Econometrica, Vol. 80, No. 5 (September, 2012), 2153–2211

ESTIMATION AND INFERENCE WITH WEAK, SEMI-STRONG, AND STRONG IDENTIFICATION

BY DONALD W. K. ANDREWS AND XU CHENG[1]

This paper analyzes the properties of standard estimators, tests, and confidence sets (CS's) for parameters that are unidentified or weakly identified in some parts of the parameter space. The paper also introduces methods to make the tests and CS's robust to such identification problems. The results apply to a class of extremum estimators and corresponding tests and CS's that are based on criterion functions that satisfy certain asymptotic stochastic quadratic expansions and that depend on the parameter that determines the strength of identification. This covers a class of models estimated using maximum likelihood (ML), least squares (LS), quantile, generalized method of moments, generalized empirical likelihood, minimum distance, and semi-parametric estimators.

The consistency/lack-of-consistency and asymptotic distributions of the estimators are established under a full range of drifting sequences of true distributions. The asymptotic sizes (in a uniform sense) of standard and identification-robust tests and CS's are established. The results are applied to the ARMA(1, 1) time series model estimated by ML and to the nonlinear regression model estimated by LS. In companion papers, the results are applied to a number of other models.

KEYWORDS: Asymptotic size, confidence set, estimator, identification, nonlinear models, strong identification, test, weak identification.

1. INTRODUCTION

THE MAIN CONTRIBUTIONS of this paper are as follows. (i) We provide a unified treatment of a class of models in which lack of identification and weak identification occurs in part of the parameter space. (ii) We analyze the asymptotic properties of extremum estimators, and t and quasi-likelihood ratio (QLR) tests and confidence sets (CS's). The results extend standard results for extremum estimators under high-level conditions to allow for singularity of the variance matrix. (iii) We introduce tests and CS's that are robust to identification issues. (iv) We provide asymptotic results that are uniform over distributions that generate the observations. This requires results for what we call the region of *semi-strong* identification, which bridges the gap between weak and strong identification. (v) We give a detailed analysis of the effects of identification weakness in the workhorse (autoregressive moving average) ARMA(1, 1) time series model.

The main technical innovations of the paper are the following. (i) For the weak identification asymptotic results, we do a quadratic approximation

[1]Andrews gratefully acknowledges the research support of the National Science Foundation via Grants SES-0751517 and SES-1058376. The authors thank a co-editor, three referees, Xiaohong Chen, Patrik Guggenberger, Sukjin Han, Yuichi Kitamura, Ulrich Müller, Peter Phillips, Eric Renault, Frank Schorfheide, Yixiao Sun, and Ed Vytlacil for helpful comments.

 DOI: 10.3982/ECTA9456

around the point of lack of identification, rather than around the true parameter.[2] (ii) In the semi-strong identification case, we obtain consistency using a nonstochastic limit of the criterion function that has not appeared before in the literature.[3] (iii) To obtain the asymptotic distribution in the semi-strong identification case, we use a quadratic expansion of the criterion function that is novel in that it only holds in a rapidly shrinking (as $n \to \infty$) neighborhood of the true parameter, combined with a key rate of convergence result for the estimator.[4]

We consider models in which the parameter θ of interest is of the form $\theta = (\beta, \zeta, \pi)$, where π is identified if and only if $\beta \neq 0$, ζ is not related to the identification of π, and $\psi = (\beta, \zeta)$ is always identified.[5] This is a canonical parametrization that may or may not hold in the natural parametrization of the model, but is assumed to hold after suitable reparametrization.

We suppose θ is estimated by minimizing a criterion function $Q_n(\theta)$ over a parameter space Θ, where n denotes the sample size. The true distribution that generates the data is indexed by a parameter $\gamma^* = (\theta^*, \phi^*)$ with parameter space Γ. Here θ^* denotes the true value of θ and ϕ^* indexes the part of the distribution of the data that is not determined by θ^*. A key assumption used in the paper is the following.

ASSUMPTION A: *If* $\beta = 0$, $Q_n(\theta)$ *does not depend on* π $\forall \theta = (\beta, \zeta, \pi) = (0, \zeta, \pi) \in \Theta$, $\forall n \geq 1$, *for any true parameter* $\gamma^* \in \Gamma$.[6]

Under Assumption A (and other conditions given below), $Q_n(\theta)$ is (relatively) flat with respect to (w.r.t.) π when β is close to 0. This causes difficulties with standard asymptotic approximations because the second derivative matrix of $Q_n(\theta)$ is singular or near singular and standard asymptotic approximations involve the inverse of this matrix.

[2]In consequence, the leading term of the expansion does not depend on the unidentified parameter, which is key to determining the asymptotic properties of the extremum estimator. This introduces a bias in the first derivative in the expansion—its mean is not zero.

[3]This limit is a nonstochastic quadratic form in the bias vector of the first derivative that appears in the quadratic approximation in part (i). See the function $\eta(\pi; \gamma_0, \omega_0)$ in (3.8) below.

[4]The shrinking neighborhood depends on the strength of identification. The rate of convergence result for the estimator establishes that the estimator lies in the shrinking neighborhood with probability that goes to 1. It is based on a different quadratic expansion—the quadratic expansion used for the weak-identification results in part (i).

[5]The parameters β, ζ, and π may be scalars or vectors.

[6]Throughout the paper, we use the term identification/lack of identification in the sense of identification by a criterion function $Q_n(\theta)$, as specified in Assumption A. Lack of identification by the criterion function $Q_n(\theta)$ is not the same as lack of identification in the usual or strict sense of the term, although there is a close relationship. For example, with a likelihood criterion function, the former implies the latter. See Sargan (1983) for a related distinction between lack of identification in the strict sense and lack of first-order identification.

EXAMPLE 1: Consider the nonlinear regression model $Y_i = \beta^* h(X_i, \pi^*) + Z_i'\zeta^* + U_i$ and the least squares criterion function $Q_n(\theta) = n^{-1}\sum_{i=1}^n (Y_i - \beta h(X_i, \pi) - Z_i'\zeta)^2$.[7] The parameter π^* is not identified when $\beta^* = 0$. Assumption A holds. The first derivative of $Q_n(\theta)$ w.r.t. π is proportional to β. Hence, when β is close to zero, the criterion function $Q_n(\theta)$ is relatively flat in the direction of π.

EXAMPLE 2: Consider the ARMA(1, 1) model estimated by (quasi-) maximum likelihood (ML). In this model, the autoregressive (AR) and the moving average (MA) parameters are not identified when their values are equal. This occurs when the time series is serially uncorrelated—a case of considerable interest in many practical applications.[8] By definition, the observed ARMA(1, 1) time series $\{Y_t : 0 \le t \le n\}$ satisfies

$$Y_t = (\pi^* + \beta^*)Y_{t-1} + \varepsilon_t - \pi^*\varepsilon_{t-1} \quad \text{for} \quad t = \ldots, 0, 1, \ldots, \tag{1.1}$$

where the true MA parameter is π^*, the true AR parameter is $\pi^* + \beta^*$, the innovations $\{\varepsilon_t : t = \ldots, 0, 1, \ldots\}$ are independent and identically distributed (i.i.d.) with mean zero and variance ζ^*, and ϕ^* is the distribution of $(\zeta^*)^{-1/2}\varepsilon_t$. When $\beta^* = 0$, the model is $Y_t = \pi^* Y_{t-1} + \varepsilon_t - \pi^*\varepsilon_{t-1}$, which is equivalent to $Y_t = \varepsilon_t$. In this case, π^* and $\pi^* + \beta^*$ are not identified.

In the ARMA(1, 1) model, the Gaussian quasi-log-likelihood function for $\theta = (\beta, \zeta, \pi)$ conditional on Y_0 and ε_0 multiplied by $-n^{-1}$ and ignoring a constant is

$$Q_n(\theta) = \frac{1}{2}\log\zeta + \frac{1}{2\zeta}n^{-1}\sum_{t=1}^n \left(Y_t - \beta\sum_{j=0}^{t-1}\pi^j Y_{t-j-1}\right)^2. \tag{1.2}$$

Assumption A holds because $Q_n(\theta)$ does not depend on π when $\beta = 0$.

The approach of this paper is to consider a general class of extremum estimators. The criterion functions considered may be smooth or nonsmooth functions of θ. We place high-level conditions on the behavior of the criterion function $Q_n(\theta)$, provide a variety of more primitive sufficient conditions, and verify the latter in several examples.

We are concerned with cases in which the model is strongly identified in part of the parameter space, but unidentified or weakly identified in another part of the parameter space. In consequence, we establish the large sample properties of extremum estimators, t and QLR tests, and CS's over the

[7]Here ϕ^* is the true distribution of (X_i, Z_i, U_i) and the latter is i.i.d. for $i = 1, \ldots, n$.

[8]Simulation results in Ansley and Newbold (1980) and Nelson and Startz (2007) demonstrate that this causes substantial bias, variance, and size problems when the AR and MA parameters are close in value. Ma and Nelson (2008) provide analogous simulation results for the nonlinear regression model when β^* is close to zero. We provide an asymptotic analysis of these problems.

full range of strength-of-identification scenarios. These large sample properties provide good approximations to the statistics' finite-sample properties under all strengths of identification, whereas standard asymptotic theory only provides good approximations under strong identification. We determine the asymptotic size of standard t and QLR tests and CS's, which often deviate from their nominal size in the presence of lack of identification at some points in the parameter space.[9]

We introduce methods of making standard tests and CS's robust to lack of identification, that is, to have correct asymptotic size (in a uniform sense). These methods include least-favorable (LF), type 1 robust, and type 2 robust critical values. With type 1 and type 2 robust critical values, the idea is to use an identification-category selection procedure to determine whether β is close to the nonidentification value 0 and, if so, to adjust the critical value to take account of the effect of nonidentification or weak identification on the behavior of the test statistic. We also introduce null-imposed (NI) and plug-in versions of these robust critical values.

These methods apply to subvectors and low dimensional functions, $r(\theta)$, of the full parameter vector θ. They allow for procedures that are asymptotically efficient when identification is not weak. In general, they do not have asymptotic optimality properties under weak identification. Nevertheless, we investigate their power in the linear instrumental variable (IV) regression model in which the conditional likelihood ratio (CLR) test of Moreira (2003) has approximate asymptotic optimality properties; see Andrews, Moreira, and Stock (2006, 2008). We find that one of the robust tests introduced here has power that is essentially the same as that of the CLR test and, hence, is approximately asymptotically optimal in a class of invariant tests. Both of these tests are based on the same test statistic and the same conditioning statistic (which is used in the construction of the data-dependent critical value). The results show that the method of constructing the data-dependent critical value considered in this paper can yield approximately asymptotically optimal tests. In addition, the robust tests are generally applicable and often have the advantage of computational ease. See Elliott, Müller, and Watson (2011) for tests that have some approximate asymptotic optimality properties in models where a nuisance parameter appears under the null hypothesis.[10]

This paper applies the general results to the ARMA(1, 1) model and the nonlinear regression model. The results for the ARMA(1, 1) model are sum-

[9]Asymptotic size is defined to be the limit of exact (i.e., finite-sample) size. For a test, exact size is the maximum rejection probability over distributions in the null hypothesis. For a confidence interval (CI), exact size is the minimum coverage probability over all distributions. Because exact size has uniformity built into its definition, so does asymptotic size as defined here.

[10]Other procedures with asymptotic optimality/admissibility properties in models with potential identification failure include those of Elliott and Müller (2007, 2008) for some change-point models. These models are not covered by this paper because the quadratic approximation condition fails.

marized as follows. The distributions of the ML estimators of the MA and AR parameters are greatly effected by weak identification, both asymptotically and in finite samples. Their distributions are bi- or trimodal, biased for nonzero true values, and far from the standard normal distribution. The asymptotic distributions for the MA and AR parameter estimators are the same under weak identification. The uniform asymptotic approximations to the finite-sample distributions are remarkably good.

Standard t CI's are found to have asymptotic and finite-sample sizes that are very poor—less than 0.60 for nominal 95% CI's concerning the MA and AR parameters. Standard CI's based on the QLR statistic and a χ^2 critical value, on the other hand, have asymptotic and finite-sample sizes that are not correct, but are far superior to those of standard $|t|$ CI's. Their asymptotic size is 0.933 for nominal 95% CI's and their finite-sample sizes are close to this. The uniform asymptotic approximations for the standard t and QLR CI's work very well.

