y}_1 - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2)^\top(\boldsymbol{M_Z} - \boldsymbol{M_W})(\boldsymbol{y}_1 - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2). \tag{5}$$

Similarly, the denominator is just

$$\begin{aligned}\tfrac{1}{n}(\boldsymbol{y}_1 - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}_{\text{IV}} - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2)^\top(\boldsymbol{y}_1 - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}_{\text{IV}} - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2)\\ = \tfrac{1}{n}(\boldsymbol{y}_1 - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2)^\top\boldsymbol{M_Z}(\boldsymbol{y}_1 - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2).\end{aligned} \tag{6}$$

Expression (5) is the numerator of the Anderson-Rubin, or AR, statistic for the hypothesis that $\beta = \hat{\beta}_{\text{IV}}$; see Anderson and Rubin (1949). The denominator of this same AR statistic is

$$\frac{1}{n-l}(\boldsymbol{y}_1 - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2)^\top\boldsymbol{M_W}(\boldsymbol{y}_1 - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2), \tag{7}$$

which may be compared to the second line of (6). We see that the Sargan statistic estimates σ_1^2 under the null hypothesis, and the AR statistic estimates it under the alternative.

Of course, AR statistics are usually calculated for the hypothesis that β takes on a specific value, say β_0, rather than $\hat{\beta}_{\text{IV}}$. Since by definition $\hat{\beta}_{\text{IV}}$ minimizes the numerator (4), it follows that the numerator of the AR statistic is always no smaller than the numerator of the Sargan statistic. Even though the AR statistic is not generally thought of as a test of the overidentifying restrictions, it could be used as such a test, because it will always reject if the restrictions are sufficiently false.

It seems natural to modify the Sargan statistic by using (7) instead of (6) as the denominator, and this was done by Basmann (1960). The usual Sargan statistic can be written as

$$S = \frac{\text{SSR}_0 - \text{SSR}_1}{\text{SSR}_0/n} = n\big(1 - \zeta(\hat{\beta}_{\text{IV}})\big) \tag{8}$$

and the Basmann statistic as

$$S' = \frac{\text{SSR}_0 - \text{SSR}_1}{\text{SSR}_1/(n-l)} = (n-l)\big(\zeta^{-1}(\hat{\beta}_{\text{IV}}) - 1\big), \tag{9}$$

where SSR_0 is the sum of squared residuals from regressing $\boldsymbol{y}_1 - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2$ on $\boldsymbol{Z}$, SSR_1 is the SSR from regressing $\boldsymbol{y}_1 - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2$ on $\boldsymbol{W}$, and $\zeta(\hat{\beta}_{\text{IV}}) \equiv \text{SSR}_1/\text{SSR}_0$. Observe

that both test statistics are simply monotonic functions of $\zeta(\hat{\beta}_{\text{IV}})$, the ratio of the two sums of squared residuals.

Another widely used test statistic for overidentification is the likelihood ratio, or LR, statistic associated with the LIML estimate $\hat{\beta}_{\text{LIML}}$. This statistic is simply $n \log \kappa(\hat{\beta}_{\text{LIML}})$, where

$$\kappa(\beta) = \frac{(\boldsymbol{y}_1 - \beta \boldsymbol{y}_2)^\top \boldsymbol{M}_{\boldsymbol{Z}} (\boldsymbol{y}_1 - \beta \boldsymbol{y}_2)}{(\boldsymbol{y}_1 - \beta \boldsymbol{y}_2)^\top \boldsymbol{M}_{\boldsymbol{W}} (\boldsymbol{y}_1 - \beta \boldsymbol{y}_2)}. \tag{10}$$

The LIML estimator $\hat{\beta}_{\text{LIML}}$ minimizes $\kappa(\beta)$ with respect to β. Since $\hat{\kappa} \equiv \kappa(\hat{\beta}_{\text{LIML}})$ is just $\zeta^{-1}(\hat{\beta}_{\text{LIML}})$, we see that the LR statistic is $-n \log \zeta(\hat{\beta}_{\text{LIML}})$. Conventional asymptotics show that $\hat{\kappa} - 1 = O_p(n^{-1/2})$ as the sample size n tends to infinity. Therefore, the LR statistic is asymptotically equivalent to the linearized likelihood ratio statistic

$$\text{LR}' \equiv (n - l)(\hat{\kappa} - 1) = (n - l)\hat{\lambda} = (n - l)\big(\zeta^{-1}(\hat{\beta}_{\text{LIML}}) - 1\big),$$

where $\hat{\lambda} \equiv \hat{\kappa} - 1$. We define LR$'$ as $(n - l)\hat{\lambda}$ rather than as $n\hat{\lambda}$ by analogy with (9). In what follows, it will be convenient to analyze LR$'$ rather than LR.

We have seen that the Sargan statistic (8), the Basmann statistic (9), and the two likelihood ratio statistics LR and LR$'$ are all monotonic functions of the ratio of SSRs $\zeta(\hat{\beta})$ for some estimator $\hat{\beta}$. Both the particular function of $\zeta(\hat{\beta})$ that is used and the choice of $\hat{\beta}$ affect the finite-sample properties of an asymptotic test. For a bootstrap test, however, it is only the choice of $\hat{\beta}$ that matters. This follows from the fact that it is only the rank of the actual test statistic in the ordered list of the actual and bootstrap statistics that determines a bootstrap P value; see Section 6 below and Davidson and MacKinnon (2006a). Therefore, for any given bootstrap data-generating process (DGP) and any estimator $\hat{\beta}$, bootstrap tests based on any monotonic transformation of $\zeta(\hat{\beta})$ yield identical results.

3. Analysis using a Simpler Model

It is clear from (5), (6), and (10) that all the statistics we have considered for testing the overidentifying restrictions depend on $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ only through their projections $\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{y}_1$ and $\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{y}_2$. We see also that $\zeta(\beta)$ is homogeneous of degree zero with respect to $\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{y}_1$ and $\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{y}_2$ separately, for any β. Thus the statistics depend on the scale of neither $\boldsymbol{y}_1$ nor $\boldsymbol{y}_2$. Moreover, the matrix $\boldsymbol{Z}$ plays no essential role. In fact, it can be shown that the distributions of the test statistics generated by the model (1) and (2) for sample size n are identical to those generated by the simpler model

$$\boldsymbol{y}_1 = \beta \boldsymbol{y}_2 + \boldsymbol{u}_1, \quad \text{and} \tag{11}$$

$$\boldsymbol{y}_2 = \boldsymbol{W}\boldsymbol{\pi} + \boldsymbol{u}_2, \tag{12}$$

where the sample size is $n-k$, the matrix $\boldsymbol{W}$ has $l-k$ columns, and $\sigma_1 = \sigma_2 = 1$. Of course, $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, and $\boldsymbol{W}$ in the simpler model (11) and (12) are not the same as in the original model. In the remainder of the paper, we deal exclusively with the former. For the original model, n and l in our results below would have to be replaced by $n-k$ and $l-k$, and $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ would have to be replaced by $\boldsymbol{M_Z y}_1$ and $\boldsymbol{M_Z y}_2$.

It is well known — see Mariano and Sawa (1972) — that all the test statistics depend on the data generated by (11) and (12) only through the six quadratic forms

$$\begin{aligned} &P_{11} \equiv \boldsymbol{y}_1^\top \boldsymbol{P_W y}_1, \quad P_{12} \equiv \boldsymbol{y}_1^\top \boldsymbol{P_W y}_2, \quad P_{22} \equiv \boldsymbol{y}_2^\top \boldsymbol{P_W y}_2, \\ &M_{11} \equiv \boldsymbol{y}_1^\top \boldsymbol{M_W y}_1, \quad M_{12} \equiv \boldsymbol{y}_1^\top \boldsymbol{M_W y}_2, \text{ and } M_{22} \equiv \boldsymbol{y}_2^\top \boldsymbol{M_W y}_2. \end{aligned} \tag{13}$$

This is also true for the general model (1) and (2), except that $\boldsymbol{P_W}$ must be replaced by $\boldsymbol{P_W} - \boldsymbol{P_Z} = \boldsymbol{P_W M_Z}$.

In this section and the next two, we make the additional assumption that the disturbances $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are normally distributed. Since the quadratic forms in (13) depend on the instruments only through the projections $\boldsymbol{P_W}$ and $\boldsymbol{M_W}$, it follows that their joint distribution depends on $\boldsymbol{W}$ only through the number of instruments l and the norm of the vector $\boldsymbol{W\pi}$. We can therefore further simplify equation (12) as

$$\boldsymbol{y}_2 = a\boldsymbol{w} + \boldsymbol{u}_2, \tag{14}$$

where the vector $\boldsymbol{w} \in \mathcal{S}(\boldsymbol{W})$ is normalized to have length unity, which implies that $a^2 = \boldsymbol{\pi}^\top \boldsymbol{W}^\top \boldsymbol{W\pi}$. Thus the joint distribution of the six quadratic forms depends only on the three parameters β, a, and ρ, and on the dimensions n and l; for the general model (1) and (2), the latter would be $n-k$ and $l-k$.

The above simplification was used in Davidson and MacKinnon (2008) in the context of tests of hypotheses about β, and further details can be found there. The parameter a determines the strength of the instruments. In weak-instrument asymptotics, $a = O(1)$, while in conventional strong-instrument asymptotics, $a = O(n^{1/2})$. Thus, by treating a as a parameter of order unity, we are in the context of weak-instrument asymptotics; see Staiger and Stock (1997). The square of the parameter a is often referred to as the (scalar) concentration parameter; see Phillips (1983, p. 470) and Stock, Wright, and Yogo (2002).

Let $\hat{\boldsymbol{u}}_1^{\text{IV}}$ denote $\boldsymbol{y}_1 - \hat{\beta}_{\text{IV}} \boldsymbol{y}_2$. The Basmann statistic (9) can be expressed as

$$(n-l) \frac{(\hat{\boldsymbol{u}}_1^{\text{IV}})^\top \boldsymbol{P_W} \hat{\boldsymbol{u}}_1^{\text{IV}}}{(\hat{\boldsymbol{u}}_1^{\text{IV}})^\top \boldsymbol{M_W} \hat{\boldsymbol{u}}_1^{\text{IV}}}.$$

The IV estimator $\hat{\beta}_{\text{IV}}$ satisfies the estimating equation $\boldsymbol{y}_2^\top \boldsymbol{P_W}(\boldsymbol{y}_1 - \hat{\beta}_{\text{IV}} \boldsymbol{y}_2) = 0$, so that $\hat{\beta}_{\text{IV}} = P_{12}/P_{22}$. Therefore

$$\begin{aligned} (\hat{\boldsymbol{u}}_1^{\text{IV}})^\top \boldsymbol{P_W} \hat{\boldsymbol{u}}_1^{\text{IV}} &= P_{11} - P_{12}^2/P_{22}, \text{ and} \\ (\hat{\boldsymbol{u}}_1^{\text{IV}})^\top \boldsymbol{M_W} \hat{\boldsymbol{u}}_1^{\text{IV}} &= M_{11} - 2P_{12}M_{12}/P_{22} + P_{12}^2 M_{22}/P_{22}^2. \end{aligned}$$

Thus we find that

$$S' = \frac{(n-l)(P_{11}P_{22} - P_{12}^2)}{M_{11}P_{22} - 2P_{12}M_{12} + P_{12}^2 M_{22}/P_{22}}. \tag{15}$$

Now consider the statistic $\mathrm{LR}' \equiv (n-l)\hat{\lambda} = (n-l)\lambda(\hat{\beta})$, where $\lambda(\beta) = \kappa(\beta) - 1$, with $\kappa(\beta)$ defined in (10). For our simplified model,

$$\lambda(\beta) = \frac{(\boldsymbol{y}_1 - \beta\boldsymbol{y}_2)^\top \boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{y}_1 - \beta\boldsymbol{y}_2)}{(\boldsymbol{y}_1 - \beta\boldsymbol{y}_2)^\top \boldsymbol{M}_{\boldsymbol{W}}(\boldsymbol{y}_1 - \beta\boldsymbol{y}_2)} = \frac{P_{11} - 2\beta P_{12} + \beta^2 P_{22}}{M_{11} - 2\beta M_{12} + \beta^2 M_{22}}. \tag{16}$$

The first-order condition for the minimization of $\lambda(\beta)$ leads to the equation

$$(\beta P_{22} - P_{12})(M_{11} - 2\beta M_{12} + \beta^2 M_{22} = (\beta M_{22} - M_{12})(P_{11} - 2\beta P_{12} + \beta^2 P_{22}).$$

It follows that

$$\hat{\beta}P_{22} - P_{12} = \hat{\lambda}(\hat{\beta}M_{22} - M_{12}),$$

whence

$$\hat{\beta} = \frac{P_{12} - \hat{\lambda}M_{12}}{P_{22} - \hat{\lambda}M_{22}} \quad \text{and} \quad \hat{\lambda} = \frac{P_{12} - \hat{\beta}P_{22}}{M_{12} - \hat{\beta}M_{22}}. \tag{17}$$

The equation above for $\hat{\lambda}$, combined with the definition (16), then shows that

$$\hat{\lambda} = \frac{P_{11} - \hat{\beta}P_{12}}{M_{11} - \hat{\beta}M_{22}} \quad \text{and} \quad \hat{\beta} = \frac{P_{11} - \hat{\lambda}M_{11}}{P_{12} - \hat{\lambda}M_{12}}. \tag{18}$$

By equating the two expressions for $\hat{\beta}$ in (17) and (18), we derive a quadratic equation satisfied by $\hat{\lambda}$, of the form $A\hat{\lambda}^2 - B\hat{\lambda} + C = 0$, where

$$A = M_{11}M_{22} - M_{12}^2, \;\; B = P_{11}M_{22} - 2P_{12}M_{12} + P_{22}M_{11}, \text{ and } C = P_{11}P_{22} - P_{12}^2. \tag{19}$$

Since we seek to minimize $\hat{\lambda}$, it must be the smaller root of this quadratic equation.

Although the quadratic forms (13) depend on the value of β, the statistics themselves do not. We are thus at liberty to set $\beta = 0$ in the subsequent analysis, without loss of generality. To see this, observe that, from (11),

$$P_{11} = \boldsymbol{y}_1^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}_1 = \boldsymbol{u}_1^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{u}_1 + 2\beta \boldsymbol{y}_2^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{u}_1 + \beta^2 P_{22},$$

where, since $\boldsymbol{y}_2$ does not depend on β, all dependence on β in this equation is explicit. By a similar argument,

$$P_{12} = \boldsymbol{y}_2^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{u}_1 + \beta P_{22},$$

and it then follows that

$$P_{11}P_{22} - P_{12}^2 = P_{22}\boldsymbol{u}_1^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{u}_1 - (\boldsymbol{y}_2^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{u}_1)^2,$$

which does not depend on β. Similar calculations show that the denominator in (15) and the coefficients A and B in the equation (19) do not depend on β, and so neither do the statistics S' and LR$'$.

In Davidson and MacKinnon (2008) it is shown that, under the assumption of normal disturbances, the six quadratic forms (13) can be expressed as functions of the three parameters β, a, and ρ and eight mutually independent random variables, the distributions of which do not depend on any of the parameters. Four of these random variables, which we denote by x_1, x_2, z_P, and z_M, are standard normal, and the other four, which we denote by t^P_{11}, t^P_{22}, t^M_{11}, and t^M_{22}, are respectively distributed as χ^2_{l-2}, χ^2_{l-1}, χ^2_{n-l}, and χ^2_{n-l-1}. In terms of these eight variables, we make the definitions

$$\begin{aligned} &Q_{11} \equiv x_1^2 + z_P^2 + t^P_{11}, \quad Q_{12} \equiv x_1 x_2 + z_P \sqrt{t^P_{22}}, \quad Q_{22} \equiv x_2^2 + t^P_{22}, \\ &N_{11} \equiv t^M_{11}, \quad N_{12} \equiv z_M \sqrt{t^M_{11}}, \text{ and } N_{22} \equiv z_M^2 + t^M_{22}. \end{aligned} \tag{20}$$

These quantities have simple interpretations: $Q_{ij} = \boldsymbol{u}_i^\top \boldsymbol{P_W} \boldsymbol{u}_j$, and $N_{ij} = \boldsymbol{u}_i^\top \boldsymbol{M_W} \boldsymbol{u}_j$, for $i = 1, 2$.

When $\beta = 0$, we find that

$$\begin{aligned} &P_{11} = Q_{11}, \quad P_{12} = a x_1 + \rho Q_{11} + r Q_{12}, \quad M_{11} = N_{11}, \\ &P_{22} = a^2 + 2a(\rho x_1 + r x_2) + \rho^2 Q_{11} + 2r\rho Q_{12} + r^2 Q_{22}, \\ &M_{22} = \rho^2 N_{11} + 2r\rho N_{12} + r^2 N_{22}, \text{ and } M_{12} = \rho N_{11} + r N_{12}, \end{aligned} \tag{21}$$

where $r = \sqrt{1 - \rho^2}$. By substituting these relations and those in (20) into (15), realizations of S' can be generated as functions of realizations of the eight independent random variables.

Realizations of LR$'$ can be generated similarly. From the standard formula for the roots of a quadratic equation, we see that

$$\text{LR}' = \frac{(n-l)(P_{11}M_{22} - 2P_{12}M_{12} + P_{22}M_{11} - \Delta^{1/2})}{2(M_{11}M_{22} - M_{12}^2)}, \tag{22}$$

where the discriminant Δ is given by

$$\Delta = (P_{11}M_{22} - 2P_{12}M_{12} + P_{22}M_{11})^2 - 4(M_{11}M_{22} - M_{12}^2)(P_{11}P_{22} - P_{12}^2).$$

4. Limits

In this section, we show that no test of the overidentifying restrictions is robust to weak instruments. In fact, the distributions of S' and LR$'$ have a singularity at the point in the parameter space at which $a = 0$ and $\rho = \pm 1$, or, equivalently, $a = r = 0$. In order to show this, we consider the limits of the expressions (15) and (22), first when $a \to 0$, and then when $r \to 0$. It is also useful to check that the finite-sample expressions have the form given by conventional (strong-instrument) asymptotics when $a \to \infty$ and $n \to \infty$.

We start by looking at the linearized likelihood ratio statistic, LR$'$. Algebra, tedious but easily handled by computer, shows that

$$\begin{aligned}
A &= M_{11}M_{22} - M_{12}^2 = r^2(N_{11}N_{22} - N_{12}^2),\\
C &= P_{11}P_{22} - P_{12}^2 = a^2(Q_{11} - x_1^2) + 2ar(Q_{11}x_2 - Q_{12}x_1) + r^2(Q_{11}Q_{22} - Q_{12}^2),\\
B &= P_{11}M_{22} - 2P_{12}M_{12} + P_{22}M_{11} = a^2N_{11} + 2ar(N_{11}x_2 - N_{12}x_1)\\
&\qquad\qquad + r^2(Q_{11}N_{22} - 2Q_{12}N_{12} + Q_{22}N_{11}),
\end{aligned}$$

where A, B, and C are the coefficients of the quadratic equation (19). It can be seen that these three coefficients are homogeneous of degree two in (a, r), and it then follows from (22) that LR$'$ is homogeneous of degree zero in these two parameters. If we set $a = 0$ in A, B, and C, therefore, these coefficients are all proportional to r^2, and so we can cancel r^2 from the numerator and denominator of (22). The value of LR$'$ for $a = 0$ is then seen to be

$$\frac{(n-l)(Q_{11}N_{22} - 2Q_{12}N_{12} + Q_{22}N_{11} - \Delta^{1/2})}{2(N_{11}N_{22} - N_{12}^2)}, \tag{23}$$

where the discriminant Δ has become

$$(Q_{11}N_{22} - Q_{22}N_{11})^2 + 4(Q_{12}N_{11} - Q_{11}N_{12})(Q_{12}N_{22} - Q_{22}N_{12}).$$

Note that (23) no longer depends on r at all. Thus the distribution of LR$'$ in the limit of completely irrelevant instruments is independent of all the model parameters.

If we set $r = 0$, then both the numerator and denominator of (22) are zero. We therefore must divide both by r^2 before setting $r = 0$. This is simple for the denominator, which is explicitly proportional to r^2. For the numerator, we proceed as follows. Let $B = B_0 + rB_1 + r^2B_2$, with B_0, B_1, and B_2 independent of r. Similarly, let $A = r^2A_2$, and let $C = C_0 + rC_1 + r^2C_2$. Then the numerator of (22) can be written as

$$B_0 + rB_1 + r^2B_2 - \big((B_0 + rB_1 + r^2B_2)^2 - 4r^2A_2(C_0 + rC_1 + r^2C_2)\big)^{1/2} \tag{24}$$

The square root above is

$$(B_0 + rB_1 + r^2B_2)\Big[1 - \frac{4r^2A_2(C_0 + rC_1 + r^2C_2)}{(B_0 + rB_1 + r^2B_2)^2}\Big]^{1/2}.$$

A Taylor expansion of this expression for small r shows that the numerator (24) is

$$\begin{aligned}\frac{2r^2A_2(C_0+rC_1+r^2C_2)}{B_0+rB_1+r^2B_2}+O(r^3) &= r^2\big(2A_2C_0/B_0+O(r)\big)\\ &= 2r^2\big[A_2(Q_{11}-x_1^2)/N_{11}+O(r)\big].\end{aligned}$$

Thus the limit of LR′ when $r \to 0$ is just

$$(n-l)(Q_{11}-x_1^2)/N_{11}. \tag{25}$$

This is independent of a, and it tends to a χ^2_{l-1} variable as $n \to \infty$.

The singularity mentioned above is a consequence of the fact that the limit at $a = r = 0$ is ill-defined, since LR′ converges to two different random variables as $r \to 0$ for $a = 0$ and as $a \to 0$ for $r = 0$. These random variables are quite different and have quite different distributions.

The limit of LR′ as $a \to \infty$, which is the limit when the instruments are strong, can be computed in a similar way, by isolating the coefficients of powers of a rather than those of r and performing a Taylor expansion for small $1/a$. The limit turns out to be, like the limit as $r \to 0$, $n(Q_{11}-x_1^2)/N_{11}$. As $n \to \infty$, $N_{11}/n \to 1$, which shows that the asymptotic distribution with strong instruments is just χ^2_{l-1}.

It would be tedious to go through analogous calculations for the statistic S'. We content ourselves with presenting the results. First, the value of S' for $a = 0$ is

$$(n-l)\frac{(Q_{11}Q_{22}-Q_{12}^2)(\rho^2Q_{11}+2r\rho Q_{12}+r^2Q_{22})}{\rho^2D_0+2r\rho D_1+r^2D_2},$$

where

$$\begin{aligned}D_0 &= Q_{12}^2N_{11}-2Q_{11}Q_{12}N_{12}+Q_{11}^2N_{22},\\ D_1 &= Q_{12}Q_{22}N_{11}-N_{12}(Q_{11}Q_{22}+Q_{12}^2)+Q_{11}Q_{12}N_{22}, \text{ and}\\ D_2 &= Q_{22}^2N_{11}-2Q_{12}Q_{22}N_{12}+Q_{12}^2N_{22}.\end{aligned}$$

This expression does depend on r, unlike the analogous expression for LR′. When $r \to 0$ with $a = 0$, it is easy to see that S' tends to the limit

$$(n-l)\frac{Q_{11}(Q_{11}Q_{22}-Q_{12}^2)}{Q_{12}^2N_{11}-2Q_{11}Q_{12}N_{12}+Q_{11}^2N_{22}}. \tag{26}$$

When $r \to 0$ with $a \neq 0$, the limit of S' is

$$\frac{(n-l)(Q_{11}-x_1^2)(a^2+2ax_1+Q_{11})}{N_{11}(a+x_1)^2}.$$

This does depend on a, and its limit as $a \to 0$ is just

$$(n-l)\frac{Q_{11}(Q_{11}-x_1^2)}{N_{11}x_1^2}, \tag{27}$$

which is quite different from (26), where the order of the limits is inverted. Lastly, as expected, the limit of S' as $a \to \infty$ is the same as that of LR$'$.

The fact that the test statistics S' and LR$'$ depend on the parameters a and ρ indicates that these statistics are not robust to weak instruments. Passing to the limit as $n \to \infty$ with weak-instrument asymptotics does not improve matters. Of the six quadratic forms on which everything depends, only the M_{ij} depend on n. Their limiting behavior is such that $M_{11}/n \to 1$, $M_{22}/n \to 1$, and $M_{12}/n \to \rho$ as $n \to \infty$. But the P_{ij} do not depend on n, and they do depend on a and ρ.

5. Finite-Sample Properties of the Tests

The discussion in the previous section was limited to the statistics S' and LR$'$. When we discuss bootstrap tests, it is enough to consider just these two, since all other statistics mentioned in Section 2 are monotonic transforms of them. But, of course, the different versions of the Sargan test and the LR test have different properties when used with (strong-instrument) asymptotic critical values. In this section, therefore, we present some Monte Carlo results on the finite-sample performance of five test statistics, including the four discussed above (S, S', LR, and LR$'$).

The fifth test statistic we examine is based on the estimator proposed by Fuller (1977). Like the IV and LIML estimators, Fuller's estimator is a K-class estimator for model (1) and (2). It takes the form

$$\begin{bmatrix} \hat{\beta} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} = \big(\boldsymbol{X}^\top(\mathbf{I} - K\boldsymbol{M}_{\boldsymbol{W}})\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top(\mathbf{I} - K\boldsymbol{M}_{\boldsymbol{W}})\boldsymbol{y}_1. \tag{28}$$

Setting $K = \hat{\kappa}$, the minimized value of the variance ratio (10), in equation (28) gives the LIML estimator, while setting $K = 1$ gives the IV estimator. Fuller's estimator sets $K = \hat{\kappa} + \eta/(n - l)$ for some nonrandom number $\eta > 0$ independent of the sample size n. We set $\eta = 1$. With this choice, Fuller's estimator $\hat{\beta}_{\mathrm{F}}$ has all moments (except when the sample size is very small) and is approximately unbiased. The corresponding test statistic is simply $-n \log \zeta(\hat{\beta}_{\mathrm{F}})$, which has the same form as the LR statistic. We will refer to this as the LRF test.

The data-generating processes, or DGPs, used for our simulations all belong to the simplified model (11) and (14). The disturbances are generated according to the relations

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

where $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are n–vectors with independent standard normal elements, and $r \equiv (1 - \rho^2)^{1/2}$. Of course, it is quite unnecessary to generate simulated samples of n observations, as it is enough to generate the six quadratic forms (13) as functions of eight mutually independent random variables, using the relations (20) and (21). The sample size n affects only the degrees of freedom of the two χ^2 random variables t_{11}^M and t_{22}^M that appear in (20). Although any DGP given by (11) and (14) involves no

explicit overidentifying restrictions, the test statistics are computed for the model (1) and (2), for which there are $q \equiv l - k - 1$ of them.

The first group of experiments is intended to provide guidance on the appropriate sample size to use in the remaining experiments. Our objective is to mimic the common situation in which the sample size is reasonably large and the instruments are quite weak. Since the behavior of our simulation DGPs is governed by weak-instrument asymptotics, we should not expect any of the test statistics to have the correct size asymptotically. However, for any given a and ρ, the rejection frequency converges as $n \to \infty$ to that given by the asymptotic distribution of the statistic used; these asymptotic distributions were discussed at the end of the previous section. In the experiments, we use sample sizes of 20, 28, 40, 56, and so on, up to 1810. Each of these numbers is larger than its predecessor by approximately $\sqrt{2}$. Each experiment used 10^6 replications.

The results of four sets of experiments are presented in Figure 1, in which we plot rejection frequencies in the experiments for a nominal level of 5%. In the top two panels, $a = 2$, so that the instruments are very weak. In the bottom two panels, $a = 8$, so that they are reasonably strong. Recall that the concentration parameter is a^2. In the two panels on the left, $\rho = 0.5$, so that there is moderate correlation between the structural and reduced form disturbances. In the two panels on the right, $\rho = 0.9$, so that there is strong correlation. Note that the vertical axis differs across most of the panels.

It is evident that the performance of all the tests varies greatly with the sample size. The Sargan (S) and Basmann (S') tests perform almost the same for large samples but very differently for small ones, with the latter much more prone to overreject than the former. For $a = 2$, the LR test and its linearized version LR$'$ perform quite differently in small samples but almost identically once $n \geq 200$. In this case, the Fuller variant of the LR test performs somewhat differently from both LR and LR$'$ for all sample sizes. In contrast, for $a = 8$, LR and LRF are so similar that we did not graph LR to avoid making the figure unreadable. LR, LR$'$, and LRF perform almost identically, and very well indeed, for large sample sizes, even though they overreject severely for small sample sizes.

As expected, all of the rejection frequencies seem to be converging to constants as $n \to \infty$. Moreover, in every case, it appears that the (interpolated) results for $n = 400$ are very similar to the results for larger values up to $n = 1810$. Accordingly, we used $n = 400$ in all the remaining experiments.

In the second group of experiments, the number of overidentifying restrictions q is varied. The four panels in Figure 2 correspond to those of Figure 1. In most cases, performance deteriorates as q increases. Sometimes, rejection frequencies seem to be converging, but by no means always. In the remaining experiments, we somewhat arbitrarily set $q = 8$. Choosing a smaller number would generally have resulted in smaller size distortions.

In the third group of experiments, the results of which are shown in Figure 3, we set $n = 400$ and $q = 8$, and we vary ρ between 0.0 and 0.99 at intervals of 0.01 for four values of a. The vertical axis is different in each of the four panels, because the tests all perform much better as a increases. For clarity, rejection frequencies for LR are not shown in the figure, because they always lie between those for LR′ and LRF. They are very close to those for LR′ when a is small, and they are very close to those for LRF when a is large.

For the smaller values of a, all of the tests can either overreject or underreject, with rejection frequencies increasing in ρ. The Sargan and Basmann tests overreject very severely when a is small and ρ is large. The LR′, LR, and LRF tests underreject severely when a is small and ρ is not large, but they overreject slightly when a is large. Based on Figure 1 and on the analysis of the previous section, we expect that this slight overrejection vanishes for larger samples.

Although the performance of all the tests is quite poor when a is small, it is worth noting that the Sargan tests are not as unreliable as t tests of the hypothesis that β has a specific value, and the LR tests are not as unreliable as LR tests for that hypothesis; see Davidson and MacKinnon (2008, 2010).

Near the Singularity

From Figures 1–3, we see that the rejection probabilities of all the tests vary considerably with the parameters a and ρ as they vary in the neighborhood of the singularity at $a = 0$, $\rho = 1$. Further insight into this phenomenon is provided by Figures 4 and 5. These are contour plots of rejection frequencies near the singularity for tests at the 0.05 level with a and ρ on the horizontal and vertical axes, respectively. Figure 4 is for the Basmann statistic S', and Figure 5 is for the LR′ statistic. Both figures are for the case dealt with in Figure 3, for which $n = 400$ and $q = 8$. The rejection frequencies are, once again, estimated using 10^6 replications.

It is clear from these figures that rejection frequencies tend to be greatest as the singularity is approached by first setting $r = 0$ and then letting a tend to zero. In this limit, S' is given by expression (27) and LR′ by expression (25). For extremely small values of a, S' actually underrejects. But, as a rises to values that are still very very small, rejection frequencies soar, sometimes to over 0.80. In contrast, LR′ underrejects severely for small values of a, values which do not have to be nearly as small as in the case of S'. In much of the figure, however, the rejection frequencies for LR′ are just a little greater than 0.05.