The nominal 95% robust CI's have asymptotic and finite-sample size that are equal to and close to 0.95, respectively. This is true even for the robust CI's based on the t statistic. The best robust CI in terms of false coverage probabilities is a type 2 robust CI based on the QLR statistic. The uniform asymptotic approximations for the robust CI's are found to work very well.

Two companion papers—Andrews and Cheng (2011a, 2011b) (hereafter AC2 and AC3, respectively) apply the results of this paper to a smooth transition threshold autoregressive (STAR) model, a smooth transition switching regression model, a nonlinear binary choice model, a nonlinear regression model with endogenous regressors, and a binary probit model with endogeneity and a linear reduced-form equation for the endogenous variable(s), as in Nelson and Olson (1978), Lee (1981), Rivers and Vuong (1988), and Magnusson (2010). Han (2009) shows that, via reparametrization, a simple bivariate probit model with endogeneity falls into the class of models considered here.[11]

Other examples covered by the results of this paper include mixed data sampling (MIDAS) regressions in empirical finance, which combine data with different sampling frequencies (see Ghysels, Sinko, and Valkanov (2007)), models with autoregressive distributed lags, continuous transition structural change models, continuous transition threshold autoregressive models (e.g., see Chan and Tsay (1998)), seasonal ARMA(1, 1) models (e.g., see Andrews, Liu, and Ploberger (1998)), models with correlated random coefficients (e.g., see Andrews (2001)), (generalized autoregressive conditional heteroskedasticity) GARCH(p, q) models, and time series models with nonlinear deterministic time trends of the form t^π or $(t^\pi - 1)/\pi$.[12]

[11]See Supplemental Appendix A in the Supplemental Material (Andrews and Cheng (2012)) for a brief discussion.

[12]Nonlinear time trends can be analyzed asymptotically in the framework considered in this paper via sample size rescaling, that is, by considering $(t/n)^\pi$ or $((t/n)^\pi - 1)/\pi$ (e.g., see Andrews and McDermott (1995)).

Not all models with lack of identification at some points in the parameter space fall into the class of models considered here. The models considered here must satisfy a set of criterion function (stochastic) quadratic approximation conditions, as described in more detail below, that do not apply to some models of interest. For example, abrupt transition structural change models, (unobserved) regime switching models, and abrupt transition threshold autoregressive models are not covered by the results of the present paper; for example, see Picard (1985), Chan (1993), Bai (1997), Hansen (2000), Liu and Shao (2003), Elliott and Müller (2007, 2008), Qu and Perron (2007), and Drton (2009) for analyses of these models. In addition, the criterion functions considered here depend on the parameter that determines the strength of identification. This differs from the criterion functions considered in the weak IV literature.

Next, we discuss the literature that is related to this paper. Cheng (2008) considers a nonlinear regression model with multiple nonlinear regressors and, hence, multiple sources of lack of identification. Here we consider a single source of lack of identification, but cover a much wider variety of models.[13]

In the models considered in this paper, a test of $H_0: \beta = 0$ versus $H_1: \beta \neq 0$, is a test for which π is a nuisance parameter that is unidentified under the null hypothesis. Testing problems of this type have been considered in the literature; for example, see Davies (1977, 1987), Andrews and Ploberger (1994, 1995), Hansen (1996), and Cho, Ishida, and White (2011). In contrast, we consider a full range of nonlinear hypotheses concerning (β, ζ, π) and CS's, where β can be 0, close to 0, or far from 0. When the null hypothesis involves (ζ, π), the identification scenario is substantially more complicated than when H_0 is $\beta = 0$.

The weak instrumental variable (IV) literature (e.g., see Nelson and Startz (1990), Dufour (1997), Staiger and Stock (1997), Stock and Wright (2000), Kleibergen (2002, 2005), Moreira (2003), and other papers referenced in Andrews and Stock (2007)) is related to the present paper because it considers weak identification. In the weak IV literature, the criterion functions considered do not have the parameters that are the source of weak identification as arguments. Thus, in linear IV models, the reduced-form parameters are not arguments of the criterion function. Similarly, in Stock and Wright (2000), which applies to nonlinear models, high-level conditions are placed on the population moment functions under which the IV's are weak for some parameters. On the other hand, in the present paper, the potential source of weak identification is an explicit part of the model.[14] In consequence, the present paper and the weak IV literature are complements.

[13] In addition, the treatment of the nonlinear regression model here allows for a whole class of error distributions, whereas Cheng (2008) considers a single error distribution.

[14] To help clarify the differences, we show in Supplemental Appendix E that Stock and Wright's (2000) Assumption C fails in the nonlinear regression model when a nonlinear regression parameter is weakly identified due to its multiplicative coefficient being close to zero.

However, in one case there is an overlap. The criterion function for the limited information maximum likelihood (LIML) estimator in the linear IV regression model can be written as either (i) a function of the parameters in the structural equation plus the parameters in the accompanying reduced-form equations, which fits the framework of the present paper, or (ii) a function of the structural equation parameters only via concentrating out the reduced-form parameters, as in Anderson and Rubin (1949) and Staiger and Stock (1997). This permits the comparison of the CLR test with the robust tests introduced here, as discussed above.

The finite-sample results of Dufour (1997) and Gleser and Hwang (1987) for CS's and tests are applicable to the models considered in this paper.[15]

Antoine and Renault (2009, 2010) and Caner (2010) consider generalized method of moment (GMM) estimation with instruments that lie in what we call the semi-strong category. Their emphasis is on asymptotic efficiency with semi-strong instruments, rather than the behavior of statistics across the full range of strengths of identification as is considered here.

In likelihood scenarios, Lee and Chesher (1986) consider Lagrange multiplier (LM) tests and Rotnitzky, Cox, Bottai, and Robins (2000) consider ML estimators and likelihood ratio tests, when the model is identified at all parameter values, but the information matrix is singular at some parameter values, such as those in the null hypothesis. This is a different scenario than considered in the present paper, because the present paper considers scenarios where identification fails at some parameter values in the parameter space, which causes the information matrix in likelihood scenarios to be singular at these parameter values. In recent papers, I. Andrews and Mikusheva (2011) and Qu (2011) consider a LM statistic in a likelihood context with weak identification.

Nelson and Startz (2007) introduce the zero-information-limit condition, which applies to the models considered in this paper, and discuss its implications. Ma and Nelson (2008) consider tests based on linearization for models of the type considered in this paper. Neither of these papers establishes the large sample properties of estimators, tests, and CS's along the lines given in this paper.

Sargan (1983) provides asymptotic results for linear-in-variables and nonlinear-in-parameters simultaneous equations models in which some parameters are unidentified. Phillips (1989) and Choi and Phillips (1992) provide finite-sample and asymptotic results for linear simultaneous equations and linear spurious regression models in which some parameters are unidentified. Their results do not overlap very much with those in this paper because the present paper is focused on nonlinear models. Their asymptotic results are pointwise

[15]This paper considers the case where the potentially unidentified parameter π lies in a bounded set Π. In this case, Corollary 3.4 of Dufour (1997) implies that if the diameter of a CS for π is as large as the diameter of Π with probability less than $1-2\alpha$, then the CS has (exact) size less than $1-\alpha$ (under certain assumptions).

in the parameters, which covers the unidentified and strongly identified categories, but not the weakly identified and semi-strongly identified categories described above.

Supplemental Appendix E (in the Supplemental Material (Andrews and Cheng (2012))) applies the results of the present paper to the nonlinear regression model with i.i.d. or stationary and ergodic regressors. One also can apply the approach of this paper to the case where the regressors are integrated. In this case, the general results given below do not apply directly. However, by using the asymptotics for nonlinear and nonstationary processes developed by Park and Phillips (1999, 2001), the approach goes through, as shown recently by Shi and Phillips (2011).[16]

The remainder of the paper is organized as follows. Section 2 introduces the extremum estimators, criterion functions, tests, confidence sets, and drifting sequences of distributions considered in the paper. Section 3 states the high-level assumptions employed and provides the asymptotic results for the extremum estimators. Section 4 establishes the asymptotic distributions of t and QLR statistics, and determines the asymptotic size of standard t and QLR CS's. Section 5 introduces methods of constructing robust tests and CS's whose asymptotic size equals their nominal size, and applies them to t and QLR tests and CS's. This section also includes the comparison of the asymptotic power of one of the robust tests with the CLR test in the linear IV regression model. Section 6 provides asymptotic and finite-sample numerical results for the ARMA$(1, 1)$ model. The Supplemental Material contains all the appendices. Supplemental Appendix A gives a verbal description of the steps in the proofs of the results in Sections 3–5 and sufficient conditions for some of the high-level conditions stated in Section 3. Supplemental Appendix B provides proofs of the results given in Sections 3–5. Supplemental Appendix C verifies the assumptions of the paper for the ARMA example. Supplemental Appendix D provides additional Monte Carlo simulation results for the ARMA example. Supplemental Appendices E and F verify the assumptions of the paper for the nonlinear regression and linear IV regression models, respectively.

AC2 provides primitive sufficient conditions for the high-level assumptions of this paper for the class of estimators based on sample averages that are smooth functions of the parameter θ, which includes ML and least squares (LS) estimators. AC3 provides sufficient conditions for the high-level assumptions for the class of GMM estimators and provides general results for Wald tests.

All limits below are taken "as $n \to \infty$." Let $X_n(\pi) = o_{p\pi}(1)$ mean that $\sup_{\pi \in \Pi} \|X_n(\pi)\| = o_p(1)$, where $\|\cdot\|$ denotes the Euclidean norm. Let "for all $\delta_n \to 0$" abbreviate "for all sequences of positive scalar constants $\{\delta_n : n \geq 1\}$

[16]Shi and Phillips (2011) employs the same method of computing asymptotic size and of constructing identification-robust CS's as was introduced in an early version of this paper and in Cheng (2008).

for which $\delta_n \to 0$." Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues, respectively, of a matrix A. All vectors are column vectors. For notational simplicity, we often write (a, b) instead of $(a', b')'$ for vectors a and b. Also, for a function $f(c)$ with $c = (a, b)$ $(= (a', b')')$, we often write $f(a, b)$ instead of $f(c)$. Let 0_d denote a d-vector of zeros. Because it arises frequently, we let 0 denote a d_β-vector of zeros, where d_β is the dimension of a parameter β. Let $\Rightarrow$ denote weak convergence of a sequence of stochastic processes indexed by $\pi \in \Pi$ for some space Π.

2. ESTIMATOR AND CRITERION FUNCTION

2.1. *Extremum Estimators*

By definition, the estimator $\widehat{\theta}_n$ (approximately) minimizes a criterion function $Q_n(\theta)$ over an "optimization parameter space" Θ:[17]

$$\text{(2.1)} \qquad \widehat{\theta}_n \in \Theta \quad \text{and} \quad Q_n(\widehat{\theta}_n) = \inf_{\theta \in \Theta} Q_n(\theta) + o(n^{-1}).$$

The function $Q_n(\theta)$ depends on the observations $\{W_i : i \leq n\}$, which may be i.i.d., independent and nonidentically distributed (i.n.i.d.), or temporally dependent.[18]

As stated above, θ is partitioned into three subvectors:

$$\text{(2.2)} \qquad \theta = (\beta, \zeta, \pi) = (\psi, \pi), \quad \text{where} \quad \psi = (\beta, \zeta).$$

The parameter $\pi \in R^{d_\pi}$ is unidentified when $\beta = 0$ $(\in R^{d_\beta})$. The parameter $\psi = (\beta, \zeta) \in R^{d_\psi}$ is always identified. The parameter $\zeta \in R^{d_\zeta}$ does not effect the identification of π.

The argument θ of the criterion function need not determine the distribution of the data. We introduce an additional parameter ϕ such that $\gamma = (\theta, \phi)$ completely determines the distribution of the data.[19] The true distribution of

[17]The $o(n^{-1})$ term in (2.1), and in (3.2) and (3.3) below, is a fixed sequence of constants that does not depend on the true parameter $\gamma \in \Gamma$ and does not depend on π in (3.2). The $o(n^{-1})$ term makes it clear that the infima in these equations need not be achieved exactly. This allows for some numerical inaccuracy in practice and also circumvents the issue of the existence of parameter values that achieve the infima. In contrast to many results in the extremum estimator literature, the $o(n^{-1})$ term is not a random $o_p(n^{-1})$ term here.