The 95% quantile of the distribution of expression (27) has the huge value of 16,285, as estimated from 9,999,999 independent realizations. In contrast, recall that the 95% quantile of the χ^2_q distribution for $q = 8$ is 15.5073. Since the distribution of S' for arbitrary a and ρ is stochastically bounded by that of (27), S' is boundedly pivotal. However, basing inference on the distribution of (27) is certain to be *extremely* conservative.

6. Bootstrap Tests

Every test statistic has a distribution which depends on the DGP that generated the sample from which it is computed. The "true" DGP that generated an observed realization of the statistic is in general unknown. However, according to the bootstrap principle, one can perform inference by replacing the unknown DGP by an estimate of it, which is called the bootstrap DGP. Because what we need for inference is the distribution of the statistic under DGPs that satisfy the null hypothesis, the bootstrap DGP must necessarily impose the null. This requirement by itself does not normally lead to a unique bootstrap DGP, and we will see in this section that, for an overidentification test, there are several plausible choices.

If the observed value of a test statistic τ is $\hat{\tau}$, and the rejection region is in the upper tail, then the bootstrap P value is the probability, under the bootstrap distribution of the statistic, that τ is greater than $\hat{\tau}$. To estimate this probability, one generates a large number, say B, of realizations of the statistic using the bootstrap DGP. Let the j^{th} realization be denoted by τ_j^*. Then the simulation-based estimate of the bootstrap P value is just the proportion of the τ_j^* greater than $\hat{\tau}$:

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

where $\mathrm{I}(\cdot)$ is the indicator function, equal to 1 when its argument is true and 0 otherwise. If this fraction is smaller than α, the level of the test, then we reject the null hypothesis. See Davidson and MacKinnon (2006a).

Parametric Bootstraps

The DGPs contained in the simple model defined by equations (11) and (14) are characterized by just three parameters, namely, β, a, and ρ, Since the value of β does not affect the distribution of the overidentification test statistics, the bootstrap DGP for a parametric bootstrap (assuming normally distributed disturbances) is completely determined by the values of a and ρ that characterize it.

The test statistic $\hat{\tau}$ itself may be any one of the overidentification statistics we have discussed. The model that is actually estimated in order to obtain $\hat{\tau}$ is not the simple model, but rather the full model given by (1) and (2). The parameters of this model include some whose values do not interest us for the purpose of defining a bootstrap DGP: β, since it has no effect on the distribution of the statistic, and $\boldsymbol{\gamma}$, since the matrix $\boldsymbol{Z}$ plays no role in the simple model, from which the bootstrap DGP is taken. There remain $\boldsymbol{\pi}$, ρ, σ_1, and σ_2.

For equation (14), the parameter a was defined as the square root of $\boldsymbol{\pi}^\top \boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{\pi}$, but that definition assumes that the vector $\boldsymbol{w}$ has unit length, and that all the variables

are scaled so that the variance of the disturbances $\boldsymbol{u}_2$ is 1. In order to take account of these facts, a suitable definition of a is

$$a = \sqrt{\boldsymbol{\pi}^\top \boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{\pi} / \sigma_2^2}. \tag{29}$$

It follows from (29) that, in order to estimate a, it is necessary also to estimate σ_2^2.

Since the parameter ρ is the correlation of the disturbances, which are not observed, any estimate of ρ must be based on the residuals from the estimation of equations (1) and (2). Let these residuals be denoted by $\ddot{\boldsymbol{u}}_1$ and $\ddot{\boldsymbol{u}}_2$. Then the obvious estimators of the parameters of the covariance matrix are

$$\ddot{\sigma}_1^2 = n^{-1} \ddot{\boldsymbol{u}}_1^\top \ddot{\boldsymbol{u}}_1, \quad \ddot{\sigma}_2^2 = n^{-1} \ddot{\boldsymbol{u}}_2^\top \ddot{\boldsymbol{u}}_2, \quad \text{and} \quad \ddot{\rho} = n^{-1} \ddot{\boldsymbol{u}}_1^\top \ddot{\boldsymbol{u}}_2 / (\ddot{\sigma}_1^2 \ddot{\sigma}_2^2)^{1/2},$$

and the obvious estimator of a is given by

$$\ddot{a}^2 = n \ddot{\boldsymbol{\pi}}^\top \boldsymbol{W}^\top \boldsymbol{W}^\top \ddot{\boldsymbol{\pi}} / \ddot{\boldsymbol{u}}_2^\top \ddot{\boldsymbol{u}}_2, \tag{30}$$

where $\ddot{\boldsymbol{\pi}}$ estimates $\boldsymbol{\pi}$. For $\ddot{\boldsymbol{u}}_1$, there are two obvious choices, the IV residuals and the LIML residuals from (1). For $\ddot{\boldsymbol{u}}_2$, the obvious choice is the vector of OLS residuals from (2), possibly scaled by a factor of $(n/(n-l))^{1/2}$ to take account of the lost degrees of freedom in the OLS estimation. However, this obvious choice is not the only one, because, if we treat the model (1) and (2) as a system, the system estimator of $\boldsymbol{\pi}$ that comes with the IV estimator of β is the three-stage least squares (3SLS) estimator, and the one that comes with the LIML estimator of β is the full-information maximum likelihood (FIML) estimator. These system estimators give rise to estimators not only of $\boldsymbol{\pi}$, but also of $\boldsymbol{u}_2$, that differ from those given by OLS.

The system estimators of $\boldsymbol{\pi}$ can be computed without actually performing a system estimation, by running the regression

$$\boldsymbol{y}_2 = \boldsymbol{W}\boldsymbol{\pi} + \varphi \ddot{\boldsymbol{u}}_1 + \text{residuals}; \tag{31}$$

see Davidson and MacKinnon (2008), where this matter is discussed in greater detail. If $\ddot{\boldsymbol{u}}_1$ is the vector of IV residuals, then the corresponding estimator $\ddot{\boldsymbol{\pi}}$ is the 3SLS estimator; if it is the vector of LIML residuals, then $\ddot{\boldsymbol{\pi}}$ is the FIML estimator.

For the purpose of computation, it is worth noting that all these estimators can be expressed as functions of the six quadratic forms (13). A short calculation shows that the estimators of a^2 and ρ based on IV residuals, scaled OLS residuals, and the OLS estimator of $\boldsymbol{\pi}$ are

$$\hat{a}^2 = (n-l)\frac{P_{22}}{M_{22}} \quad \text{and} \quad \hat{\rho} = \frac{n^{-1}(M_{12} - \hat{b}M_{22})}{\hat{\sigma}_1 \hat{\sigma}_2}, \tag{32}$$

where $\hat{b} = P_{12}/P_{22}$ is the difference between the IV estimator of β and the true β of the DGP,

$$\hat{\sigma}_1^2 = n^{-1}\big(Q_{11} + N_{11} - 2\hat{b}(P_{12} + M_{12}) + \hat{b}^2(P_{22} + M_{22})\big), \text{ and } \hat{\sigma}_2^2 = M_{22}/(n-l).$$

From (20), we can express $\hat{a}^2$ as

$$\frac{(n-l)(a^2 + 2a(\rho x_1 + r x_2) + \rho^2 Q_{11} + 2r\rho Q_{12} + r^2 Q_{22}}{\rho^2 N_{11} + 2r\rho N_{12} + r^2 N_{22}}.$$

The weak-instrument asymptotic limit of this expression replaces the denominator divided by $n - l$ by 1. The expectation of the numerator without the factor of $n - l$ is $a^2 + \rho^2 l + r^2 l = a^2 + l$. Consequently, it may be preferable to reduce bias in the estimation of a^2 by setting $\hat{a}^2 = (n - l)\max(0, P_{22}/M_{22} - l)$; see Davidson and MacKinnon (2008).

The system estimator of the vector $\boldsymbol{\pi}$, which we write as $\ddot{\boldsymbol{\pi}}$, is computed by running regression (31). A tedious calculation shows that

$$\ddot{\boldsymbol{\pi}}^\top \boldsymbol{W}^\top \boldsymbol{W} \ddot{\boldsymbol{\pi}} = P_{22} + \frac{\ddot{P}_{11}\ddot{M}_{12}^2}{\ddot{M}_{11}^2} - 2\frac{\ddot{P}_{12}\ddot{M}_{12}}{\ddot{M}_{11}}, \tag{33}$$

with

$$\begin{aligned}
\ddot{P}_{11} &= \ddot{\boldsymbol{u}}_1^\top \boldsymbol{P}_{\boldsymbol{W}} \ddot{\boldsymbol{u}}_1 = P_{11} - 2\ddot{b}P_{12} + \ddot{b}^2 P_{22}, \\
\ddot{P}_{12} &= \ddot{\boldsymbol{u}}_1^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}_2 = P_{12} - \ddot{b}P_{22}, \\
\ddot{M}_{11} &= \ddot{\boldsymbol{u}}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \ddot{\boldsymbol{u}}_1 = M_{11} - 2\ddot{b}M_{12} + \ddot{b}^2 M_{22}, \text{ and} \\
\ddot{M}_{12} &= \ddot{\boldsymbol{u}}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2 = M_{12} - \ddot{b}M_{22},
\end{aligned}$$

where $\ddot{b}$ is $\hat{b} = \hat{\beta}_{\text{IV}} - \beta$ and $\ddot{\boldsymbol{u}}_1$ is the vector of IV residuals if the structural equation (1) is estimated by IV, and $\ddot{b}$ is $\tilde{b} = \hat{\beta}_{\text{LIML}} - \beta$ and $\ddot{\boldsymbol{u}}_1$ is the vector of LIML residuals if (1) is estimated by LIML. Then, if σ_2^2 is estimated using the residuals $\boldsymbol{y}_2 - \boldsymbol{W}\ddot{\boldsymbol{\pi}}$, we find that

$$\ddot{\sigma}_2^2 = \frac{1}{n-l}\left(M_{22} + \frac{\ddot{M}_{12}^2}{\ddot{M}_{11}^2}\ddot{P}_{11}\right). \tag{34}$$

Using the results (33) and (34) in (30) allows us to write

$$\ddot{a}^2 = \frac{(n-l)(P_{22}\ddot{M}_{11}^2 + \ddot{P}_{11}\ddot{M}_{12}^2 - 2\ddot{P}_{12}\ddot{M}_{11}\ddot{M}_{12})}{\ddot{M}_{11}^2 M_{22} + \ddot{M}_{12}^2 \ddot{P}_{11}}.$$

For the parameter ρ, more calculation shows that

$$\ddot{\rho} = \ddot{M}_{12}\left[\frac{\ddot{P}_{11} + \ddot{M}_{11}}{\ddot{M}_{11}^2 M_{22} + \ddot{M}_{12}^2 \ddot{P}_{11}}\right]^{1/2}.$$

We consider four different bootstrap DGPs. The simplest, which we call the IV-R bootstrap, uses the IV and OLS estimates of a and ρ given in (32). The IV-ER bootstrap is based on 3SLS estimation of the two-equation system, that is, on IV estimation of (1), and on regression (31) with $\ddot{\boldsymbol{u}}_1$ the vector of IV residuals. Similarly,

the LIML-ER bootstrap relies on FIML estimation of the system, that is, on LIML for (1), and on (31) with $\ddot{\boldsymbol{u}}_1$ the LIML residuals. Finally, we also define the F(1)-ER bootstrap, which is the same as LIML-ER except that $\hat{\beta}_{\text{LIML}}$ is replaced at every step by Fuller's modified LIML estimator with $\eta = 1$.

It is plain that, the closer the bootstrap DGP to the true DGP, the better will be bootstrap inference; see Davidson and MacKinnon (1999). We may therefore expect that IV-ER should perform better than IV-R, and that LIML-ER should perform better than IV-ER. Between LIML-ER and F(1)-ER, there is no obvious reason *a priori* to expect that one of them would outperform the other. But, whatever the properties of these bootstraps may be when the true DGP is not in the neighborhood of the singularity at $a = 0$, $\rho = 1$, we cannot expect anything better than some improvement over inference based on asymptotic critical values, rather than truly reliable inference, in the neighborhood of the singularity.

Resampling

Any parametric bootstrap risks being unreliable if the strong assumptions used to define the null hypothesis are violated. Most practitioners would therefore prefer a more robust bootstrap method. The strongest assumption we have made so far is that the disturbances are normally distributed. It is easy to relax this assumption by using a bootstrap DGP based on resampling, in which the bivariate normal distribution is replaced by the joint empirical distribution of the residuals. The discussion of the previous subsection makes it clear that several resampling bootstraps can be defined, depending on the choice of residuals that are resampled.

The most obvious resampling bootstrap DGP in the context of IV estimation is

$$\boldsymbol{y}_1^* = \hat{\beta}_{\text{IV}} \boldsymbol{y}_2^* + \boldsymbol{Z}\hat{\boldsymbol{\gamma}}_{\text{IV}} + \hat{\boldsymbol{u}}_1^* \tag{35}$$

$$\boldsymbol{y}_2^* = \boldsymbol{W}\hat{\boldsymbol{\pi}} + \hat{\boldsymbol{u}}_2^*, \tag{36}$$

where $\boldsymbol{y}_1^*$ and $\boldsymbol{y}_2^*$ are n–vectors of bootstrap observations, $\hat{\boldsymbol{u}}_1^*$ and $\hat{\boldsymbol{u}}_2^*$ are n–vectors of bootstrap disturbances with typical elements $\hat{u}_{1i}^*$ and $\hat{u}_{2i}^*$, respectively, and $\hat{\boldsymbol{\pi}}$ is the OLS estimate from (2). The bootstrap disturbances are drawn in pairs from the bivariate empirical distribution of the structural residuals $\hat{u}_{1i}^{\text{IV}}$ and the rescaled reduced-form residuals $\big(n/(n-l)\big)^{1/2}\hat{u}_{2i}^{\text{OLS}}$:

$$\begin{bmatrix} \hat{u}_{1i}^* \\ \hat{u}_{2i}^* \end{bmatrix} \sim \text{EDF}\left(\begin{matrix} \hat{u}_{1i}^{\text{IV}} \\ \big(n/(n-l)\big)^{1/2}\hat{u}_{2i}^{\text{OLS}} \end{matrix} \right). \tag{37}$$

Here EDF stands for "empirical distribution function". The rescaling of the reduced form residuals $\hat{u}_{2i}^{\text{OLS}}$ ensures that the distribution of the $\hat{u}_{2i}^*$ has variance equal to the unbiased OLS variance estimator.

Since all of the overidentification test statistics are invariant to the values of β and $\boldsymbol{\gamma}$, we may replace the bootstrap DGP for $\boldsymbol{y}_1^*$ given by (35) by

$$\boldsymbol{y}_1^* = \hat{\boldsymbol{u}}_1^*. \tag{38}$$

The bootstrap statistics generated by (38) and (36) are identical to those generated by (35) and (36). We will refer to the bootstrap DGP given by (38), (36), and (37) as the IV-R resampling bootstrap. It is a semiparametric bootstrap, because it uses parameter estimates of the reduced-form equation, but it does not assume a specific functional form for the joint distribution of the disturbances. The empirical distribution of the residuals has a covariance matrix which is exactly that used to estimate a and ρ by the IV-R parametric bootstrap; hence our nomenclature.

The IV-ER resampling bootstrap draws pairs from the joint EDF of the IV residuals $\hat{\boldsymbol{u}}_1^{\text{IV}}$ from equation (1) and the residuals $\boldsymbol{y}_2 - \boldsymbol{W}\hat{\boldsymbol{\pi}}_{\text{IV}}$ computed by running regression (31) with $\hat{\boldsymbol{u}}_1^{\text{IV}}$ replacing $\ddot{\boldsymbol{u}}_1$. It also uses the resulting estimator $\hat{\boldsymbol{\pi}}_{\text{IV}}$ in (36) instead of the OLS estimator $\hat{\boldsymbol{\pi}}$. Note that the residuals $\boldsymbol{y}_2 - \boldsymbol{W}\hat{\boldsymbol{\pi}}_{\text{IV}}$ are *not* the residuals from (31), but rather those residuals plus $\ddot{\varphi}\hat{\boldsymbol{u}}_1^{\text{IV}}$.

The LIML-ER resampling bootstrap is very similar to the IV-ER one, except that it uses $\hat{\boldsymbol{u}}_1^{\text{LIML}}$ both directly and in regression (31). Formally, the resampling draws pairs from the bivariate empirical distribution of

$$\begin{bmatrix} \hat{u}_{i1}^{\text{LIML}} \\ \hat{u}_{i2}^{\text{LIML}} \end{bmatrix} = \begin{bmatrix} y_{i1} - \hat{\beta}_{\text{LIML}} y_{2i} - \boldsymbol{Z}_i \hat{\boldsymbol{\gamma}}_{\text{LIML}} \\ y_{i2} - \boldsymbol{W}_i \hat{\boldsymbol{\pi}}_{\text{LIML}} \end{bmatrix}. \tag{39}$$

Similarly, for the F(1)-ER resampling bootstrap, the structural equation (1) is estimated by Fuller's estimator with $\eta = 1$, and the residuals from this used both for resampling and in the regression (31).

A word of caution is advisable here. Although the values of overidentification test statistics are invariant to β, thereby allowing us to use (38) instead of (35) in the bootstrap DGP, the residuals from which we resample in (37) and (39) do depend on the estimate of β, as does the estimate of $\boldsymbol{\pi}$ if it is based on any variant of equation (31). But the test statistics depend on the estimate of β only through the residuals and the estimate of $\boldsymbol{\pi}$.

7. Performance of Bootstrap Tests

In principle, any of the bootstrap DGPs discussed in the previous section can be combined with any of the test statistics discussed in Section 2. However, there is no point considering both S and S', or both LR and LR$'$, because in each case one test statistic is simply a monotonic transformation of the other. If both the statistics in each pair are bootstrapped using the same bootstrap DGP, they must therefore yield identical results.

All of our experiments involve 100,000 replications for each set of parameter values, and the bootstrap tests mostly use $B = 399$. This is a smaller number than should generally be used in practice, but it is perfectly satisfactory for simulation experiments, because experimental randomness in the bootstrap P values tends to average out across replications. Although the disturbances of the true DGPs are taken to be

normally distributed, the bootstrap DGPs we investigate in the main experiments are resampling ones, because we believe they are the ones that will be used in practice.

Figures 6, 7, and 8 present the results of a large number of Monte Carlo experiments. Figure 6 concerns Sargan tests, Figure 7 concerns LR tests, and Figure 8 concerns Fuller LR tests. Each of the figures shows rejection frequencies as a function of ρ for 34 values of ρ, namely, 0.00, 0.03, 0.06, ..., 0.99. The four panels correspond to $a = 2$, 4, 6, and 8. Note that the scale of the vertical axis often differs across panels within each figure and across figures for panels corresponding to the same value of a. It is important to keep this in mind when interpreting the results.

As we have already seen, for small and moderate values of a, Sargan tests tend to overreject severely when ρ is large and to underreject modestly when it is small. It is evident from Figure 6 that, for $a = 2$, using either the IV-R or IV-ER bootstrap improves matters only slightly. However, both these methods do provide a more and more noticeable improvement as a increases. For $a = 8$, the improvement is very substantial. If we were increasing n as well as a, it would be natural to see this as evidence of an asymptotic refinement.

There seems to be no advantage to using IV-ER rather than IV-R. In fact, the latter always works a bit better when ρ is very large. This result is surprising in the light of the findings of Davidson and MacKinnon (2008, 2010) for bootstrapping t tests on β. However, the bootstrap methods considered in those papers imposed the null hypothesis that $\beta = \beta_0$, while the ones considered here do not. Apparently, this makes a difference.

Using the LIML-ER and F(1)-ER bootstraps with the Sargan statistic yields entirely different results. The former underrejects very severely for all values of ρ when a is small, but the extent of the underrejection drops rapidly as a increases. The latter always underrejects less severely than LIML-ER (it actually overrejects for large values of ρ when $a = 2$), and it performs surprisingly well for $a \geq 6$. Of course, it may seem a bit strange to bootstrap a test statistic based on IV estimation using a bootstrap DGP based on LIML or its Fuller variant.

In Figure 7, we see that, in contrast to the Sargan test, the LR test generally underrejects, often very severely when both ρ and a are small. Its performance improves rapidly as a increases, however, and it actually overrejects slightly when ρ and a are both large. All of the bootstrap methods improve matters, and the extent of the improvement increases with a. For $a = 8$, all the bootstrap methods work essentially perfectly. For small values of a, the IV-R bootstrap actually seems to be the best in many cases, although it does lead to modest overrejection when ρ is large.

In Figure 8, we see that the Fuller LR test never underrejects as much as the LR test, and it actually overrejects quite severely when ρ is large and $a = 2$. However, that is the only case in which it overrejects much. This is the only test for which its own bootstrap DGP, namely, F(1)-ER, is arguably the best one to use. Except when the asymptotic test already works perfectly, using that bootstrap method almost always

improves the performance of the test. The bottom two panels of Figure 8 look very similar to the corresponding panels of Figure 7, except that the bootstrapped Fuller test tends to underreject just a bit. It is evident that, as a increases, the LR test and its Fuller variant become almost indistinguishable.

Figures 6, 7, and 8 provide no clear ranking of tests and bootstrap methods. There seems to be a preference for the LR and Fuller LR tests, and for the LIML-ER and F(1)-ER bootstrap DGPs. In no case does any combination of those tests and those bootstrap DGPs overreject anything like as severely as the Sargan test bootstrapped using IV-R or IV-ER. Provided the instruments are not very weak, any of these combinations should yield reasonably accurate, but perhaps somewhat conservative, inferences in most cases.

The rather mixed performance of the bootstrap tests can be understood by using the concept of "bootstrap discrepancy," which is a function of the nominal level of the test, say α. The bootstrap discrepancy is simply the actual rejection rate for a bootstrap test at level α minus α itself. Davidson and MacKinnon (2006b) shows that the bootstrap discrepancy at level α is a conditional expectation of the random variable

$$q(\alpha) \equiv R\big(Q(\alpha, \mu^*), \mu\big) - \alpha, \tag{40}$$

where $R(\alpha, \mu)$ is the probability, under the DGP μ, that the test statistic is in the rejection region for nominal level α, and $Q(\alpha, \mu)$ is the inverse function that satisfies the equation

$$R\big(Q(\alpha, \mu), \mu\big) = \alpha = Q\big(R(\alpha, \mu), \mu\big).$$

Thus $Q(\alpha, \mu)$ is the true level-α critical value of the asymptotic test under μ. The random element in (40) is μ^*, the bootstrap DGP. If $\mu^* = \mu$, then we see clearly that $q(\alpha) = 0$, and the bootstrap discrepancy vanishes. For more detail, see Davidson and MacKinnon (2006b).

Suppose now that the true DGP μ_0 is near the singularity. The bootstrap DGP can reasonably be expected also to be near the singularity, but most realizations are likely to be farther away from the singularity than μ_0 itself. If μ_0 were actually at the singularity, then any bootstrap DGP would necessarily be farther away. If the statistic used is S, then we see from Figure 4 that rejection frequencies fall as the DGP moves away from the singularity in most, but not all, directions. Thus, for most such bootstrap DGPs, $Q(\alpha, \mu^*)$ is smaller than $Q(\alpha, \mu_0)$ for any α, and so the probability mass $R\big(Q(\alpha, \mu^*), \mu_0\big)$ in the distribution generated by μ_0 is greater than α. This means that $q(\alpha)$ is positive, and so the bootstrap test overrejects. However, if the statistic used is LR, the reverse is the case, as we see from Figure 5, and the bootstrap test underrejects. This is just what we see in Figures 6 through 8.

Figures 9 and 10 are contour plots similar to Figures 4 and 5, but they are for bootstrap rather than asymptotic tests. The IV-R parametric bootstrap is used for the Sargan test in Figure 9, and the LIML-ER parametric bootstrap is used for the LR test in Figure 10. In both cases, there are 100,000 replications, and $B = 199$. Figure 9 looks

remarkably like Figure 4, with low rejection frequencies for extremely small values of a, then a ridge where rejection frequencies are very high for slightly larger values of a. The ridge is not quite as high as the one in Figure 4, and the rejection frequencies diminish more rapidly as a increases.

Similarly, Figure 10 looks like Figure 5, but the severe underrejection in the far left of the figure occurs over an even smaller region, and there is an area of modest overrejection nearby. Both these size distortions can be explained by Figure 5. When a is extremely small, the estimate used by the bootstrap DGP tends on average to be larger, so the bootstrap critical values tend, on average, to be overestimates. This leads to underrejection. However, there is a region where a is not quite so small in which the bootstrap DGP uses estimates of a that are sometimes too small and sometimes too large. The former causes overrejection, the latter underrejection. Because of the curvature of the rejection probability function, the net effect is modest overrejection; see Davidson and MacKinnon (1999). This is actually the case for most of the parameter values shown in the figure, but the rejection frequencies are generally not much greater than 0.05.

8. Power Considerations

Overidentification tests are performed in order to check whether some of the assumptions for the two-equation model (1) and (2) to be correctly specified are valid. Those assumptions are not valid if the DGP for equation (1) is actually

$$\boldsymbol{y}_1 = \boldsymbol{Z}\boldsymbol{\gamma}_1 + \boldsymbol{W}_1\boldsymbol{\delta} + \beta\boldsymbol{y}_2 + \boldsymbol{u}_1, \tag{41}$$

where the columns of the matrix $\boldsymbol{W}_1$ are in the span of the columns of the matrix $\boldsymbol{W}$ and are linearly independent of those of $\boldsymbol{Z}$. As in Section 3, we can eliminate $\boldsymbol{Z}$ from the model, replacing all other variables and the disturbances by their projections onto the orthogonal complement of the span of the columns of $\boldsymbol{Z}$. The simpler model of equations (11) and (14) becomes

$$\boldsymbol{y}_1 = \beta\boldsymbol{y}_2 + \delta\,\boldsymbol{w}_p + \boldsymbol{v}_1 \tag{42}$$

$$\boldsymbol{y}_2 = a\boldsymbol{w}_1 + \boldsymbol{u}_2 = a\boldsymbol{w}_1 + \rho\boldsymbol{v}_1 + r\boldsymbol{v}_2. \tag{43}$$

The vector $\boldsymbol{W}\boldsymbol{\pi}$ is now written as $a\boldsymbol{w}_1$ instead of $a\boldsymbol{w}$, and the vector $\boldsymbol{W}_1\boldsymbol{\delta}$ is written as $\delta\,\boldsymbol{w}_p$. As before, we make the normalizations that $\|\boldsymbol{w}_1\|^2 = 1$ and $a^2 = \boldsymbol{\pi}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{\pi}$. In addition, we normalize so that $\|\boldsymbol{w}_p\|^2 = 1$ and $\delta^2 = \boldsymbol{\delta}^\top\boldsymbol{W}_1^\top\boldsymbol{W}_1\boldsymbol{\delta}$.

The disturbance vector $\boldsymbol{u}_1$ of (11) is written as $\boldsymbol{v}_1$. In the rightmost expression of equations (43), the vector $\boldsymbol{u}_2$ has been replaced by $\rho\boldsymbol{v}_1 + r\boldsymbol{v}_2$, where $r \equiv (1-\rho^2)^{1/2}$. The vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are independent $\mathrm{N}(\boldsymbol{0}, \mathbf{I})$. Further, since $\boldsymbol{w}_1$ and $\boldsymbol{w}_p$ are not in general orthogonal, we write

$$\boldsymbol{w}_p = \theta\boldsymbol{w}_1 + t\boldsymbol{w}_2,$$

where $\boldsymbol{w}_2^\top\boldsymbol{w}_1 = 0$, and $t = (1-\theta^2)^{1/2}$.

The Basmann statistic S' is still given by equation (15), which is simply an algebraic consequence of the definition (9). Since the DGP for $\boldsymbol{y}_2$ is unchanged, the quantities denoted in (9) by P_{22} and M_{22} are the same under the alternative as under the null. Since the DGP for $\boldsymbol{M_W y}_1$ is also the same under the null and the alternative, so are M_{11} and M_{12}. Thus only P_{11} and P_{12} differ from the expressions for them in equations (21). It is easy to check that neither the numerator nor the denominator of S' in (15) depends on β under the alternative, and so in our computations we set $\beta = 0$ without loss of generality.

In order to analyze the asymptotic power of the Sargan test in Basmann form, we seek to express its limiting asymptotic distribution as a chi-squared variable that is non-central under the alternative. As usual, in order for the non-centrality parameter (NCP) to have a finite limit, we invoke a Pitman drift. With our normalization of $\boldsymbol{w}_p$, this just means that δ is constant as the sample size n tends to infinity. Again, we cannot expect to find a limiting chi-squared distribution with weak-instrument asymptotics, and so our asymptotic construction supposes that $a \to \infty$ as $n \to \infty$.

Under the null and the alternative, the denominator of (15), divided by $(n-l)P_{22}$, is simply an estimate of the variance of $\boldsymbol{v}_1$. For the purposes of the asymptotic analysis of the simpler model, it can therefore be replaced by 1. The quantity of which the limiting distribution is expected to be chi-squared is therefore $P_{11} - P_{12}^2/P_{22}$. Recall that this is just the numerator of both the S and S' statistics.

With $\beta = 0$, we compute as follows:

$$\begin{aligned}
P_{11} &= \boldsymbol{y}_1^\top \boldsymbol{P_W y}_1 = \delta^2 + 2\delta\theta x_1 + 2\delta t z_1 + \boldsymbol{v}_1^\top \boldsymbol{P_W v}_1, \\
P_{12} &= \boldsymbol{y}_1^\top \boldsymbol{P_W y}_2 = a(x_1 + \delta\theta) + O_p(1), \text{ and} \\
P_{22} &= \boldsymbol{y}_2^\top \boldsymbol{P_W y}_2 = a^2 + 2a(\rho x_1 + r x_2) + O_p(1),
\end{aligned}$$

where the symbol $O_p(1)$ means of order unity as $a \to \infty$. Also, $z_i = \boldsymbol{w}_2^\top \boldsymbol{v}_i$ and, as before, $x_i = \boldsymbol{w}_1^\top \boldsymbol{v}_i$, $i = 1, 2$. Thus the limit as $a \to \infty$ of $P_{11} - P_{12}^2/P_{22}$ is

$$\begin{aligned}
&\delta^2 + 2\delta\theta x_1 + 2\delta t z_1 + \boldsymbol{v}_1^\top \boldsymbol{P_W v}_1 - (x_1 + \delta\theta)^2 \\
&\quad = \boldsymbol{v}_1^\top \boldsymbol{P_W v}_1 - x_1^2 + \delta^2 t^2 + 2\delta t z_1.
\end{aligned} \tag{44}$$

In equation (20), we introduced the quantity Q_{11}, equal to $\boldsymbol{v}_1^\top \boldsymbol{P_W v}_1$ and distributed as χ^2_l. It was expressed as the sum of three mutually independent random variables, x_1^2, z_P^2, and t_{11}^P. Now we separate out both the terms x_1^2 and z_1^2 to obtain

$$Q_{11} = x_1^2 + z_1^2 + z_P^2 + t_{11}^{P_0}, \tag{45}$$

where all four random variables above are independent, with x_1, z_1, and z_P standard normal, and $t_{11}^{P_0}$ distributed as χ^2_{l-3}. Note that $t_{11}^{P_0}$ is not to be confused with t_{11}^P in equations (20), which is distributed as χ^2_{l-2}.