[18]The indices i and t are interchangeable in this paper. For the general results and cross section examples, the observations are indexed by i $(= 1, \ldots, n)$. To conform with standard notation, the observations are indexed by t $(= 1, \ldots, n$ or $= -r, \ldots, n$ for some $r \geq 0)$ in time series examples, such as the ARMA(1, 1) example.

[19]In a nonlinear regression model estimated by least squares, θ indexes the regression function and possibly a finite-dimensional feature of the distribution of the errors, such as its variance, and ϕ indexes the remaining characteristics of the distribution of the errors, which may be infinite dimensional. In an unconditional likelihood scenario, no parameter ϕ appears. In a

the observations $\{W_i : i \leq n\}$ is denoted F_γ where $\gamma \in \Gamma$. We let P_γ and E_γ denote probability and expectation under F_γ.

The parameter space Γ for the true parameter γ, referred to as the "true parameter space," is assumed to be compact and of the form

$$\Gamma = \{\gamma = (\theta, \phi) : \theta \in \Theta^*, \phi \in \Phi^*(\theta)\}, \tag{2.3}$$

where the true parameter space for θ, Θ^*, is a compact subset of R^{d_θ} and $\Phi^*(\theta) \subset \Phi^*\ \forall \theta \in \Theta^*$ for some compact metric space Φ^* with a metric that induces weak convergence of the bivariate distributions (W_i, W_{i+m}) for all $i, m \geq 1$.[20–22]

2.2. *Confidence Sets and Tests*

We are interested in the effect of lack of identification or weak identification on the behavior of the extremum estimator $\widehat{\theta}_n$. In addition, we are interested in its effects on CS's for various functions $r(\theta)$ of θ and on tests of null hypotheses of the form $H_0 : r(\theta) = v$.

A CS is obtained by inverting a test. For example, a nominal $1 - \alpha$ CS for $r(\theta)$ is

$$\text{CS}_n = \{v : \mathcal{T}_n(v) \leq c_{n,1-\alpha}(v)\}, \tag{2.4}$$

where $\mathcal{T}_n(v)$ is a test statistic, such as a t, Wald, or QLR statistic, and $c_{n,1-\alpha}(v)$ is a critical value for testing $H_0 : r(\theta) = v$. Critical values considered in this paper may depend on the null value v of $r(\theta)$ as well as on the sample size n. The coverage probability of a CS for $r(\theta)$ is

$$P_\gamma(r(\theta) \in \text{CS}_n) = P_\gamma\big(\mathcal{T}_n(r(\theta)) \leq c_{n,1-\alpha}(r(\theta))\big). \tag{2.5}$$

This paper focuses on the smallest finite-sample coverage probability of a CS over the parameter space, that is, the finite-sample size of the CS. It is

conditional likelihood scenario, with conditioning variables $\{X_i : i \geq 1\}$, ϕ indexes the distribution of $\{X_i : i \geq 1\}$. In a moment condition model, θ is a finite-dimensional parameter that appears in the moment functions and ϕ indexes those aspects of the distribution of the observations that are not determined by θ.

[20]The true parameter space Θ^* is the space of parameter values that the researcher specifies as including the true value. The optimization parameter space Θ is the space over which the researcher optimizes the sample criterion function. For reasons stated below (see the discussion preceding Assumption B1), we allow for a difference between Θ and Θ^*.

[21]The metric d_{Φ^*} on Φ^* must satisfy the condition if $\gamma \to \gamma_0$, then (W_i, W_{i+m}) under γ converges in distribution to (W_i, W_{i+m}) under γ_0. Note that Γ is a metric space with metric $d_\Gamma(\gamma_1, \gamma_2) = \|\theta_1 - \theta_2\| + d_{\Phi^*}(\phi_1, \phi_2)$, where $\gamma_j = (\theta_j, \phi_j) \in \Gamma$ for $j = 1, 2$.

[22]The asymptotic results below give uniformity results over the parameter space Γ. If one has a noncompact parameter space Φ_1^* for the parameter ϕ, instead of Φ^*, then one can apply the results established here to show that the uniformity results hold for all compact subsets Φ^* of Φ_1^* that satisfy the given conditions.

approximated by the asymptotic size, which is defined to be

$$\text{(2.6)} \qquad \text{AsySz} = \liminf_{n\to\infty} \inf_{\gamma\in\Gamma} P_\gamma(r(\theta)\in \text{CS}_n) = \liminf_{n\to\infty} \inf_{\gamma\in\Gamma} P_\gamma\big(\mathcal{T}_n(r(\theta)) \le c_{n,1-\alpha}(r(\theta))\big).$$

For a test, we are interested in its maximum null rejection probability, which is the size of the test. A test's asymptotic size is an approximation to the latter. The test's null rejection probability is $P_\gamma(\mathcal{T}_n(v) > c_{n,1-\alpha}(v))$ for $\gamma = (\theta, \phi) \in \Gamma$ with $r(\theta) = v$ and its asymptotic size is $\text{AsySz} = \limsup_{n\to\infty} \sup_{\gamma\in\Gamma: r(\theta)=v} P_\gamma(\mathcal{T}_n(v) > c_{n,1-\alpha}(v))$.

2.3. *Drifting Sequences of Distributions*

In (2.6), the uniformity over $\gamma \in \Gamma$ for any given sample size n is crucial for the asymptotic size to be a good approximation to the finite-sample size. The value of γ at which the finite-sample size of a CS or test is attained often varies with the sample size. Therefore, to determine the asymptotic size, we need to derive the asymptotic distribution of the test statistic $\mathcal{T}_n(v_n)$ under sequences of true parameters $\gamma_n = (\theta_n, \phi_n)$ and $v_n = r(\theta_n)$ that may depend on n.[23] Similarly, to investigate the finite-sample behavior of the extremum estimator under weak identification, we need to consider its asymptotic behavior under drifting sequences of true distributions, as in Staiger and Stock (1997), Stock and Wright (2000), and numerous other papers that consider weak instruments.

Suppose the true value of the parameter is $\theta_n = (\beta_n, \zeta_n, \pi_n)$ for $n \ge 1$, where n indexes the sample size. The behavior of extremum estimators and tests depends on the magnitude of $\|\beta_n\|$, and varies across the three categories of sequences $\{\beta_n : n \ge 1\}$ defined in Table I.[24] In consequence, the following sequences $\{\gamma_n\}$ are key:

$$\text{(2.7)} \qquad \Gamma(\gamma_0) = \{\{\gamma_n \in \Gamma : n \ge 1\} : \gamma_n \to \gamma_0 \in \Gamma\},$$
$$\Gamma(\gamma_0, 0, b) = \big\{\{\gamma_n\} \in \Gamma(\gamma_0) : \beta_0 = 0 \text{ and } n^{1/2}\beta_n \to b \in (R \cup \{\pm\infty\})^{d_\beta}\big\},$$

[23]Drifting sequences of parameters have been shown to play a crucial role in the literature on the (uniform) asymptotic size properties of tests and CS's when the statistics of interest display discontinuities in their pointwise asymptotic distributions; see Mikusheva (2007), Andrews and Guggenberger (2009, 2010), and Andrews, Cheng, and Guggenberger (2009). The situation considered here is an example of the latter phenomenon.

[24]Hahn and Kuersteiner (2002) and Antoine and Renault (2009, 2010) refer to sequences in our *semi-strong* category as *nearly weak*. For this paper at least, we prefer our terminology because estimators are consistent and asymptotically normal under semi-strong sequences, just as under sequences in the strong category. The only difference is that their rate of convergence is slower.

TABLE I

IDENTIFICATION CATEGORIES

Category	$\{\beta_n\}$ Sequence	Identification Property of π
I(a)	$\beta_n = 0 \ \forall n \geq 1$	Unidentified
I(b)	$\beta_n \neq 0$ and $n^{1/2}\beta_n \to b \in R^{d_\beta}$ (and, hence, $\|\beta_n\| = O(n^{-1/2})$)	Weakly identified
II	$\beta_n \to 0$ and $n^{1/2}\|\beta_n\| \to \infty$	Semi-strongly identified
III	$\beta_n \to \beta_0 \neq 0$	Strongly identified

$$\Gamma(\gamma_0, \infty, \omega_0) = \big\{\{\gamma_n\} \in \Gamma(\gamma_0) : n^{1/2}\|\beta_n\| \to \infty \text{ and } \beta_n/\|\beta_n\| \to \omega_0 \in R^{d_\beta}\big\},$$

where $\gamma_0 = (\beta_0, \zeta_0, \pi_0, \phi_0)$ and $\gamma_n = (\beta_n, \zeta_n, \pi_n, \phi_n)$.[25]

The sequences in $\Gamma(\gamma_0, 0, b)$ are in categories I and II, and are sequences for which $\{\beta_n\}$ is *close* to 0: $\beta_n \to 0$. When $\|b\| < \infty$, $\{\beta_n\}$ is within $O(n^{-1/2})$ of 0 and the sequence is in category I. The sequences in $\Gamma(\gamma_0, \infty, \omega_0)$ are in categories II and III and are more *distant* from $\beta = 0$: $n^{1/2}\|\beta_n\| \to \infty$. The sets $\Gamma(\gamma_0, 0, b)$ and $\Gamma(\gamma_0, \infty, \omega_0)$ are *not* disjoint. Both contain sequences in category II.

Throughout the paper, we use the terminology "under $\{\gamma_n\} \in \Gamma(\gamma_0)$" to mean "when the true parameters are $\{\gamma_n\} \in \Gamma(\gamma_0)$ for any $\gamma_0 \in \Gamma$"; "under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$" to mean "when the true parameters are $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ for any $\gamma_0 \in \Gamma$ with $\beta_0 = 0$ and any $b \in (R \cup \{\pm\infty\})^{d_\beta}$"; and "under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$" to mean "when the true parameters are $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ for any $\gamma_0 \in \Gamma$ and any $\omega_0 \in R^{d_\beta}$ with $\|\omega_0\| = 1$."

Lemma 2.1 below shows that the AsySz of a sequence of CS's is determined by the asymptotic coverage probabilities of the CS's under the drifting sequences of distributions in $\Gamma(\gamma_0, 0, b)$ and $\Gamma(\gamma_0, \infty, \omega_0)$. We emphasize that asymptotic coverage probabilities for sequences in all three categories I–III, including the semi-strong category II, are needed to establish the asymptotic size of a sequence of CS's.

Consider the CS for $r(\theta)$ in (2.4). Denote the coverage probability of the CS under $\gamma_n = (\theta_n, \phi_n)$ by $\mathrm{CP}_n(\gamma_n) = P_{\gamma_n}(\mathcal{T}_n(r(\theta_n)) \leq c_{n,1-\alpha}(r(\theta_n)))$. Let

$$h = (b, \gamma_0) \quad \text{and} \quad H = \{h = (b, \gamma_0) : \|b\| < \infty, \gamma_0 \in \Gamma \text{ with } \beta_0 = 0\}. \tag{2.8}$$

ASSUMPTION ACP: (i) *For any* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $\|b\| < \infty$, $\mathrm{CP}_n(\gamma_n) \to \mathrm{CP}(h)$ *for some* $\mathrm{CP}(h) \in [0, 1]$, *where* $h = (b, \gamma_0) \in H$.

[25]Note that the 0 in $\Gamma(\gamma_0, 0, b)$ and the ∞ in $\Gamma(\gamma_0, \infty, \omega_0)$ stand for different things. In the former, $\beta_0 = 0$; in the latter, $n^{1/2}\|\beta_n\| \to \infty$.

(ii) *For any* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $\liminf_{n\to\infty} \mathrm{CP}_n(\gamma_n) \geq \mathrm{CP}_\infty$ *for some* $\mathrm{CP}_\infty \in [0, 1]$.
(iii) *For some* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $\mathrm{CP}_n(\gamma_n) \to \mathrm{CP}_\infty$.
(iv) *For some* $\delta > 0$, $\gamma = (\beta, \zeta, \pi, \phi) \in \Gamma$ *with* $0 \leq \|\beta\| < \delta$ *implies that* $\widetilde{\gamma} = (\widetilde{\beta}, \zeta, \pi, \phi) \in \Gamma$ *for all* $\widetilde{\beta} \in R^{d_\beta}$ *with* $0 \leq \|\widetilde{\beta}\| < \delta$.

Here ACP abbreviates asymptotic coverage probability.

LEMMA 2.1: *Suppose Assumption* ACP *holds*. *Then*

$$\mathrm{AsySz} = \min\{\inf_{h\in H} \mathrm{CP}(h), \mathrm{CP}_\infty\}.$$

COMMENTS: (i) Assumption ACP is verified below for standard t and QLR CS's, as well as several CS's that are robust to weak identification. Lemma 2.1 then gives their AsySz. Note that Assumption ACP(ii) requires asymptotic results for the semi-strongly identified category II sequences, not just the strongly identified category III sequences.

(ii) The sets $\Gamma(\gamma_0, 0, b)$ and $\Gamma(\gamma_0, \infty, \omega_0)$ are distinguished by whether $n^{1/2}\|\beta_n\| \to \|b\|$ with $\|b\| < \infty$ or $\|b\| = \infty$. Similarly, Assumptions ACP(i) and ACP(ii) and (iii) are distinguished by $\|b\| < \infty$ and $\|b\| = \infty$. The reason this distinction arises and is important is that the asymptotic behavior of the normalized (generalized) stochastic first derivative of the criterion function $Q_n(\theta)$ depends on whether $\|b\| < \infty$ or $\|b\| = \infty$. If $\|b\| < \infty$, its limit is the sum of deterministic and stochastic terms, because the signal and noise are of the same order of magnitude. If $\|b\| = \infty$, its limit is deterministic, because the signal dominates the noise (see (3.7) below).

(iii) Lemma 2.1 is proved by showing that one can reduce uniform coverage probability results to coverage probability results under suitable subsequences. Then one shows that results under such subsequences are implied by results under suitable full sequences. The proof follows the lines of the argument in Andrews and Guggenberger (2010).

3. ASSUMPTIONS AND ESTIMATION RESULTS

3.1. *Parameter Space Assumptions*

First, we specify conditions on the parameter spaces Θ and Γ. To obtain asymptotic size results for tests and CS's, the parameter space must be specified precisely. Without loss of generality (w.l.o.g.), the optimization parameter space Θ can be written as

$$\Theta = \{\theta = (\psi, \pi) : \psi \in \Psi(\pi), \pi \in \Pi\}, \quad \text{where}$$
$$\Pi = \{\pi : (\psi, \pi) \in \Theta \text{ for some } \psi\},$$
$$\Psi(\pi) = \{\psi : (\psi, \pi) \in \Theta\} \quad \text{for} \quad \pi \in \Pi. \tag{3.1}$$

Allowing $\Psi(\pi)$ to depend on π is needed in the ARMA$(1, 1)$ example, among others.[26]

We consider the case where the optimization parameter space Θ includes Θ^* in its interior (Assumption B1(i) below). Because Θ is user selected, often this can be accomplished by the choice of Θ. Given $\text{int}(\Theta) \supset \Theta^*$, the true value of θ cannot lie on the boundary of the optimization parameter space. In consequence, the asymptotic distribution of $\widehat{\theta}_n$ is not affected by boundary constraints for any sequence of true parameters in Θ^*. This allows us to focus in this paper on the effects of weak identification, independently from boundary constraints, on the behavior of estimators, tests, and CS's.[27]

Define $\Theta^*_\delta = \{\theta \in \Theta^* : \|\beta\| < \delta\}$, where Θ^* is the true parameter space for θ. The optimization parameter space Θ satisfies the following assumption.

ASSUMPTION B1: (i) $\text{int}(\Theta) \supset \Theta^*$.

(ii) *For some* $\delta > 0$, $\Theta \supset \{\beta \in R^{d_\beta} : \|\beta\| < \delta\} \times \mathcal{Z}^0 \times \Pi \supset \Theta^*_\delta$ *for some nonempty open set* $\mathcal{Z}^0 \subset R^{d_\zeta}$ *and* Π *as in* (3.1).

(iii) Π *is compact*.

Assumption B1(ii) ensures that Θ is compatible with Assumptions C1, C3, and C5 below.[28]

The true parameter space Γ satisfies the next assumption.

ASSUMPTION B2: (i) Γ *is compact and* (2.3) *holds*.

(ii) *For some* $\delta > 0$, $\gamma = (\beta, \zeta, \pi, \phi) \in \Gamma$ *with* $0 \leq \|\beta\| < \delta$ *implies that* $\widetilde{\gamma} = (\widetilde{\beta}, \zeta, \pi, \phi) \in \Gamma$ *for all* $\widetilde{\beta} \in R^{d_\beta}$ *with* $0 \leq \|\widetilde{\beta}\| < \delta$.

(iii) *For* $\delta > 0$ *as in* (ii), $\exists \gamma = (\beta, \zeta, \pi, \phi) \in \Gamma$ *with* $0 < \|\beta\| < \delta$.

Assumption B2(ii) ensures that Assumption ACP(iv) holds. Assumptions B2(ii) and (iii) guarantee that there exist elements γ of Γ whose β values are nonzero but are arbitrarily close to zero, which is the region of near lack of identification, and that Γ is compatible with Assumption C5 below.

[26]We write Θ in terms of the sets Π and $\Psi(\pi)$, rather than sets Ψ and $\Pi(\psi)$, because below we carry out quadratic expansions of $Q_n(\psi, \pi)$ w.r.t. ψ for each $\pi \in \Pi$ and this yields stochastic processes that are indexed by the fixed set Π and that converge weakly as processes on Π.

[27]If the true and optimization parameters spaces both equal a set Θ, then the uniform results of this paper apply to any subset Θ^* of Θ that satisfies the conditions listed below, but they do not apply to the entire true parameter space Θ because of boundary effects.

[28]Assumption B1(iii) is used to show that certain continuous functions on Π introduced in Assumptions C6 and C7 below, which have unique minima on Π, satisfy "identifiable uniqueness" properties. Assumption B1(iii) could be avoided by imposing identifiable uniqueness properties directly in Assumptions C6 and C7.

3.2. *Concentrated Estimator and Probability Limit Results*

Define the concentrated extremum estimator $\widehat{\psi}_n(\pi)$ ($\in \Psi(\pi)$) of ψ for given $\pi \in \Pi$ by

$$(3.2) \qquad Q_n(\widehat{\psi}_n(\pi), \pi) = \inf_{\psi \in \Psi(\pi)} Q_n(\psi, \pi) + o(n^{-1}).$$

Let $Q_n^c(\pi)$ denote the concentrated sample criterion function $Q_n(\widehat{\psi}_n(\pi), \pi)$. Define an extremum estimator $\widehat{\pi}_n$ ($\in \Pi$) by

$$(3.3) \qquad Q_n^c(\widehat{\pi}_n) = \inf_{\pi \in \Pi} Q_n^c(\pi) + o(n^{-1}).$$

We assume that the extremum estimator $\widehat{\theta}_n$ in (2.1) can be written as $\widehat{\theta}_n = (\widehat{\psi}_n(\widehat{\pi}_n), \widehat{\pi}_n)$.[29]

Next, we specify the limit of the sample criterion function $Q_n(\theta)$ along drifting sequences of true parameters $\{\gamma_n\} \in \Gamma(\gamma_0)$ whose limit is $\gamma_0 \in \Gamma$ and determine the probability limit of $\widehat{\theta}_n$.

ASSUMPTION B3: (i) *For some nonstochastic real-valued function* $Q(\theta; \gamma_0)$ *on* $\Theta \times \Gamma$, $\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta; \gamma_0)| \to_p 0$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0)$ $\forall \gamma_0 \in \Gamma$.

(ii) *When* $\beta_0 = 0$, *for every neighborhood* Ψ_0 ($\subset R^{d_\psi}$) *of* $\psi_0 = (\beta_0, \zeta_0)$, $\inf_{\pi \in \Pi}(\inf_{\psi \in \Psi(\pi)/\Psi_0} Q(\psi, \pi; \gamma_0) - Q(\psi_0, \pi; \gamma_0)) > 0$ $\forall \gamma_0 = (\psi_0, \pi_0, \phi_0) \in \Gamma$.

(iii) *When* $\beta_0 \neq 0$, *for every neighborhood* Θ_0 ($\subset \Theta$) *of* $\theta_0 = (\beta_0, \zeta_0, \pi_0)$, $\inf_{\theta \in \Theta/\Theta_0} Q(\theta; \gamma_0) - Q(\theta_0; \gamma_0) > 0$ $\forall \gamma_0 = (\theta_0, \phi_0) \in \Gamma$.

Assumption B3(i) defines the (asymptotic) population criterion function $Q(\theta; \gamma_0)$. Assumption B3(ii) provides a condition for the identification of β and ζ despite the nonidentification of π when $\beta_0 = 0$. Uniformity over Π is required due to the nonidentification of π. A condition of this type also is used in Andrews (1993) for the uniform consistency of a family of estimators. Assumption B3(iii) is a standard identification condition for θ when $\beta_0 \neq 0$. A condition of this sort is verified for various extremum estimators in Newey and McFadden (1994).

A set of primitive sufficient conditions for Assumption B3(ii) and (iii) is given in Assumption B3* in Supplemental Appendix A.

LEMMA 3.1: *Suppose Assumptions* A *and* B3 *hold. Under* $\{\gamma_n\} \in \Gamma(\gamma_0)$, *where* $\gamma_0 = (\beta_0, \zeta_0, \pi_0, \phi_0)$, *there are two alternatives*:

(a) *When* $\beta_0 = 0$, $\sup_{\pi \in \Pi} \|\widehat{\psi}_n(\pi) - \psi_n\| \to_p 0$ *and* $\widehat{\psi}_n - \psi_n \to_p 0$.

(b) *When* $\beta_0 \neq 0$, $\widehat{\theta}_n - \theta_n \to_p 0$.

COMMENT: When $\beta_0 = 0$, the asymptotic behavior of $\widehat{\pi}_n$ is determined below.

[29] If (3.2) and (3.3) hold and $\widehat{\theta}_n = (\widehat{\psi}_n(\widehat{\pi}_n), \widehat{\pi}_n)$, then (2.1) automatically holds.

3.3. *Close to $\beta = 0$ Assumptions and Estimation Results*

The following Assumptions C1–C8 are used to determine the asymptotic distributions of estimators and test statistics under sequences of true parameters $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$ and to establish the consistency of $\widehat{\pi}_n$ under sequences $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| = \infty$. The "C" denotes that the sequences of parameters $\{\gamma_n\}$ considered are *close* to the point of nonidentification.

The first assumption, Assumption C1, requires that the criterion function $Q_n(\theta)$ has a stochastic quadratic expansion in ψ around the nonidentification point $\psi_{0,n} = (0, \zeta_n)$ uniformly in $\pi \in \Pi$. Assumptions C2 and C3 concern the behavior of the (generalized) first derivative in the expansion. Assumption C4 concerns the behavior of the (generalized) second derivative. Assumptions C5 and C7 arise because the quadratic expansion is about the nonidentification point $\psi_{0,n}$, rather than the true value ψ_n. Assumptions C6–C8 are used when determining the asymptotic behavior of $\widehat{\pi}_n$.

We now define a sequence of scalar constants $\{a_n(\gamma_n) : n \geq 1\}$ that provides the normalization required so that the (generalized) first derivative in the quadratic expansion in Assumption C1 is nondegenerate asymptotically.[30] These constants appear in the conditions on the remainder term of the approximation in Assumption C1. Define

$$(3.4) \qquad a_n(\gamma_n) = \begin{cases} n^{1/2}, & \text{if } \{\gamma_n\} \in \Gamma(\gamma_0, 0, b) \text{ and } \|b\| < \infty, \\ \|\beta_n\|^{-1}, & \text{if } \{\gamma_n\} \in \Gamma(\gamma_0, 0, b) \text{ and } \|b\| = \infty. \end{cases}$$

Note that $\|\beta_n\|^{-1} < n^{1/2}$ for n large when $\|b\| = \infty$, because $n^{1/2}\|\beta_n\| \to \infty$.[31] Hence, $a_n(\gamma_n) \leq n^{1/2}$ for n large.