It is legitimate to write Q_{11} in this way because it can be constructed as the sum of the squares of the l independent N(0,1) variables $\boldsymbol{w}_j^\top \boldsymbol{v}_1$, where the $\boldsymbol{w}_j$ form an arbitrary orthonormal basis of the span of the columns of $\boldsymbol{W}$. Using (45), the right-hand side of (44) can be written as

$$z_P^2 + t_{11}^{P_0} + (z_1 + \delta t)^2.$$

This is the sum of three independent random variables. The first is χ_1^2, the second is χ_{l-3}^2, and the last is noncentral $\chi_1^2(\delta^2 t^2)$. It follows that, when a^2 and the sample size both tend to infinity, which implies that the instruments are not weak, the numerator of the test statistic is distributed as $\chi_{l-1}^2(\delta^2 t^2)$. Note that, if $\theta = 1$, so that $\boldsymbol{w}_p = \boldsymbol{w}_1$, the NCP $\delta^2 t^2$ vanishes.

For the general model (1) and (2), with DGP given by equation (41), it can be shown that the NCP is

$$\frac{1}{\sigma_1^2} \boldsymbol{\delta}^\top \boldsymbol{W}_1^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{W}_1 \boldsymbol{\delta} - \frac{1}{\sigma_1^2} \frac{(\boldsymbol{\pi}^\top \boldsymbol{W}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{W}_1 \boldsymbol{\delta})^2}{\boldsymbol{\pi}^\top \boldsymbol{W}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{W} \boldsymbol{\pi}}. \tag{46}$$

For the simpler model given by equations (42) and (43), the first term here collapses to δ^2 and the second term, which arises because β has to be estimated, collapses to $-\theta^2 \delta^2$. Therefore, expression (46) as a whole corresponds to $\delta^2 t^2$ for the simpler model.

Finite-sample concerns

The asymptotic result that S' follows the $\chi_{l-1}^2(\delta^2 t^2)$ distribution strongly suggests that S, LR, and LR$'$ must do so as well, because all these statistics are asymptotically equivalent. In fact, a more tedious calculation than that in equations (44) and (45) shows that the limiting distribution of LR$'$ as both n and a tend to infinity is the same as for S', namely $\chi_{l-1}^2(\delta^2 t^2)$. Because these results are only asymptotic, however, it is necessary to resort to simulation to investigate behavior under the alternative in finite samples.

Under the null, we were able in Section 3 to express all six quantities, the P_{ij} and the M_{ij}, for $i, j = 1, 2$, in terms of eight independent random variables. Under the alternative, we require ten of these variables. For the M_{ij}, there is no need to change the expressions for them in (21), where we use the three variables t_{11}^M, t_{22}^M, and z_M, distributed respectively as χ_{n-l}^2, χ_{n-l-1}^2, and N(0,1). These represent the projections of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ onto the orthogonal complement of the span of the instruments. For the P_{ij}, however, we decompose as follows:

$$\begin{aligned} Q_{11} &= x_1^2 + z_1^2 + z_P^2 + t_{11}^{P_0}, \\ Q_{12} &= x_1 x_2 + z_1 z_2 + z_P \sqrt{t_{22}^{P_0}}, \text{ and} \\ Q_{22} &= x_2^2 + z_2^2 + t_{22}^{P_0}. \end{aligned}$$

Here x_i, z_i, $i = 1, 2$, and z_P are standard normal, t_{11}^P is χ_{l-3}^2, and t_{22}^P is χ_{l-2}^2, all seven variables being mutually independent. We can simulate both S' and LR$'$ very

cheaply, by drawing ten random variables, independently of either the sample size n or the degree of overidentification $l - 1$, because all the statistics are deterministic functions of the P_{ij} and the M_{ij}, and, of course, n and l. The relations in (21) hold except those for P_{11} and P_{12}. These are replaced by

$$P_{11} = Q_{11} + \delta^2 + 2\delta\theta x_1 + 2\delta t z_1,$$

and

$$P_{12} = a x_1 + \rho Q_{11} + r Q_{12} + \delta(a\theta + \rho\theta x_1 + \rho t z_1 + r\theta x_2 + r t z_2).$$

These equations differ from the corresponding ones in (21) only by terms proportional to a positive power of δ.

Simulation evidence

Since we have seen that the LR′ test often has much better finite-sample properties than the S' test, even when both are bootstrapped, it is important to see whether the superior performance of LR′ comes at the expense of power. In this section, we employ simulation methods to do so.

Given the considerable size distortion of the asymptotic tests for most of that part of the parameter space considered in Section 7, we limit attention to parametric bootstrap tests. In this, we follow Horowitz and Savin (2000), which argues that the best way to proceed, as long as the rejection probability of a test is far removed from its nominal level, is to consider a bootstrap test. But that proposition is based on the assumption that the bootstrap discrepancy is small enough to be ignored, which is not the case for the overidentification tests we have considered in the neighborhood of the singularity. Because of that, and because it is unreasonable to expect that there is much in the way of usable power near the singularity, it is primarily of interest to investigate power for situations in which the instruments are not too weak.

As before, all the simulation results are presented graphically. These results are based on 200,000 replications with 399 bootstrap repetitions. The same random variables are used for every set of parameter values. These experiments would have been extremely computationally demanding without the theoretical results of Sections 6 and the first part of this section, which allow us to calculate everything very cheaply after we have generated and stored $200,000 \times 10$ plus $200,000 \times 399 \times 8$ random variables. The first set of random variables is used to calculate the actual test statistics and the estimates of a and ρ, and the second set is used to calculate the bootstrap statistics.

We report results only for S' bootstrapped using the IV parameter estimates and for LR′ bootstrapped using the LIML estimates. Recall from Section 2 that the former results apply to S as well as S', and the latter apply to LR as well as LR′, because the test statistics in each pair are monotonically related.

Figure 11 shows power functions for $q = 8$, $\rho = 0.5$, and four values of a. When $a = 2$, LR′ rejects much less frequently than S', both under the null and under the

alternative. Both power functions level out as δ becomes large, and it appears that neither test rejects with probability one as $\delta \to \infty$. As a increases, the two power functions converge, and both tests do seem to reject with probability one for large δ.

The top two panels of Figure 12 are comparable to the top two panels of Figure 11, but with $q = 2$. When $a = 2$, S' now rejects less often that it did before, but LR$'$ rejects more often. When $a = 4$, LR$'$ rejects very much more often than it did before, and the two power functions are quite close. We also obtained results for $a = 6$, $a = 8$, and $a = 16$, which are not shown. For $a = 6$, the power functions for S' and LR$'$ are extremely similar, and for $a \geq 8$ they are visually indistinguishable.

The bottom two panels of Figure 12 are comparable to the top right panel, except that $\rho = 0.1$ or $\rho = 0.9$ instead of $\rho = 0.5$. It is evident that the shapes of the power functions depend on ρ, but for most values of δ the dependence is moderate. This justifies our use of $\rho = 0.5$ in most of the experiments. Using other values of ρ would not change the main results.

When one power function is always above another, as is the case in all the panels of Figures 11 and 12, it is difficult to conclude that one test is genuinely more powerful than the other. Perhaps greater power is just an artifact of greater rejection frequencies whether or not the null hypothesis is true.

One way to compare such tests is to graph rejection frequencies under the alternative against rejection frequencies under the null. Each point on such a "size-power curve" corresponds to some nominal level for the bootstrap test, with levels running from 0 to 1. The abscissa is the rejection frequency when the DGP satisfies the null, the ordinate the rejection frequency when the DGP belongs to the alternative. For a level of 0, the test never rejects, since bootstrap P values cannot be negative. If the level is 1, the test always rejects. As the nominal level increases from 0 to 1, we expect power (on the vertical axis) to increase more rapidly than the rejection frequency under the null (on the horizontal axis). See Davidson and MacKinnon (1998).

The top two panels of Figure 13 show size-power curves for $q = 2$, $a = 4$, and four values of δ. Perhaps surprisingly, the curves for LR$'$ in the left-hand panel look remarkably similar to the ones for S' in the right-hand panel. The apparently greater power of S', which is evident in the top right panel of Figure 12, seems to be almost entirely accounted for by its greater tendency to reject under the null.

The bottom two panels of Figure 13 show size-power curves for $q = 2$, $\delta = 4$, and four values of a. It is clear that power increases with a, but at a decreasing rate. As $a \to \infty$, the curves converge to the one given by asymptotic theory, where the distribution under the null is central χ^2_{l-1} and the one under the alternative is noncentral $\chi^2_{l-1}(\delta^2 t^2)$. This curve is graphed in the figure and labelled $a = \infty$.

The asymptotic result that the test statistics follow the $\chi^2_{l-1}(\delta^2 t^2)$ distribution suggests that only the product $\delta t = \delta(1-\theta^2)^{1/2}$ influences power, and that, in particular, there should be no power beyond the level of the test when $\theta = 1$. In finite samples,

things turn out to be more complicated, as can be seen from Figure 14, which plots power against θ for $\delta = 4$. The top two panels show results for $a = 2$ and $a = 4$. The S' test has substantial power when $\theta = 1$ and $a = 2$, which presumably reflects its tendency to overreject severely under the null when the instruments are weak. Those panels also show, once again, that S' can reject far more often than LR$'$ when the instruments are weak. This is much less evident in the bottom two panels, which show results for larger values of a (6 and 8).

One surprising feature of Figure 14 is that, in all cases, power initially increases as θ increases from 0, even though $\delta(1 - \theta^2)^{1/2}$ declines. This is true even for quite large values of a, such as $a = 16$, although, of course, it is not true for extremely large values.

9. Relaxing the IID Assumption

The resampling bootstraps that we looked at in Section 7 do not implicitly make the assumption that the disturbances are normal. They do, however, assume that the disturbances are pairwise IID. If instead the disturbances are heteroskedastic, then the covariance matrix of their bivariate distribution may be different for each observation. In that case, all the test statistics we have studied have distributions that depend on the pattern of heteroskedasticity, and so they are no longer approximately pivotal for the model (1) and (2) under either weak-instrument or strong-instrument asymptotics.

Andrews, Moreira, and Stock (2004) proposes heteroskedasticity-robust versions of test statistics for tests about the value of β that are robust to weak instruments. Note that, although Andrews, Moreira, and Stock (2006) is based on the 2004 paper and has almost the same title, it does not contain this material. However, this work cannot be applied here, because, as we have seen, the overidentification tests are not robust to weak instruments.

The role of the denominators of the statistics S, S', and LR$'$ is simply to provide non-robust estimates of the scale of the numerators. In order to make those statistics robust to heteroskedasticity, we have to provide robust measures instead. The numerators of all three statistics can be written as

$$\hat{\boldsymbol{u}}_1^\top \boldsymbol{P}_{\boldsymbol{W}} \hat{\boldsymbol{u}}_1 = \hat{\boldsymbol{u}}_1^\top \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top \hat{\boldsymbol{u}}_1, \tag{47}$$

where the vector $\hat{\boldsymbol{u}}_1$ denotes either $\boldsymbol{y}_1 - \boldsymbol{Z}\hat{\gamma}_{\text{IV}} - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2$, in the case of S and S', or $\boldsymbol{y}_1 - \boldsymbol{Z}\hat{\gamma}_{\text{LIML}} - \hat{\beta}_{\text{LIML}}\boldsymbol{y}_2$, in the case of LR$'$. Expression (47) is a quadratic form in the l–vector $\boldsymbol{W}^\top \hat{\boldsymbol{u}}_1$. The usual estimate of the covariance matrix of that vector is $\boldsymbol{W}^\top \hat{\boldsymbol{\Omega}} \boldsymbol{W}$, where $\hat{\boldsymbol{\Omega}} = \operatorname{diag} \hat{u}_{1i}^2$. Thus the heteroskedasticity-robust variant of all three test statistics is the quadratic form

$$\hat{\boldsymbol{u}}_1^\top \boldsymbol{W} (\boldsymbol{W}^\top \hat{\boldsymbol{\Omega}} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \hat{\boldsymbol{u}}_1. \tag{48}$$

There would be no point in using a heteroskedasticity-robust statistic along with a bootstrap DGP that imposed homoskedasticity. The natural way to avoid doing so is to use the wild bootstrap. In Davidson and MacKinnon (2010), the wild bootstrap is shown to have good properties when used with tests about the value of β. The disturbances of the wild bootstrap DGP are given by

$$\begin{bmatrix} u_{1i}^* \\ u_{2i}^* \end{bmatrix} = \begin{bmatrix} \hat{u}_{1i}\nu_i^* \\ \hat{u}_{2i}\nu_i^* \end{bmatrix}, \tag{49}$$

where ν_i^* is an auxiliary random variable with expectation 0 and variance 1. The easiest choice for the distribution of the ν_i^* is the Rademacher distribution, which sets ν_i^* to $+1$ or -1, each with probability one half. This is also probably the best choice in most cases; see Davidson and Flachaire (2008).

The IID assumption can, of course, be relaxed in other ways. In particular, it would be easy to modify the test statistic (48) to allow for clustered data by replacing the middle matrix with one that resembles the middle matrix for the usual cluster robust covariance matrix. We could then use a variant of the cluster robust wild bootstrap of Cameron et al. (2008) that allows for simultaneity. The Rademacher random variable associated with each cluster, the analog of ν_i^* in equation (49), would then multiply the residuals for all observations within that cluster for both equations.

10. Concluding Remarks

We have shown that the well-known Sargan test for overidentification in a linear simultaneous-equations model estimated by instrumental variables often overrejects severely when the instruments are weak. In the same circumstances, the likelihood ratio test often underrejects severely. We provide a finite-sample analysis that explains these facts and shows that the distributions of the different test statistics we consider have a singularity when the concentration parameter vanishes and the absolute value of the correlation between the disturbances of the structural and reduced-form equations tends to one. Thus it can be risky to use asymptotic tests in this situation. We have proposed a new test based on Fuller's modified LIML estimator, which often outperforms the ordinary LR test.

We have also proposed four bootstrap methods which can be applied to all three of these tests. Although bootstrapping does not help much when the instruments are extremely weak, especially when the disturbances of the two equations are highly correlated, it does help substantially when the instruments are only moderately weak. In particular, using a bootstrap DGP based on Fuller's estimator generally leads to much more accurate inferences than simply using asymptotic theory in this case.

There is a cost in terms of power to using a bootstrap test based on any version of the likelihood ratio statistic relative to a test based on the conventional Sargan or Basmann statistics. This cost generally seems to be very modest, except when the instruments are very weak.

References

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 (2004). "Optimal invariant similar tests for instrumental variables regression", NBER Technical Working Paper 299.

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.

Basmann, R. L. (1960). "On the finite sample distributions of generalized classical identifiability test statistics", *Journal of the American Statistical Association*, **55**, 650–659.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). "Bootstrap-based improvements for inference with clustered errors", *Review of Economics and Statistics*, **80**, 414–427.

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

Davidson, R. and J. G. MacKinnon (1998). "Graphical methods for investigating the size and power of hypothesis tests", *The Manchester School*, **66**, 1–26.

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 T. C. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics*, Volume 1, pp. 812–838. Basingstoke: Palgrave-Macmillan.

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 (2008). "Bootstrap inference in a linear equation estimated by instrumental variables", *Econometrics Journal*, **11**, 443–477.

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

Fuller, W. A. (1977). "Some properties of a modification of the limited information estimator", *Econometrica*, **45**, 939–953.

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.

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. (2009), "Tests with correct size when instruments can be arbitrarily weak", *Journal of Econometrics*, 152, 131–140.

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

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

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

Stock, J. H., J. H. Wright, and M. Yogo (2002). "A survey of weak instruments and weak identification in generalized method of moments", *Journal of Business and Economic Statistics*, **20**, 518–529.

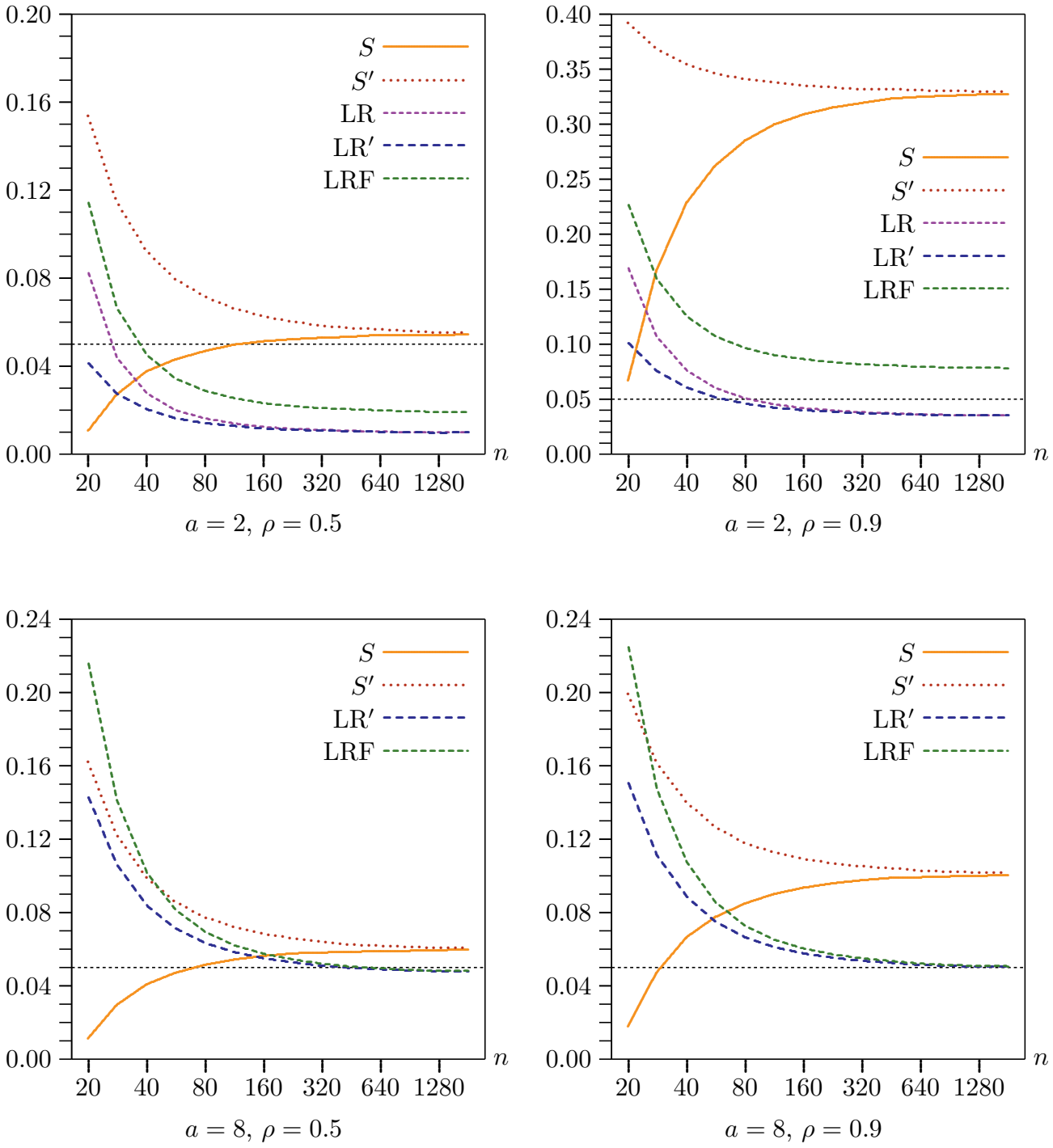

Figure 1. Rejection frequencies for asymptotic tests as a function of n for $q = 8$

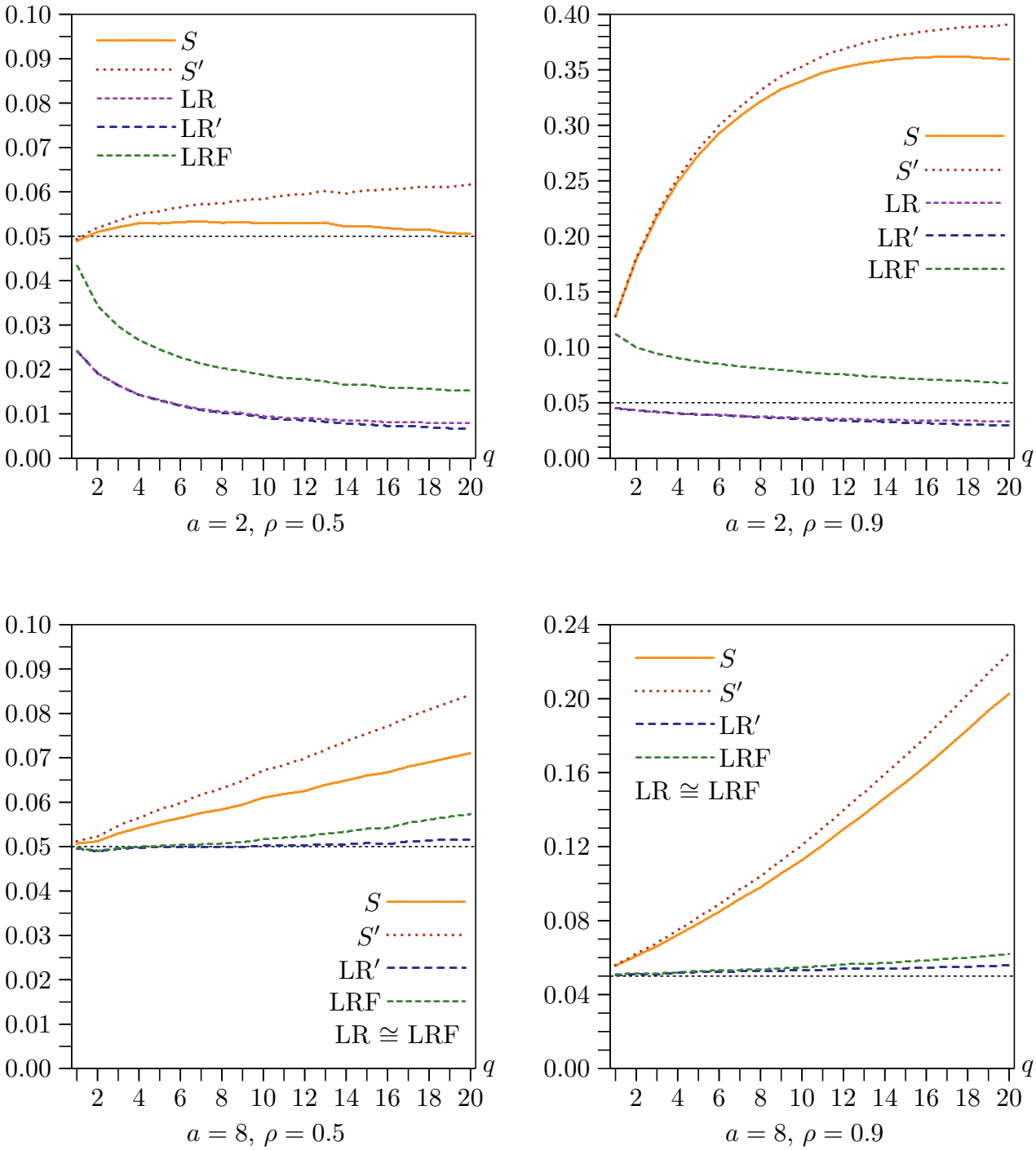

Figure 2. Rejection frequencies for asymptotic tests as functions of q for $n = 400$

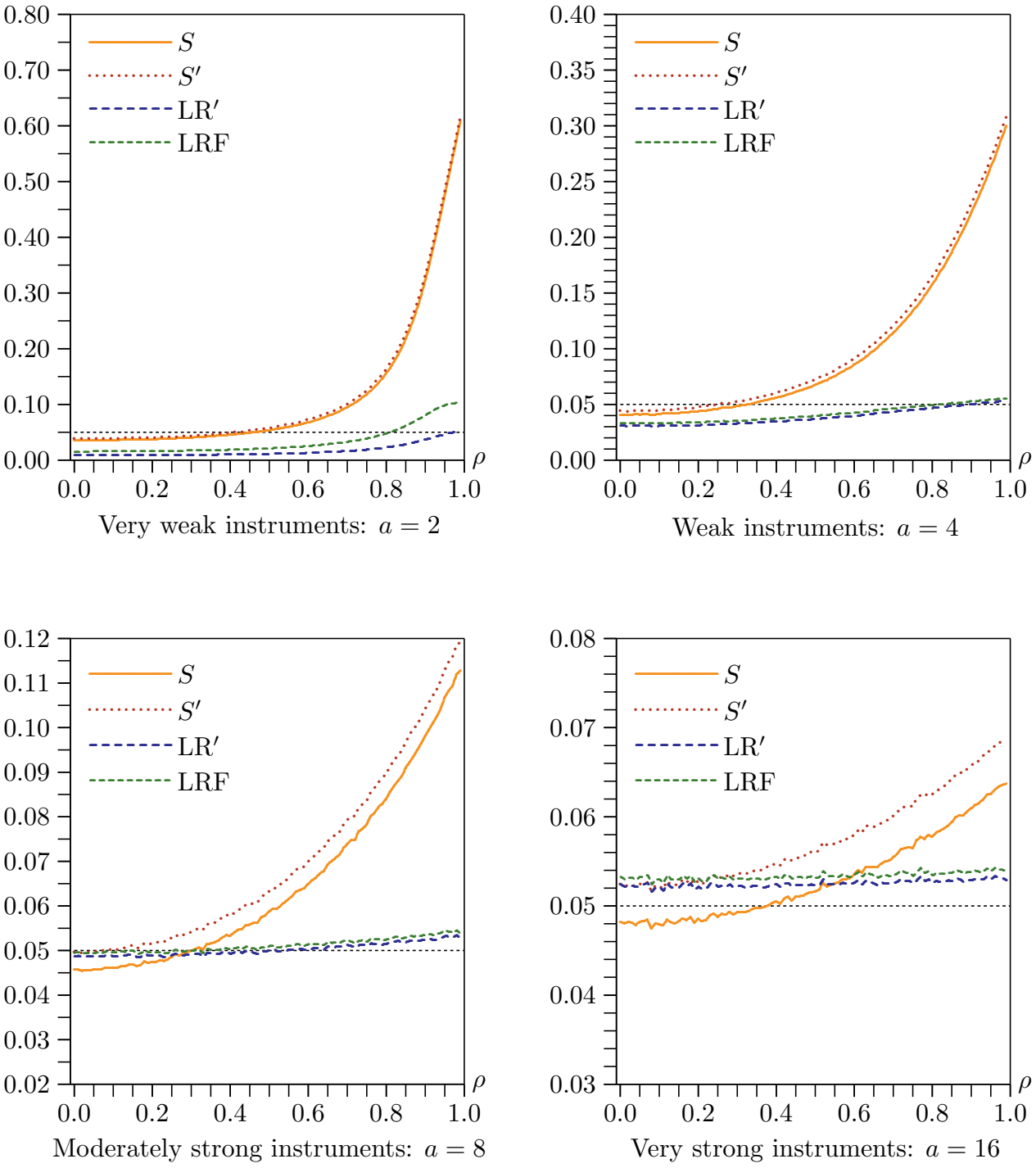

Figure 3. Rejection frequencies for asymptotic tests as functions of ρ for $q = 8$ and $n = 400$

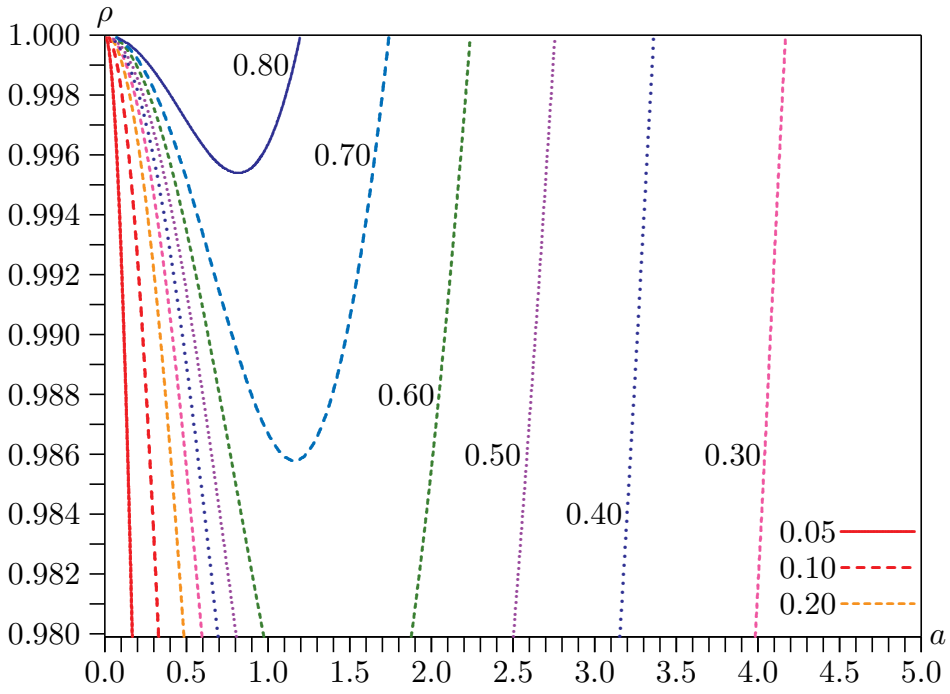

Figure 4. Contours of rejection frequencies for S' tests with $q = 8$ and $n = 400$

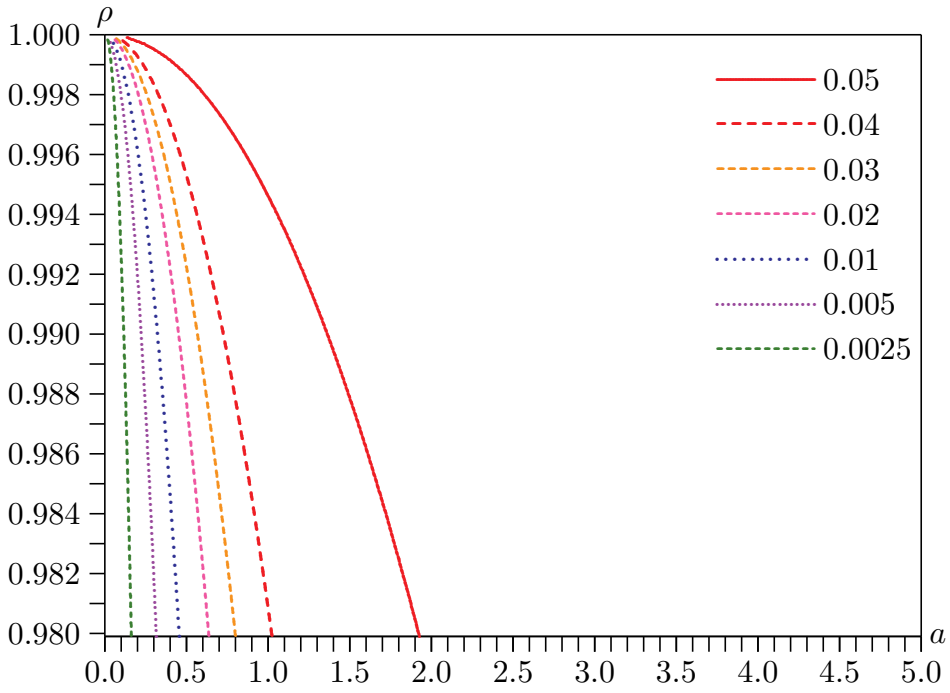

Figure 5. Contours of rejection frequencies for LR$'$ tests with $q = 8$ and $n = 400$