ASSUMPTION C1: *Under* $\{\gamma_n = (\beta_n, \zeta_n, \pi_n, \phi_n)\} \in \Gamma(\gamma_0, 0, b)$, *for some* $\delta > 0$, $\forall \theta = (\psi, \pi) \in \Theta_\delta = \{\theta \in \Theta : \|\beta\| < \delta\}$, *the following statements hold*:

(i) *The sample criterion function* $Q_n(\psi, \pi)$ *has a quadratic expansion in* ψ *around* $\psi_{0,n} = (0, \zeta_n)$ *for given* π,

$$Q_n(\psi, \pi) = Q_n(\psi_{0,n}, \pi) + D_\psi Q_n(\psi_{0,n}, \pi)'(\psi - \psi_{0,n}) + \frac{1}{2}(\psi - \psi_{0,n})' D_{\psi\psi} Q_n(\psi_{0,n}, \pi)(\psi - \psi_{0,n}) + R_n(\psi, \pi),$$

where $D_\psi Q_n(\psi_{0,n}, \pi) \in R^{d_\psi}$ *is a stochastic generalized first partial-derivative vector, and* $D_{\psi\psi} Q_n(\psi_{0,n}, \pi) \in R^{d_\psi \times d_\psi}$ *is a generalized second partial-derivative matrix that is symmetric and may be stochastic or nonstochastic*.

(ii) *The remainder*, $R_n(\psi, \pi)$, *satisfies*

$$\sup_{\psi \in \Psi(\pi) : \|\psi - \psi_{0,n}\| \leq \delta_n} \frac{|a_n^2(\gamma_n) R_n(\psi, \pi)|}{(1 + \|a_n(\gamma_n)(\psi - \psi_{0,n})\|)^2} = o_{p\pi}(1)$$

[30] See Lemma 9.1 in Supplemental Appendix B.

[31] The quantity $a_n(\gamma_n)$ actually depends on the entire sequence $\{\gamma_n\}$ because b depends on $\{\gamma_n\}$.

for all constants $\delta_n \to 0$,

(iii) $D_\zeta Q_n(\theta)$ *and* $D_{\zeta\zeta} Q_n(\theta)$ *do not depend on* π *when* $\beta = 0$, *where* $\theta = (\beta, \zeta, \pi) \in \Theta$, $D_\zeta Q_n(\theta)$ *denotes the last* d_ζ *elements of* $D_\psi Q_n(\theta)$, *and* $D_{\zeta\zeta} Q_n(\theta)$ *is the lower* $d_\zeta \times d_\zeta$ *block of* $D_{\psi\psi} Q_n(\theta)$.

Because the expansion in Assumption C1 is about the point of lack of identification $\psi_{0,n}$, rather than the true value ψ_n, the leading term $Q_n(\psi_{0,n}, \pi)$ does not depend on π by Assumption A. This is key. It implies that $\widehat{\theta}_n = (\widehat{\psi}_n, \widehat{\pi}_n)$ not only minimizes $Q_n(\psi, \pi)$, but also $Q_n(\psi, \pi) - Q_n(\psi_{0,n}, \pi)$. The latter has the quadratic expansion in Assumption C1 with linear and quadratic terms whose asymptotic properties one can determine using Assumptions C2–C5 below.

Sufficient conditions for Assumption C1 when $Q_n(\theta)$ is a sample average that is smooth in θ are given in Lemma 8.6 in Supplemental Appendix A. In this case, $D_\psi Q_n(\theta)$ and $D_{\psi\psi} Q_n(\theta)$ are the pointwise partial and second partial derivatives of $Q_n(\theta)$. For the nonsmooth sample average case, sufficient conditions are given in Lemma 8.7 in Supplemental Appendix A. In this case, $D_\psi Q_n(\theta)$ is a "stochastic derivative" of $Q_n(\theta)$, which typically equals the pointwise derivative for points where the latter exists, and $D_{\psi\psi} Q_n(\theta)$ is the (nonstochastic) second partial derivative of the expected value of $Q_n(\theta)$. This case covers quantile estimators and ML and LS estimators in continuous, but not smooth, threshold autoregressive models, as in Chan and Tsay (1998). Sufficient conditions for Assumption C1 when $Q_n(\theta)$ is a GMM or minimum distance (MD) criterion function, smooth or nonsmooth in θ, are given in AC3.

If $D_\psi Q_n(\theta)$ and $D_{\psi\psi} Q_n(\theta)$ are the pointwise partial and second partial derivatives of $Q_n(\theta)$, then Assumption C1(iii) is implied by Assumption A. Otherwise, in the presence of Assumption A, Assumption C1(iii) is not restrictive.

Note that Assumption C1 is compatible with semiparametric estimators.

The (generalized) first derivative of $Q_n(\theta)$ w.r.t. ψ is assumed to satisfy the following assumption.

ASSUMPTION C2: (i) $D_\psi Q_n(\theta)$ *takes the form* $D_\psi Q_n(\theta) = n^{-1}\sum_{i=1}^n m(W_i, \theta)$ *for some function* $m(W_i, \theta) \in R^{d_\psi}$ $\forall \theta \in \Theta_\delta$, *for any true parameter* $\gamma^* \in \Gamma$.

(ii) $E_{\gamma^*} m(W_i, \psi^*, \pi) = 0$ $\forall \pi \in \Pi$, $\forall i \geq 1$ *when the true parameter is* γ^* $\forall \gamma^* = (\psi^*, \pi^*, \phi^*) \in \Gamma$ *with* $\beta^* = 0$.[32]

Define an empirical process $\{G_n(\pi) : \pi \in \Pi\}$ by

$$G_n(\pi) = n^{-1/2}\sum_{i=1}^n \big(m(W_i, \psi_{0,n}, \pi) - E_{\gamma_n} m(W_i, \psi_{0,n}, \pi)\big). \tag{3.5}$$

[32]In some time series examples, $D_\psi Q_n(\theta)$ is of the form $n^{-1}\sum_{i=1}^n m_i(\theta)$, where $m_i(\theta)$ depends on $\{W_j : \forall 1 \leq j \leq i\}$. Assumption C2 can be relaxed to cover such cases without any changes to the results of the paper. In such cases, Assumption C3 below still can hold provided $\{m_i(\theta) : i \leq n\}$ satisfies a suitable "asymptotic weak dependence" condition, such as near-epoch dependence.

The recentered and rescaled (generalized) first derivative of $Q_n(\theta)$ w.r.t. ψ is assumed to satisfy an empirical process central limit theorem (CLT):

ASSUMPTION C3: *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, $G_n(\cdot) \Rightarrow G(\cdot; \gamma_0)$, *where* $G(\cdot; \gamma_0)$ *is a mean zero Gaussian process indexed by* $\pi \in \Pi$ *with bounded continuous sample paths and some covariance kernel* $\Omega(\pi_1, \pi_2; \gamma_0)$ *for* $\pi_1, \pi_2 \in \Pi$.

Numerous empirical process results in the literature can be used to verify this assumption, including results in Pollard (1984, 1990), Andrews (1994), and van der Vaart and Wellner (1996).

The (generalized) second derivative of $Q_n(\theta)$ w.r.t. ψ is assumed to satisfy the following assumption.

ASSUMPTION C4: (i) *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, $\sup_{\pi\in\Pi} \|D_{\psi\psi} Q_n(\psi_{0,n}, \pi) - H(\pi; \gamma_0)\| \to_p 0$ *for some nonstochastic symmetric* $d_\psi \times d_\psi$ *matrix-valued function* $H(\pi; \gamma_0)$ *on* $\Pi \times \Gamma$ *that is continuous on* Π $\forall \gamma_0 \in \Gamma$.

(ii) $\lambda_{\min}(H(\pi; \gamma_0)) > 0$ *and* $\lambda_{\max}(H(\pi; \gamma_0)) < \infty$ $\forall \pi \in \Pi$, $\forall \gamma_0 \in \Gamma$ *with* $\beta_0 = 0$.

Define the $d_\psi \times d_\beta$ matrix of partial derivatives of the average population moment function w.r.t. the true β value, β^*, to be

$$K_n(\theta; \gamma^*) = n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \beta^{*\prime}} E_{\gamma^*} m(W_i, \theta). \tag{3.6}$$

The domain of the function $K_n(\theta; \gamma^*)$ is $\Theta_\delta \times \Gamma_0$, where $\Theta_\delta = \{\theta \in \Theta : \|\beta\| < \delta\}$, $\Gamma_0 = \{\gamma_a = (a\beta, \zeta, \pi, \phi) \in \Gamma : \gamma = (\beta, \zeta, \pi, \phi) \in \Gamma \text{ with } \|\beta\| < \delta \text{ and } a \in [0, 1]\}$, and $\delta > 0$ is as in Assumption B2(ii). The set Γ_0 is not empty by Assumptions B2(ii) and (iii).

ASSUMPTION C5: (i) $K_n(\theta; \gamma^*)$ *exists* $\forall (\theta, \gamma^*) \in \Theta_\delta \times \Gamma_0$, $\forall n \geq 1$.

(ii) *For some nonstochastic* $d_\psi \times d_\beta$ *matrix-valued function* $K(\psi_0, \pi; \gamma_0)$, $K_n(\overline{\psi}_n, \pi; \widetilde{\gamma}_n) \to K(\psi_0, \pi; \gamma_0)$ *uniformly over* $\pi \in \Pi$ *for all nonstochastic sequences* $\{\overline{\psi}_n\}$ *and* $\{\widetilde{\gamma}_n\}$ *such that* $\widetilde{\gamma}_n \in \Gamma$, $\widetilde{\gamma}_n \to \gamma_0 = (0, \zeta_0, \pi_0, \phi_0)$ *for some* $\gamma_0 \in \Gamma$, $(\overline{\psi}_n, \pi) \in \Theta$, *and* $\overline{\psi}_n \to \psi_0 = (0, \zeta_0)$.

(iii) $K(\psi_0, \pi; \gamma_0)$ *is continuous on* Π $\forall \gamma_0 \in \Gamma$ *with* $\beta_0 = 0$.

Assumption C5 is not restrictive. A set of primitive sufficient conditions for Assumption C5 is given in Supplemental Appendix A.

For simplicity, $K(\psi_0, \pi; \gamma_0)$ is abbreviated as $K(\pi; \gamma_0)$. Note that $(\overline{\psi}_n, \widetilde{\gamma}_n)$ in Assumption C5(ii) is in $\Theta_\delta \times \Gamma_0$ for n large.

Due to the expansion about $\psi_{0,n}$, rather than about the true value ψ_n, in Assumption C1, a bias is introduced in the first derivative $D_\psi Q_n(\psi_{0,n}, \pi)$: its mean is not zero. In consequence, its behavior differs between category I and

II sequences. With category I sequences, it converges (after suitable normalization) to the sum of the stochastic term $G(\pi)$ and the nonstochastic term $K(\pi;\gamma_0)b$ due to the bias, and the two are of the same order of magnitude. With category II sequences, the true β_n is farther from the point of expansion 0 than with category I sequences and, in consequence, the nonstochastic bias term is of a larger order of magnitude than the stochastic term. In this case, the limit is $K(\pi;\gamma_0)\omega_0$, which is nonstochastic.

Specifically, Assumptions C2, C3, and C5 are used to show the key result

$$
\begin{aligned}
(3.7)\qquad & a_n(\gamma_n)D_\psi Q_n(\psi_{0,n},\pi) \\
&= \big[G_n(\pi)+(K_n(\psi_{0,n},\pi;\gamma_n)+o(1))n^{1/2}\beta_n\big]n^{-1/2}a_n(\gamma_n) \\
&\Rightarrow \begin{cases} G(\pi;\gamma_0)+K(\pi;\gamma_0)b, \\ \quad \text{if } n^{1/2}\beta_n \to b\in R^{d_\beta}, \\ K(\pi;\gamma_0)\omega_0, \\ \quad \text{if } \|n^{1/2}\beta_n\|\to\infty \text{ and } \beta_n/\|\beta_n\|\to\omega_0, \end{cases}
\end{aligned}
$$

where the convergence holds under $\{\gamma_n\}\in\Gamma(\gamma_0,0,b)$.[33]

Next, we introduce the limits of the concentrated criterion function $Q_n^c(\pi)=Q_n(\widehat{\psi}_n(\pi),\pi)$ after suitable normalization. Define a "weighted noncentral chi-square" process $\{\xi(\pi;\gamma_0,b):\pi\in\Pi\}$ and a nonstochastic function $\{\eta(\pi;\gamma_0,\omega_0):\pi\in\Pi\}$ by

$$
\begin{aligned}
(3.8)\qquad \xi(\pi;\gamma_0,b) &= -\frac{1}{2}(G(\pi;\gamma_0)+K(\pi;\gamma_0)b)'H^{-1}(\pi;\gamma_0) \\
&\quad \times (G(\pi;\gamma_0)+K(\pi;\gamma_0)b), \\
\eta(\pi;\gamma_0,\omega_0) &= -\frac{1}{2}\omega_0'K(\pi;\gamma_0)'H^{-1}(\pi;\gamma_0)K(\pi;\gamma_0)\omega_0.
\end{aligned}
$$

Under Assumptions C3, C4, and C5(iii), $\{\xi(\pi;\gamma_0,b):\pi\in\Pi\}$ has bounded continuous sample paths almost surely (a.s.).

Let $Q_{0,n}=Q_n(\psi_{0,n},\pi)$, where $\psi_{0,n}=(0,\zeta_n)$ as in Assumption C1. Note that $Q_{0,n}$ does not depend on π by Assumption A.

LEMMA 3.2: *Suppose Assumptions* A, B1–B3, *and* C1–C5 *hold*. *Under* $\{\gamma_n\}\in\Gamma(\gamma_0,0,b)$, *there are two alternatives*:

(a) *When* $\|b\|<\infty$, $n(Q_n^c(\cdot)-Q_{0,n})\Rightarrow\xi(\cdot;\gamma_0,b)$.

(b) *When* $\|b\|=\infty$ *and* $\beta_n/\|\beta_n\|\to\omega_0$ *for some* $\omega_0\in R^{d_\beta}$ *with* $\|\omega_0\|=1$, $\|\beta_n\|^{-2}(Q_n^c(\pi)-Q_{0,n})\to_p\eta(\pi;\gamma_0,\omega_0)$ *uniformly over* $\pi\in\Pi$.

To obtain the asymptotic distribution of $\widehat{\pi}_n$ when $\beta_n=O(n^{-1/2})$ via the continuous mapping theorem, we use the following assumption.

[33]See Lemma 9.1 in Supplemental Appendix B.

ASSUMPTION C6: *Each sample path of the stochastic process* $\{\xi(\pi;\gamma_0,b):\pi\in\Pi\}$ *in some set* $A(\gamma_0,b)$ *with* $P_{\gamma_0}(A(\gamma_0,b))=1$ *is minimized over* Π *at a unique point* (*which typically depends on the sample path*), *denoted* $\pi^*(\gamma_0,b)$ $\forall\gamma_0\in\Gamma$ *with* $\beta_0=0$ $\forall b$ *with* $\|b\|<\infty$.

In Assumption C6, $\pi^*(\gamma_0,b)$ is random. In Supplemental Appendix A, we provide a primitive sufficient condition for Assumption C6 for the case when β is a scalar, that is, $d_\beta=1$, which covers many cases of interest.

Define the Gaussian process $\{\tau(\pi;\gamma_0,b):\pi\in\Pi\}$ by

$$\tau(\pi;\gamma_0,b)=-H^{-1}(\pi;\gamma_0)(G(\pi;\gamma_0)+K(\pi;\gamma_0)b)-(b,0_{d_\zeta}), \tag{3.9}$$

where $(b,0_{d_\zeta})\in R^{d_\psi}$. Note that, by (3.8) and (3.9), $\xi(\pi;\gamma_0,b)=-(1/2)(\tau(\pi;\gamma_0,b)+(b,0_{d_\zeta}))'H(\pi;\gamma_0)(\tau(\pi;\gamma_0,b)+(b,0_{d_\zeta}))$. As in Assumption C6, $\pi^*(\gamma_0,b)=\arg\min_{\pi\in\Pi}\xi(\pi;\gamma_0,b)$.

The following theorem is one of the main results of this paper. It provides the asymptotic distribution of the estimator $\widehat{\theta}_n$ and the optimized objective function $Q_n(\widehat{\theta}_n)$ for category I sequences.

THEOREM 3.1: *Suppose Assumptions* A, B1–B3, *and* C1–C6 *hold*. *Under* $\{\gamma_n\}\in\Gamma(\gamma_0,0,b)$ *with* $\|b\|<\infty$, *the following results hold*:

(a)

$$\begin{pmatrix} n^{1/2}(\widehat{\psi}_n-\psi_n) \\ \widehat{\pi}_n \end{pmatrix} \to_d \begin{pmatrix} \tau(\pi^*(\gamma_0,b);\gamma_0,b) \\ \pi^*(\gamma_0,b) \end{pmatrix}.$$

(b) $n(Q_n(\widehat{\theta}_n)-Q_{0,n})\to_d \inf_{\pi\in\Pi}\xi(\pi;\gamma_0,b)$.

COMMENTS: (i) Define the Gaussian process $\{\tau_\beta(\pi;\gamma_0,b):\pi\in\Pi\}$ by

$$\tau_\beta(\pi;\gamma_0,b)=S_\beta\tau(\pi;\gamma_0,b)+b, \tag{3.10}$$

where $S_\beta=[I_{d_\beta}:0_{d_\beta\times d_\zeta}]$ is the $d_\beta\times d_\psi$ selector matrix that selects β out of ψ. The asymptotic distribution of $n^{1/2}\widehat{\beta}_n$ (without centering at β_n) under $\Gamma(\gamma_0,0,b)$ with $\|b\|<\infty$ is given by $\tau_\beta(\pi^*(\gamma_0,b);\gamma_0,b)$. This quantity appears in the asymptotic distributions of t statistics below.

(ii) Assumption C6 is not needed for Theorem 3.1(b).

(iii) Using Theorem 3.1, Figure 1 provides the asymptotic and finite-sample densities of the ML estimator of the MA parameter π in the ARMA(1, 1) model when the true π value, π_0, is 0.4. It gives the densities for $b=0$, -2, -4, and -12, where b indexes the magnitude of the difference β between the AR and MA parameters.[34] Specifically, for the finite-sample results, $b=n^{1/2}\beta$,

[34]The asymptotic density in Figure 1 is invariant to the sign of b.

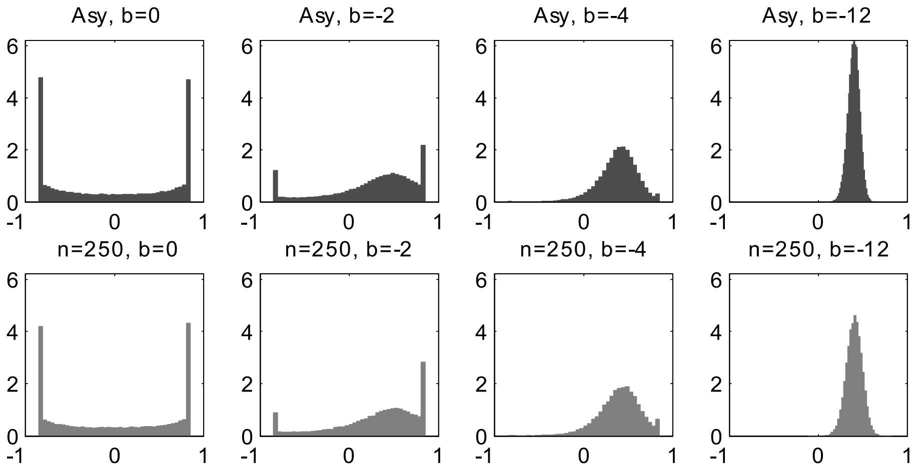

FIGURE 1.—Asymptotic and finite-sample ($n = 250$) densities of the estimator of the MA parameter π in the ARMA(1, 1) model when $\pi_0 = 0.4$.

$n = 250$, and $\varepsilon_t \sim N(0, 1)$. Note that for $n = 250$, the values $b = 0$, -2, -4, and -12 correspond to $\beta = 0.0$, -0.13, -0.25, and -0.76, respectively. For $n = 100$, these b values correspond to $\beta = 0.0$, -0.2, -0.4, and -1.2, respectively. The optimization parameter spaces for the MA and AR parameters are $[-0.85, 0.85]$ and $[-0.90, 0.90]$, respectively. The true parameter spaces are $[-0.80, 0.80]$ and $[-0.85, 0.85]$, respectively.[35] For the asymptotic and finite-sample results 50,000 simulation repetitions are used.

Figure 1 shows that the ML estimator has a distribution that is very far from a normal distribution in the unidentified and weakly identified cases. In these cases, there is a buildup of mass at the boundaries of the optimization space. There also is a bias toward 0. Figure 1 indicates that the asymptotic approximations based on Theorem 3.1 work strikingly well. There are some differences between the asymptotic and finite-sample densities, but they are small.

(iv) Figure 2 provides analogous results to those of Figure 1 for the ML estimator of β, the difference between the AR and MA parameters. Figure 2 shows a very pronounced bimodal distribution in the unidentified case and a side lobe in one weakly identified case. As in Figure 1, the asymptotic approximations are found to work exceptionally well.

3.4. *Intermediate Assumptions and Estimation Results*

Next, we specify an assumption that is used in the proof of consistency of $\widehat{\pi}_n$ in the "less close, local to $\beta = 0$" case in which $\beta_n \to 0$ and $n^{1/2}\|\beta_n\| \to \infty$.

[35]These choices cover a broad range of parameters, but avoid unit root and boundary effects. These parameter spaces satisfy Assumptions B1 and B2.

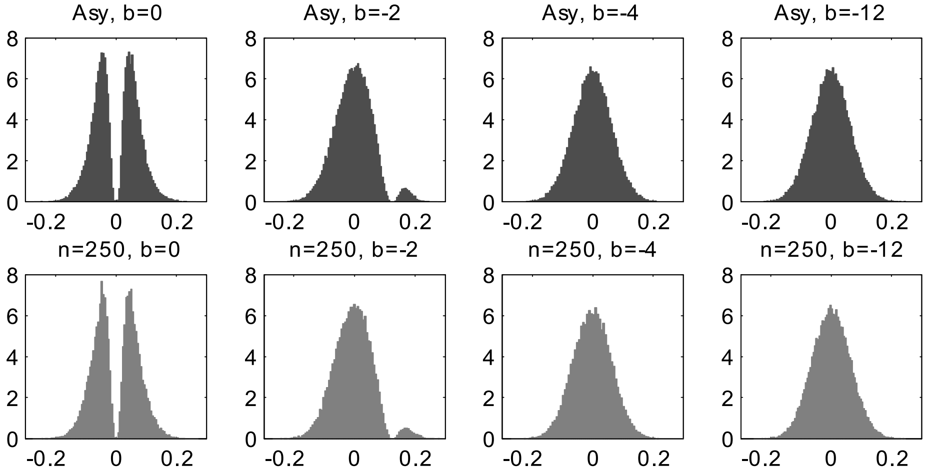

FIGURE 2.—Asymptotic and finite-sample ($n = 250$) densities of the estimator of β (centered at the true value) in the ARMA(1, 1) model when $\pi_0 = 0.4$.

ASSUMPTION C7: *The nonstochastic function* $\eta(\pi; \gamma_0, \omega_0)$ *is uniquely minimized over* $\pi \in \Pi$ *at* π_0 $\forall \gamma_0 \in \Gamma$ *with* $\beta_0 = 0$.

In Assumption C7, the minimizing value π_0 is nonrandom. In some examples, such as the ARMA(1, 1) example, Assumption C7 can be verified directly. In other examples, Assumption C7 can be verified using the Cauchy–Schwarz inequality or a matrix version of it (see Tripathi (1999)) when $K(\pi; \gamma_0)$ and $H(\pi; \gamma_0)$ take proper forms. For example, see the verification of Assumption C7 for the nonlinear regression example in Supplemental Appendix E and the verification of Assumption C7 for GMM estimators in AC3.

Lemma 9.3 in Supplemental Appendix B shows that when $\pi = \pi_0$, $K(\pi; \gamma_0) = -H(\pi; \gamma_0)S_\beta'$, where $S_\beta = [I_{d_\beta} : 0] \in R^{d_\beta \times d_\psi}$, whereas this relationship does not hold for $\pi \neq \pi_0$ in general.

LEMMA 3.3: *Suppose Assumptions* A, B1–B3, C1–C5, *and* C7 *hold*. *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, (a) $\widehat{\pi}_n - \pi_n \to_p 0$ *and* (b) $\widehat{\psi}_n - \psi_n \to_p 0$.

The following assumption is used when obtaining a key rate of convergence result for $\widehat{\psi}_n$ for sequences $\{\gamma_n\}$ for which $\beta_n \to 0$ and $n^{1/2}\|\beta_n\| \to \infty$.

ASSUMPTION C8: *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, $\frac{\partial}{\partial \psi'} E_{\gamma_n} D_\psi Q_n(\psi, \pi_n)|_{\psi = \psi_n} \to H(\pi_0; \gamma_0)$.