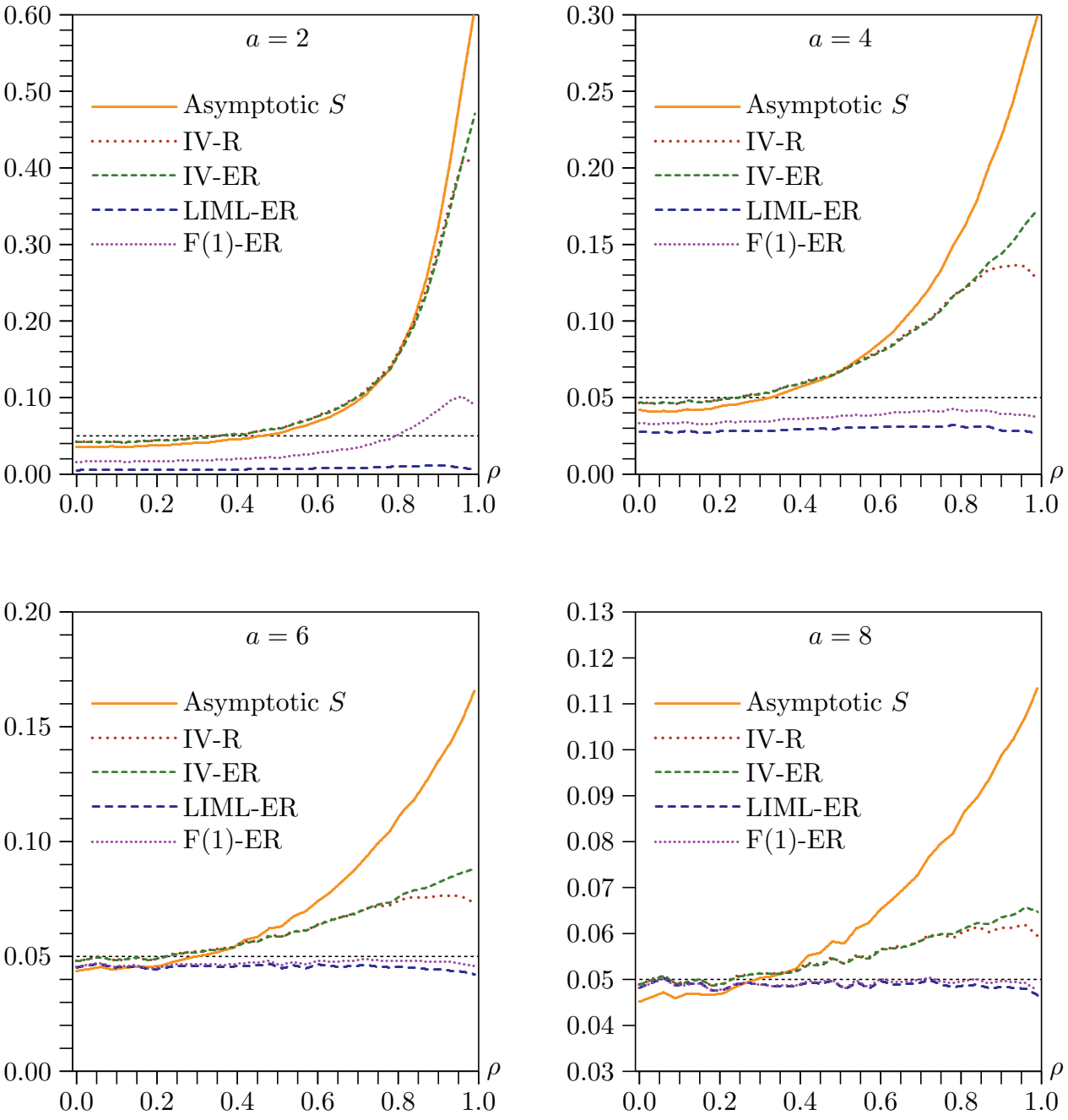

Figure 6. Rejection frequencies for Sargan tests as functions of ρ for $q = 8$ and $n = 400$

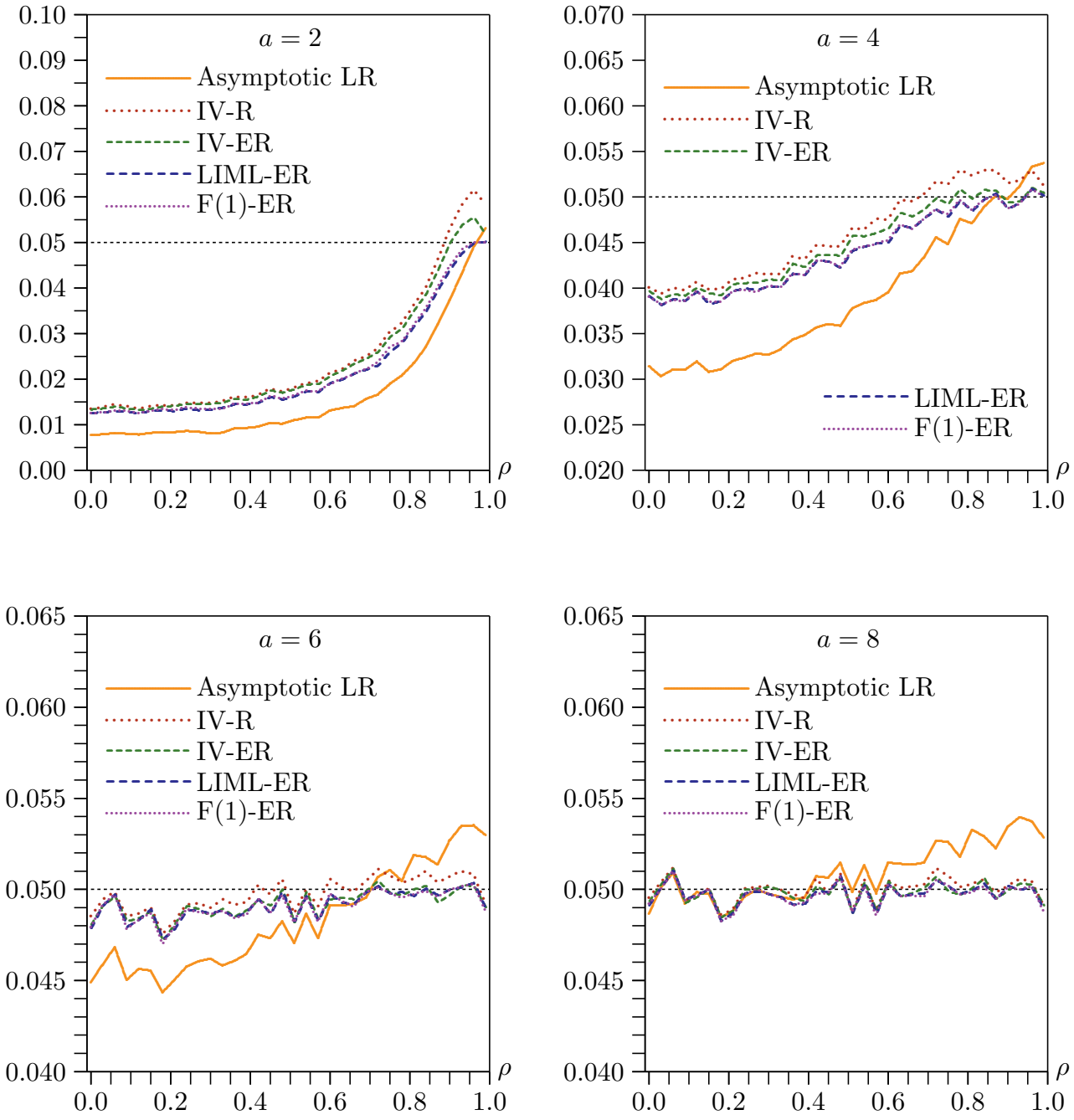

Figure 7. Rejection frequencies for LR tests as functions of ρ for $q = 8$ and $n = 400$

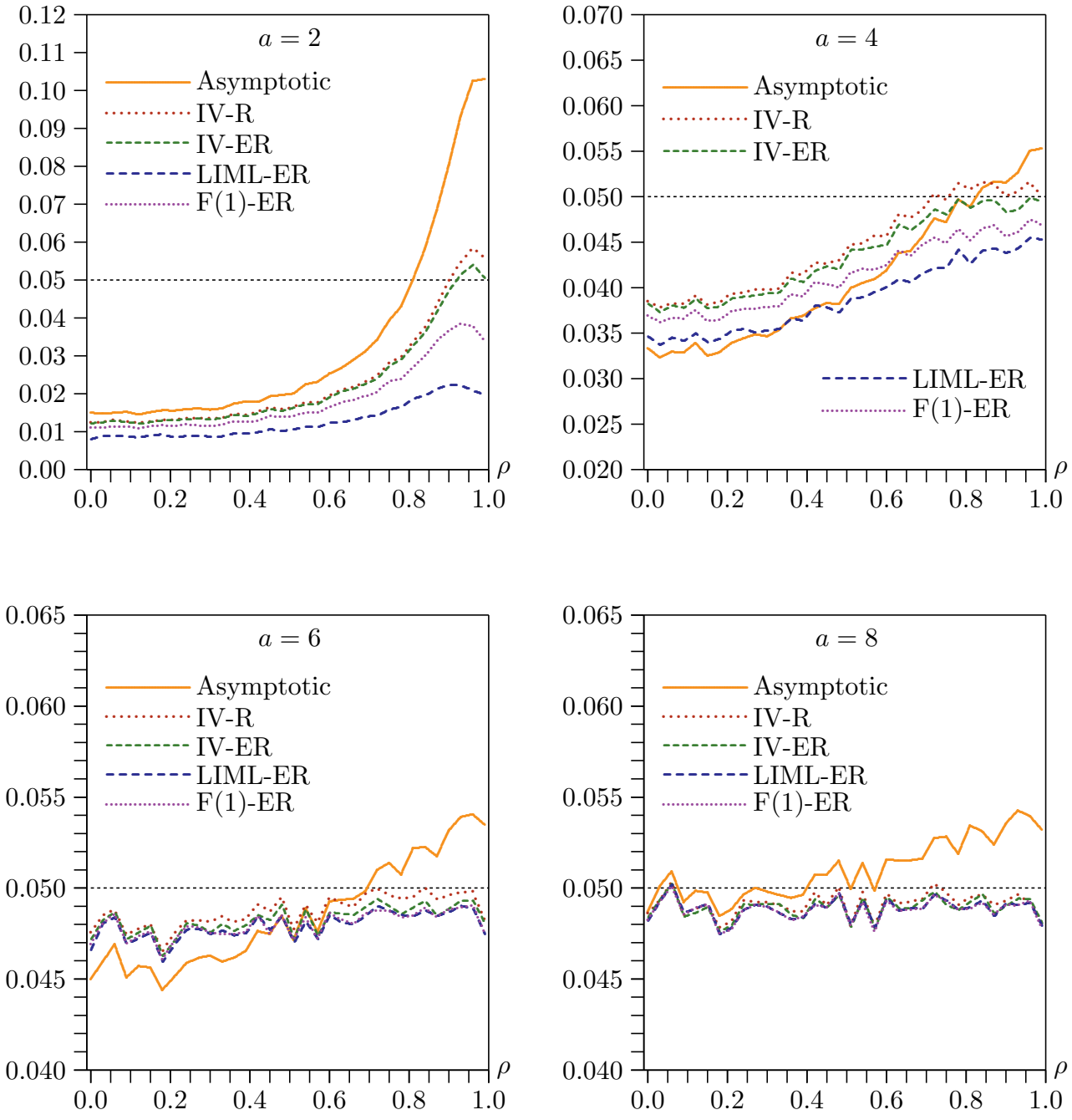

Figure 8. Rejection frequencies for Fuller LR tests as functions of ρ for $q = 8$ and $n = 400$