By Assumption C4(i), $H(\pi; \gamma_0)$ is the probability limit of $D_{\psi\psi} Q_n(\psi_{0,n}, \pi_n)$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$. When $Q_n(\theta)$ is a twice differentiable sample average,

$D_\psi Q_n(\theta)$ and $D_{\psi\psi} Q_n(\theta)$ are its first- and second-order partial derivatives w.r.t. ψ, respectively. One can switch E and ∂ under certain regularity conditions, so that $(\partial/\partial\psi')E_{\gamma_n} D_\psi Q_n(\psi_n, \pi_n)$ is the expectation of $D_{\psi\psi} Q_n(\psi_n, \pi_n)$ in this case. Hence, Assumption C8 can be verified by a uniform law of large numbers (LLN) and the continuity of $D_{\psi\psi} Q_n(\psi, \pi)$ in ψ.[36]

LEMMA 3.4: *Suppose Assumptions* A, B1–B3, C1–C5, C7, *and* C8 *hold*. *Then*, $\|\beta_n\|^{-1} \times (\widehat{\psi}_n - \psi_n) = o_p(1)$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ *with* $\beta_0 = 0$.

COMMENT: Lemma 3.4 is a key result because it allows one to apply the quadratic expansion in Assumption D1 below, which only holds in a rapidly shrinking neighborhood of the true value for category II sequences $\{\gamma_n\}$.

3.5. *Distant From* $\beta = 0$ *Assumptions and Estimation Results*

Assumptions D1–D3 below are used to derive asymptotic distributions under sequences of true parameters $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$. The "D" denotes that the sequences of true parameters considered are more *distant* from the point of nonidentification than are the sequences in the C assumptions.

We define a matrix $B(\beta)$ that is used to normalize the (generalized) second-derivative matrix $D^2 Q_n(\theta_n)$ of $Q_n(\theta_n)$ (which is introduced in Assumption D1 below) so that it is nonsingular asymptotically, as specified in Assumption D2. Let

$$B(\beta) = \begin{bmatrix} I_{d_\psi} & 0_{d_\psi \times d_\pi} \\ 0_{d_\pi \times d_\psi} & \iota(\beta) I_{d_\pi} \end{bmatrix} \in R^{d_\theta \times d_\theta}, \quad \text{where} \tag{3.11}$$

$$\iota(\beta) = \begin{cases} \beta, & \text{if } \beta \text{ is a scalar,} \\ \|\beta\|, & \text{if } \beta \text{ is a vector.} \end{cases}$$

We use a different definition of $B(\beta)$ in the scalar and vector β cases because in the scalar case the use of β, rather than $\|\beta\|$, produces noticeably simpler (but equivalent) formulae, but in the vector case $\|\beta\|$ is required.

ASSUMPTION D1: *When the true parameters are* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, *the following statements hold*:

(i) *The sample criterion function* $Q_n(\theta)$ *has a quadratic expansion in* θ *around* θ_n:

$$\begin{aligned} Q_n(\theta) = {} & Q_n(\theta_n) + DQ_n(\theta_n)'(\theta - \theta_n) \\ & + \frac{1}{2}(\theta - \theta_n) D^2 Q_n(\theta_n)(\theta - \theta_n) + R_n^*(\theta), \end{aligned}$$

[36]When $Q_n(\theta)$ is nonsmooth, one can show that $E_{\gamma_n} D_\psi Q_n(\theta)$ is close to the first-order partial derivative of $Q(\theta; \gamma_0)$ w.r.t. ψ, roughly by switching E_{γ_n} and D_ψ under some regularity conditions, and $D_{\psi\psi} Q_n(\theta)$ is typically taken to be the second-order partial derivative of $Q(\theta; \gamma_0)$ w.r.t. ψ in this case.

where $DQ_n(\theta_n) \in R^{d_\theta}$ *is a stochastic generalized first derivative vector and* $D^2Q_n(\theta_n) \in R^{d_\theta \times d_\theta}$ *is a generalized second derivative matrix that is symmetric and may be stochastic or nonstochastic*.

(ii) *The remainder*, $R_n^*(\theta)$, *satisfies*

$$\sup_{\theta \in \Theta_n(\delta_n)} \frac{|nR_n^*(\theta)|}{(1 + \|n^{1/2}B(\beta_n)(\theta - \theta_n)\|)^2} = o_p(1)$$

for all constants $\delta_n \to 0$, *where* $\Theta_n(\delta_n) = \{\theta \in \Theta : \|\psi - \psi_n\| \leq \delta_n \|\beta_n\|$ *and* $\|\pi - \pi_n\| \leq \delta_n\}$.

The quadratic approximation in Assumption D1 only holds for θ in a neighborhood $\Theta_n(\delta_n)$ of θ_n whose radius shrinks as the sample size gets larger. In particular, the distance between ψ and ψ_n shrinks faster than $\|\beta_n\|$ when $\beta_n \to 0$. It is for this reason that the rate of convergence result in Lemma 3.4 is a key result.[37]

The sufficient conditions for Assumption C1 referenced in the previous subsection also are sufficient for Assumption D1. The quantities $DQ_n(\theta_n)$ and $D^2Q_n(\theta_n)$ take similar forms to $D_\psi Q_n(\psi_{0,n}, \pi)$ and $D_{\psi\psi} Q_n(\psi_{0,n}, \pi)$ (see the discussion following Assumption C1), but involve derivatives w.r.t. θ, not ψ, and hence are not functions of π.

The next assumption requires good behavior of the (generalized) second derivative of $Q_n(\theta_n)$ after it has been rescaled to eliminate its singularity when $\beta_n \to 0$.

ASSUMPTION D2: *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $J_n = B^{-1}(\beta_n) D^2 Q_n(\theta_n) \times B^{-1}(\beta_n) \to_p J(\gamma_0) \in R^{d_\theta \times d_\theta}$, *where* $J(\gamma_0)$ *is nonsingular and symmetric*.

The next assumption requires the rescaled (generalized) first derivative to satisfy a CLT.

ASSUMPTION D3: (i) *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $n^{1/2} B^{-1}(\beta_n) DQ_n(\theta_n) \to_d G^*(\gamma_0) \sim N(0_{d_\theta}, V(\gamma_0))$ *for some symmetric* $d_\theta \times d_\theta$ *matrix* $V(\gamma_0)$.[38]

(ii) $V(\gamma_0)$ *is positive definite* $\forall \gamma_0 \in \Gamma$.

The following theorem is a key result. It provides the asymptotic distribution of the estimator $\widehat{\theta}_n$ and the optimized objective function $Q_n(\widehat{\theta}_n)$ for category II and III sequences.

[37]The quadratic approximation requires $\theta \in \Theta_n(\delta_n)$ because for such $\theta = (\beta, \zeta, \pi)$, one has $\|\beta\|/\|\beta_n\| = 1 + o(1)$ and, hence, the rescaling that enters the Hessian is asymptotically equivalent whether it is based on β or the true value β_n.

[38]In the vector β case, $J(\gamma_0)$ and $V(\gamma_0)$ may depend on ω_0 as well as γ_0.

THEOREM 3.2: *Suppose Assumptions* A, B1–B3, C1–C5, C7, C8, *and* D1–D3 *hold*. *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, *the following results hold*:

(a) $n^{1/2}B(\beta_n)(\widehat{\theta}_n - \theta_n) \to_d -J^{-1}(\gamma_0)G^*(\gamma_0) \sim N(0_{d_\theta}, J^{-1}(\gamma_0)V(\gamma_0)J^{-1}(\gamma_0))$.

(b) $n(Q_n(\widehat{\theta}_n) - Q_n(\theta_n)) \to_d -\frac{1}{2}G^*(\gamma_0)'J^{-1}(\gamma_0)G^*(\gamma_0)$.

In sum, the asymptotic results of this paper for $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{\zeta}_n, \widehat{\pi}_n)$ are as follows: The estimator $\widehat{\psi}_n = (\widehat{\beta}_n, \widehat{\zeta}_n)$ is $n^{1/2}$ consistent for all categories of sequences $\{\beta_n\}$ in Table I. The estimator $\widehat{\pi}_n$ is inconsistent for category I sequences and consistent for categories II and III. The asymptotic distribution of $n^{1/2}(\widehat{\psi}_n - \psi_n)$ $(= n^{1/2}((\widehat{\beta}_n, \widehat{\zeta}_n) - (\beta_n, \zeta_n)))$ is a functional of a Gaussian process with a mean that is (typically) nonzero for category I sequences (due to the inconsistency of $\widehat{\pi}_n$) and is normal with mean zero for categories II and III. The asymptotic distribution of $\widehat{\pi}_n$ is a functional of the same Gaussian process for category I sequences. These estimation results permit the calculation of the asymptotic biases of $(\widehat{\beta}_n, \widehat{\zeta}_n, \widehat{\pi}_n)$ for category I sequences as a function of the strength of identification. The asymptotic distribution of $n^{1/2}\|\beta_n\|(\widehat{\pi}_n - \pi_n)$ is normal with mean zero for category II sequences. The asymptotic distribution of $n^{1/2}(\widehat{\pi}_n - \pi_n)$ is normal with mean zero for category III sequences.

4. *t* AND QLR CONFIDENCE SETS AND TESTS

In this section, we determine the asymptotic size of standard CS's for a function $r(\theta)$ $(\in R^{d_r})$ of θ obtained by inverting t and QLR tests of the hypothesis $H_0 : r(\theta) = v$ for $v \in r(\Theta)$. We also consider standard t and QLR tests of H_0. In Section 5 below, we introduce robust CS's whose asymptotic size is guaranteed to equal their nominal size. For brevity, results for Wald CS's for vector-valued functions $r(\theta)$ are given in AC3.

4.1. *t Statistics*

The t statistic is defined as follows. Let

$$\Sigma(\gamma_0) = J^{-1}(\gamma_0)V(\gamma_0)J^{-1}(\gamma_0) \quad \text{and} \quad \widehat{\Sigma}_n = \widehat{J}_n^{-1}\widehat{V}_n\widehat{J}_n^{-1}, \tag{4.1}$$

where $\widehat{J}_n$ and $\widehat{V}_n$ are estimators of $J(\gamma_0)$ and $V(\gamma_0)$ that do not depend on the nuisance parameter ϕ.

The t statistic is defined when $r(\theta)$ is real-valued, that is, $d_r = 1$. It takes the form

$$T_n(v) = \frac{n^{1/2}(r(\widehat{\theta}_n) - v)}{(r_\theta(\widehat{\theta}_n)B^{-1}(\widehat{\beta}_n)\widehat{\Sigma}_n B^{-1}(\widehat{\beta}_n)r_\theta(\widehat{\theta}_n)')^{1/2}}, \tag{4.2}$$

where $r_\theta(\theta) = (\partial/\partial\theta')r(\theta) = [r_\psi(\theta) : r_\pi(\theta)] \in R^{d_r \times d_\theta}$, $r_\psi(\theta) = (\partial/\partial\psi')r(\theta) \in R^{d_r \times d_\psi}$, and $r_\pi(\theta) = (\partial/\partial\pi')r(\theta) \in R^{d_r \times d_\pi}$.

Although this definition of the t statistic involves $B^{-1}(\widehat{\beta}_n)$, it is the same as the standard definition used in practice. By Theorem 3.2(a), when $\beta_0 \neq 0$, $B^{-1}(\beta_0)\Sigma(\gamma_0)B^{-1}(\beta_0)$ is the asymptotic covariance matrix of $\widehat{\theta}_n$. In the t statistic, the asymptotic covariance is replaced by the estimator $B^{-1}(\widehat{\beta}_n)\widehat{\Sigma}_n B^{-1}(\widehat{\beta}_n)$. The same form of the t statistic is used under all sequences of true parameters $\{\gamma_n\} \in \Gamma(\gamma_0)$.

In the results below, we consider the behavior of the t statistic when the null hypothesis holds. Thus, under a sequence $\{\gamma_n\}$, we consider the sequence of null hypotheses $H_0 : r(\theta) = v_n$, where v_n equals $r(\theta_n)$ and $\gamma_n = (\theta_n, \phi_n)$. We employ the notational simplification

$$(4.3) \qquad T_n = T_n(v_n), \quad \text{where} \quad v_n = r(\theta_n).$$

The function of interest, $r(\theta)$, satisfies the following assumption.