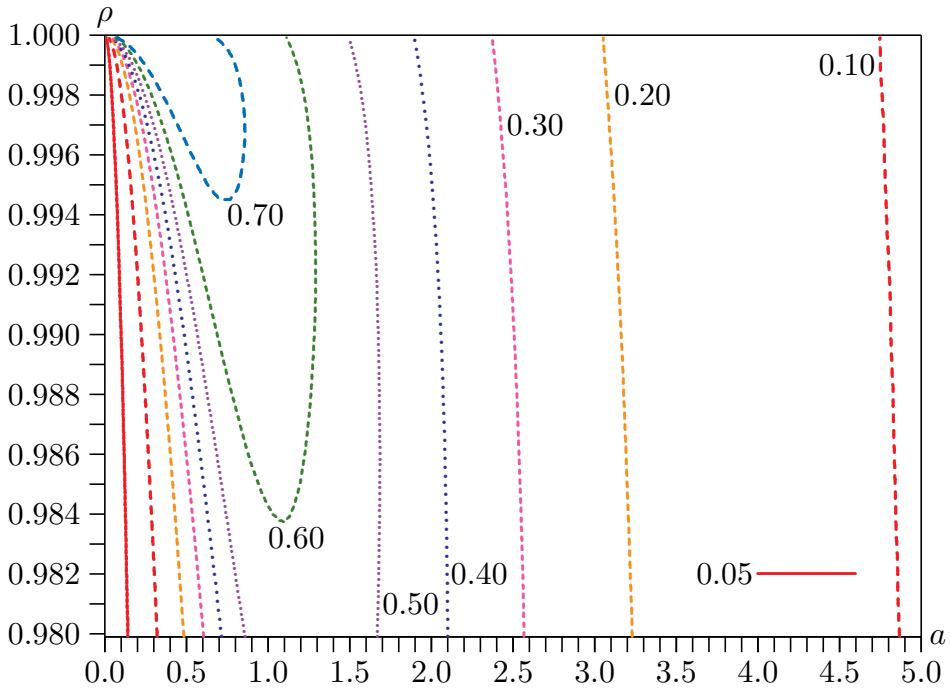

Figure 9. Contours of rejection frequencies for IV-R bootstrap Sargan tests

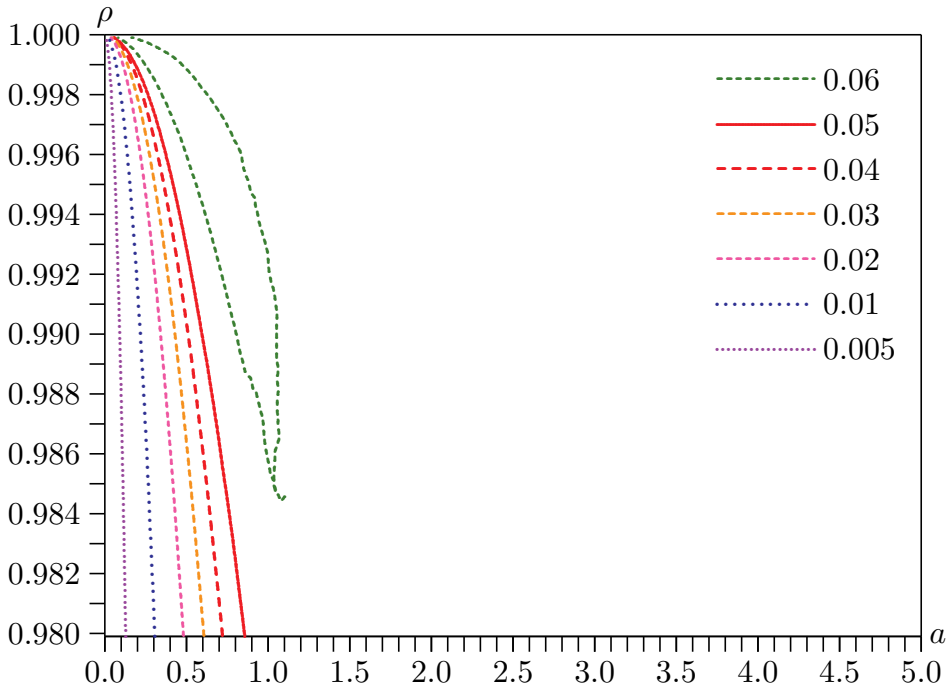

Figure 10. Contours of rejection frequencies for LIML-ER bootstrap LR tests

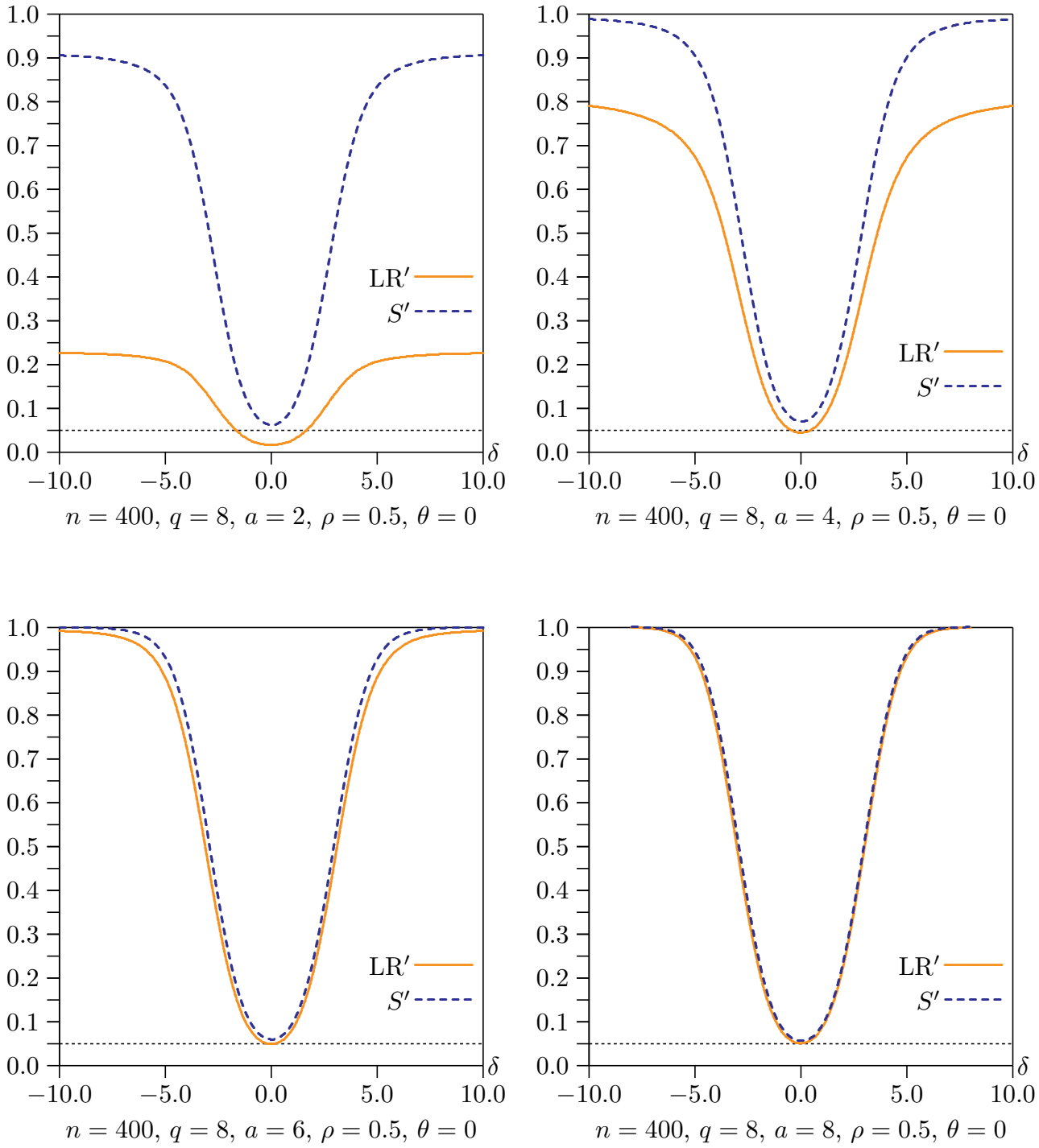

Figure 11. Power of bootstrap tests as functions of δ with $t = 1$

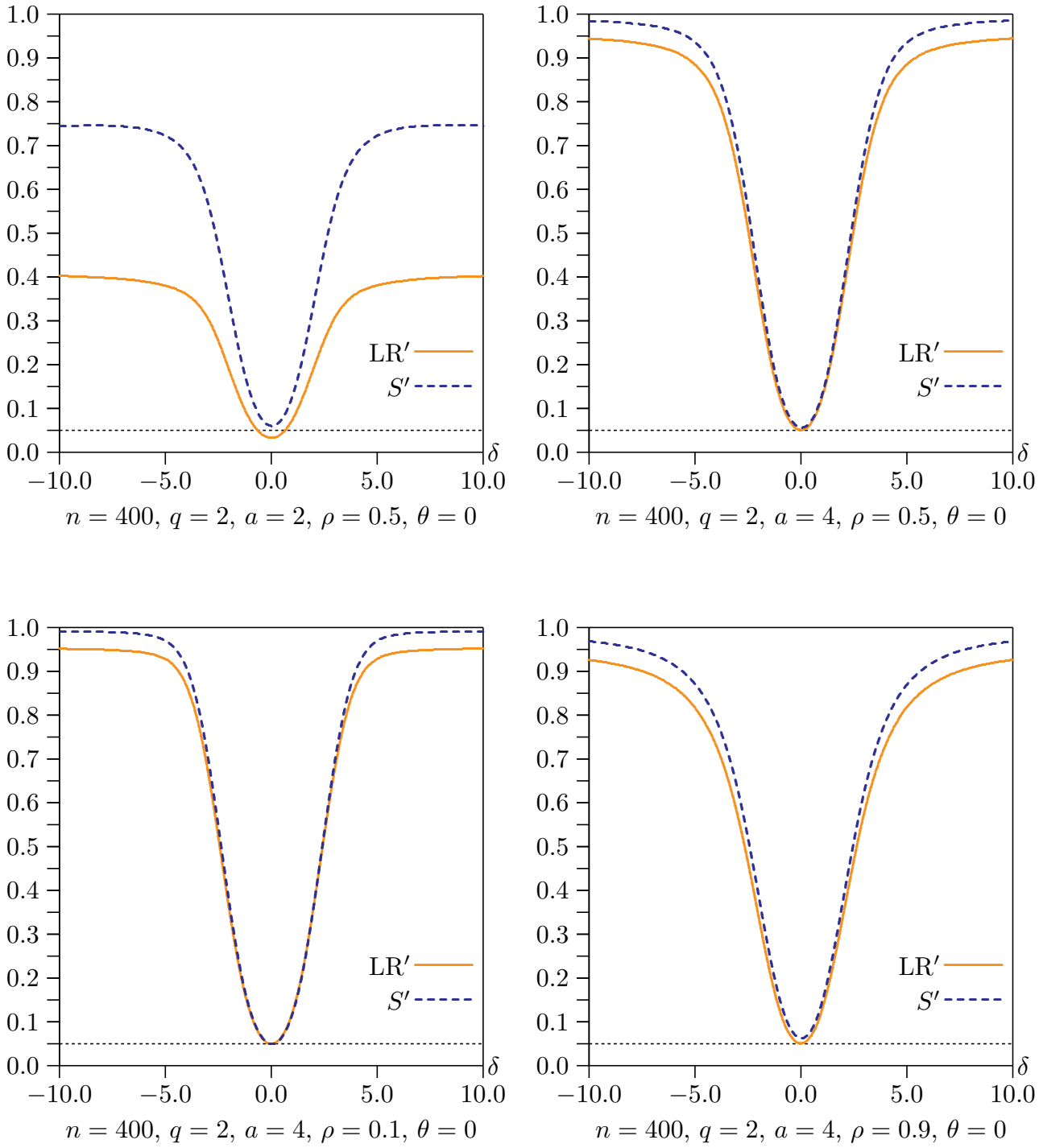

Figure 12. Power of bootstrap tests as functions of δ with $t = 1$

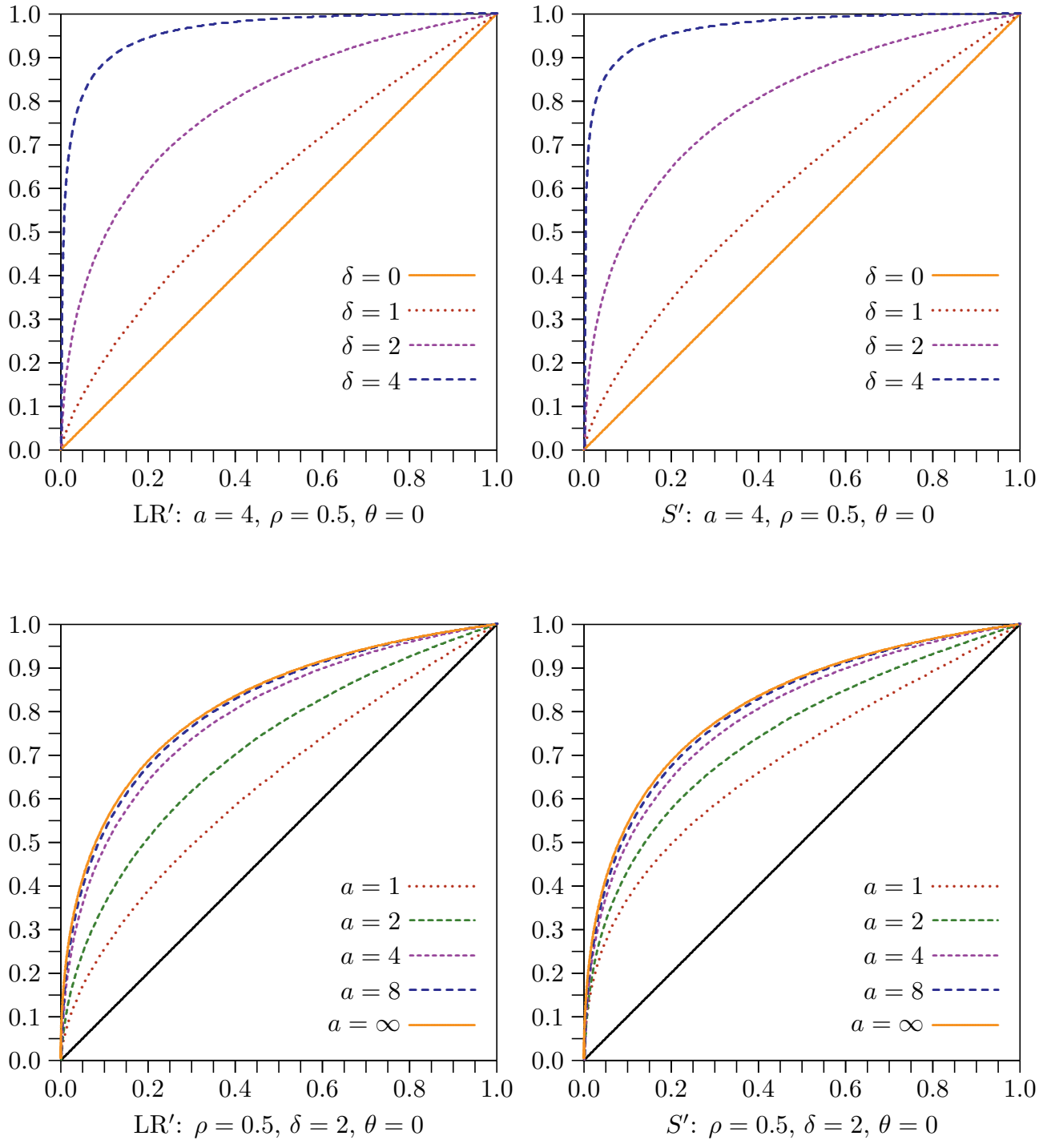

Figure 13. Size-power curves, $q = 2$, $n = 400$

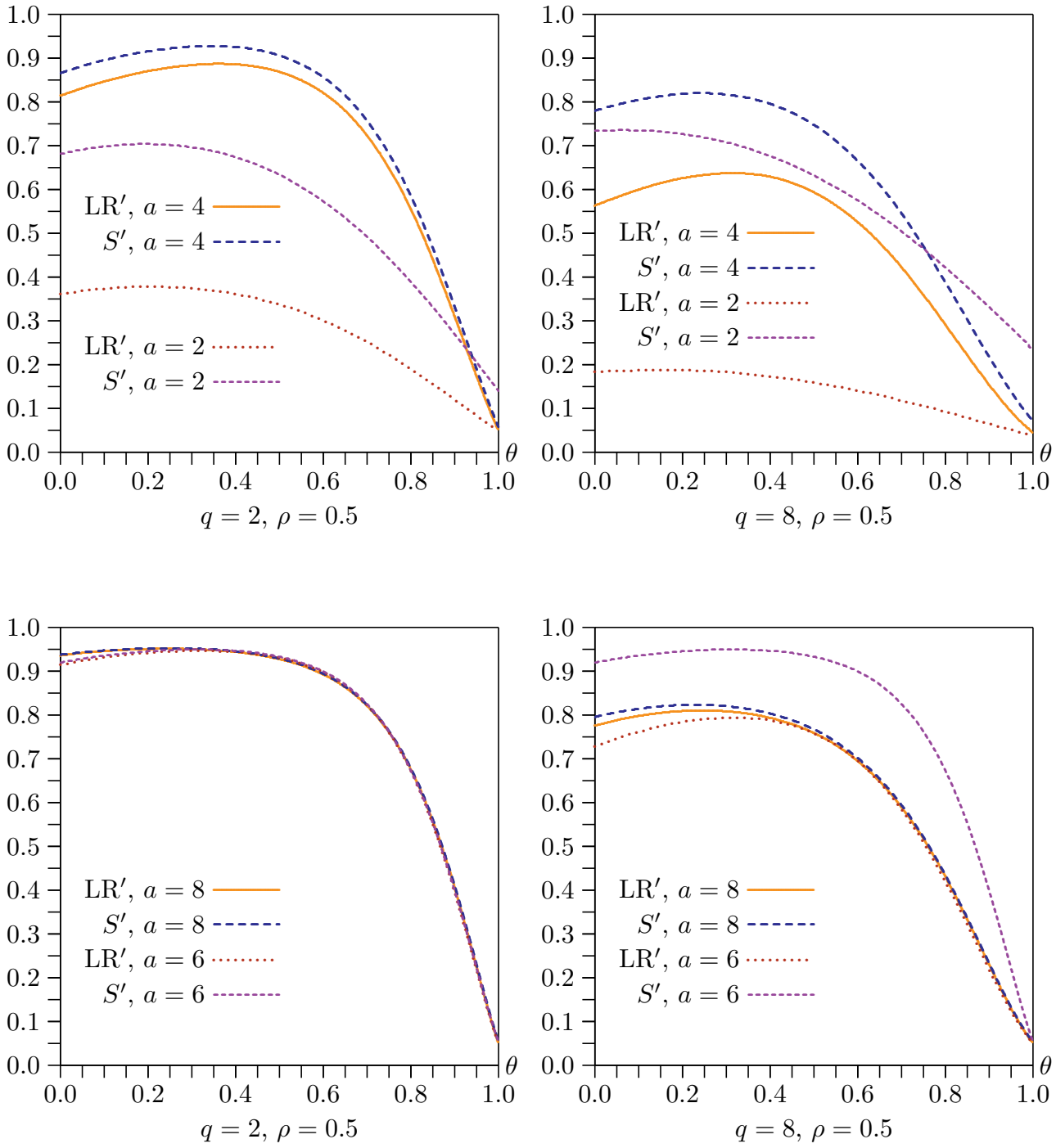

Figure 14. Power as a function of θ for $n = 400$ and $\delta = 4$