ASSUMPTION R: (i) $r(\theta) \in R$ *is continuously differentiable on* Θ.
(ii) $r_\theta(\theta) \neq 0_{d_\theta}$ $\forall \theta \in \Theta$.
(iii) $\text{rank}(r_\pi(\theta)) = d_\pi^*$ *for some constant* $d_\pi^* \leq \min(d_r, d_\pi)$ $\forall \theta \in \Theta_\delta = \{\theta \in \Theta : \|\beta\| < \delta\}$ *for some* $\delta > 0$.

A sufficient condition for Assumption R is $r(\theta) = R_1'\theta$, where $R_1 \in R^{d_\theta}$ and $R_1 \neq 0$.

4.2. *Variance Matrix Estimators*

The estimators of the components of the asymptotic variance matrix are assumed to satisfy the following Assumptions V1 and V2. Two forms of Assumption V1 are needed: one for scalar β and one for vector β.[39] For brevity, we only state the scalar β version here; the vector β version is given in Supplemental Appendix A.

When β is a scalar, let $J(\theta; \gamma_0)$ and $V(\theta; \gamma_0)$ for $\theta \in \Theta$ be some nonstochastic $d_\theta \times d_\theta$ matrix-valued functions such that $J(\theta_0; \gamma_0) = J(\gamma_0)$ and $V(\theta_0; \gamma_0) = V(\gamma_0)$, where $J(\gamma_0)$ and $V(\gamma_0)$ are as in Assumptions D2 and D3. Let

$$(4.4) \qquad \Sigma(\theta; \gamma_0) = J^{-1}(\theta; \gamma_0)V(\theta; \gamma_0)J^{-1}(\theta; \gamma_0) \quad \text{and}$$
$$\Sigma(\pi; \gamma_0) = \Sigma(\psi_0, \pi; \gamma_0).$$

Let $\Sigma_{\beta\beta}(\pi; \gamma_0)$ denote the upper left $(1, 1)$ element of $\Sigma(\pi; \gamma_0)$.

Assumption V1 below applies when β is a scalar.

[39]The reason for the difference is that the normalizing matrix $B(\beta)$ is different in these two cases.

ASSUMPTION V1—Scalar β: (i) *$\widehat{J}_n = \widehat{J}_n(\widehat{\theta}_n)$ and $\widehat{V}_n = \widehat{V}_n(\widehat{\theta}_n)$ for some* (*stochastic*) *$d_\theta \times d_\theta$ matrix-valued functions $\widehat{J}_n(\theta)$ and $\widehat{V}_n(\theta)$ on Θ that satisfy* $\sup_{\theta\in\Theta}\|\widehat{J}_n(\theta) - J(\theta;\gamma_0)\| \to_p 0$ *and* $\sup_{\theta\in\Theta}\|\widehat{V}_n(\theta) - V(\theta;\gamma_0)\| \to_p 0$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $\|b\| < \infty$.

(ii) *$J(\theta;\gamma_0)$ and $V(\theta;\gamma_0)$ are continuous in θ on Θ $\forall\gamma_0 \in \Gamma$ with $\beta_0 = 0$.*

(iii) $\lambda_{\min}(\Sigma(\pi;\gamma_0)) > 0$ *and* $\lambda_{\max}(\Sigma(\pi;\gamma_0)) < \infty$ $\forall\pi \in \Pi$, $\forall\gamma_0 \in \Gamma$ *with* $\beta_0 = 0$.

The following assumption applies with both scalar and vector β.

ASSUMPTION V2: *Under $\Gamma(0,\infty,\omega_0)$, $\widehat{J}_n \to_p J(\gamma_0)$ and $\widehat{V}_n \to_p V(\gamma_0)$.*

4.3. *Asymptotic Distribution of the t Statistic*

Next, we provide the asymptotic distribution of the t statistic under H_0. Define

$$(4.5)\qquad T_\psi(\pi;\gamma_0,b) = \frac{r_\psi(\pi)\tau(\pi;\gamma_0,b)}{(r_\psi(\pi)\overline{\Sigma}_{\psi\psi}(\pi;\gamma_0,b)r_\psi(\pi)')^{1/2}},$$

where $r_\psi(\pi) = r_\psi(\psi_0,\pi) \in R^{1\times d_\psi}$, $\tau(\pi;\gamma_0,b) \in R^{d_\psi}$, $\overline{\Sigma}_{\psi\psi}(\pi;\gamma_0,b)$ is the upper left $d_\psi \times d_\psi$ block of $\overline{\Sigma}(\pi;\gamma_0,b)$, $\overline{\Sigma}(\pi;\gamma_0,b) = \Sigma(\pi;\gamma_0)$ in the scalar β case (and is defined differently in the vector β case; see (8.2) in Supplemental Appendix A), $\Sigma(\pi;\gamma_0)$ is defined in (4.4), and $\tau_\beta(\pi;\gamma_0,b)$ is defined in (3.10). Also, define

$$(4.6)\qquad T_\pi(\pi;\gamma_0,b) = \frac{\|\tau_\beta(\pi;\gamma_0,b)\|(r(\psi_0,\pi) - r(\psi_0,\pi_0))}{(r_\pi(\pi)\overline{\Sigma}_{\pi\pi}(\pi;\gamma_0,b)r_\pi(\pi)')^{1/2}},$$

where $\overline{\Sigma}_{\pi\pi}(\pi;\gamma_0,b)$ is the lower right $d_\pi \times d_\pi$ block of $\overline{\Sigma}(\pi;\gamma_0,b)$ and $r_\pi(\pi) = r_\pi(\psi_0,\pi)$.

The following theorem provides the asymptotic null distribution of the t statistic for a scalar restriction. (The null holds by the definition $T_n = T_n(v_n)$ in (4.3).)

THEOREM 4.1: *Suppose Assumptions* A, B1–B3, C1–C8, D1–D3, R, V1, *and* V2 *hold and* $d_r = 1$.

(a) *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $\|b\| < \infty$ *and* $d_\pi^* = 0$, $T_n \to_d T_\psi(\pi^*(\gamma_0,b);\gamma_0,b)$.

(b) *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $\|b\| < \infty$ *and* $d_\pi^* = 1$, $T_n \to_d T_\pi(\pi^*(\gamma_0,b);\gamma_0,b)$.

(c) *Under* $\{\gamma_n\} \in \Gamma(\gamma_0,\infty,\omega_0)$, $T_n \to_d N(0,1)$.

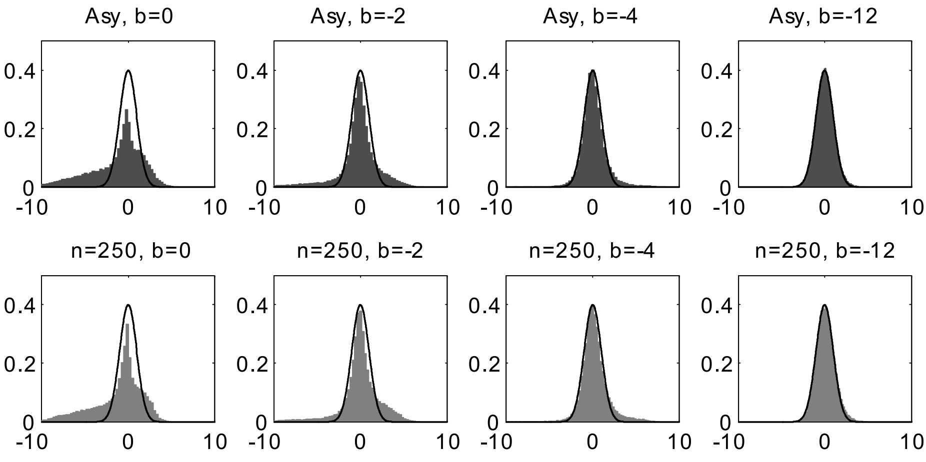

FIGURE 3.—Asymptotic and finite-sample ($n = 250$) densities of the t statistic for the MA parameter π in the ARMA$(1, 1)$ model when $\pi_0 = 0.4$ and with standard normal density (black line).

COMMENTS: (i) When $d^*_\pi = 0$, the scalar restriction only involves ψ by Assumption R(iii). When $d^*_\pi = 1$, the restriction involves π and possibly ψ. However, the randomness in $\widehat{\psi}_n$ is dominated by that in $\widehat{\pi}_n$ under the conditions of Theorem 4.1(b) because $\widehat{\psi}_n$ is consistent but $\widehat{\pi}_n$ is not. In consequence, the asymptotic distribution in Theorem 4.1(b) is as if the restriction is only on π.

(ii) Using Theorem 4.1, Figure 3 provides the asymptotic and finite-sample ($n = 250$) densities of the t statistic for tests concerning the MA parameter π in the ARMA$(1, 1)$ model for $\pi_0 = 0.4$ and $b = 0$, -2, -4, and -12. The black line in Figure 3 is the standard normal density, which is the strong-identification asymptotic density of the t statistic. Figure 3 shows that the t statistic has a noticeably nonnormal shape due to skewness and kurtosis for small $|b|$, although it is much less nonnormal than the distribution of the corresponding estimator.[40]

(iii) Figure 4(a) provides graphs of the 0.95 asymptotic quantiles of the $|t|$ statistic for π as a function of $|b|$.[41] For small to medium $|b|$ values, the graphs exceed the 0.95 quantile under strong identification (given by the horizontal black line). This implies that $|t|$ tests and CI's that employ the stan-

[40]The distributions of the estimator of π and the t statistic for π are not the same up to a scale shift, even asymptotically. This occurs because the variance estimator that appears in the t statistic involves an estimator of π, that is not consistent when $|b| < \infty$; it is random even in the limit.

[41]The asymptotic quantiles are invariant to the sign of b, but the finite-sample quantiles are not.

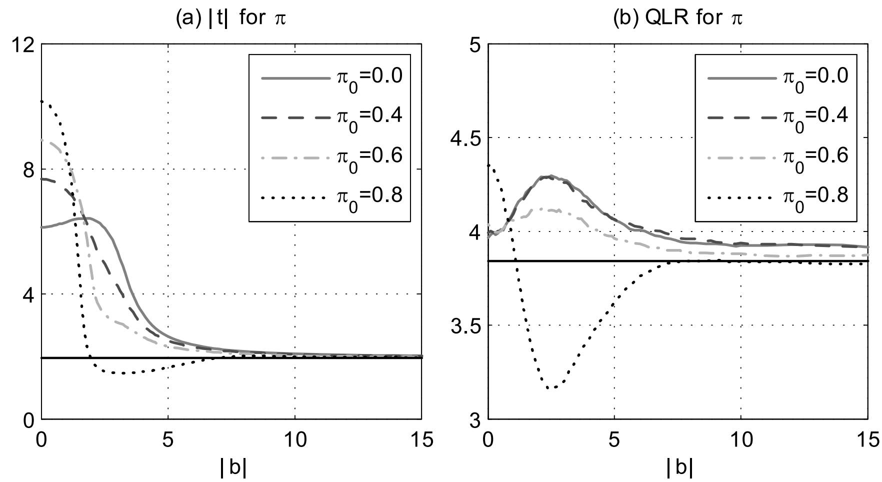

FIGURE 4.—Asymptotic 0.95 quantiles of the $|t|$ and QLR statistics for tests concerning the MA parameter π in the ARMA$(1, 1)$ model.

dard critical value (based on the normal distribution) have incorrect asymptotic size. The exceedance is very large. For example, for $\pi_0 = 0.8$ and $b = 0$, the quantile is roughly 10, whereas for strong identification ($|b| = \infty$), it is roughly 2.

(iv) The results of Theorem 4.1 are used below to obtain the asymptotic size of standard and robust t CI's. But first, we provide analogous results for the QLR statistic.

4.4. *QLR Statistics*

Here, we consider the quasi-likelihood ratio (QLR) statistic. In this subsection, $d_r \geq 1$. The function $r(\theta)$ is assumed to be smooth and to be of the form

$$(4.7) \qquad r(\theta) = \begin{bmatrix} r_1(\psi) \\ r_2(\pi) \end{bmatrix},$$

where $r_1(\psi) \in R^{d_{r_1}}$, $d_{r_1} \geq 0$ is the number of restrictions on ψ, $r_2(\pi) \in R^{d_{r_2}}$, $d_{r_2} \geq 0$ is the number of restrictions on π, and $d_r = d_{r_1} + d_{r_2}$.

Given the form in (4.7), our results for the QLR statistic do not cover the case where a single restriction depends on both ψ and π. This can be restrictive. However, in some cases, it is possible to obtain results for restrictions of this type by a simple reparametrization; see Comment (iii) to Theorem 4.2 below.

For $v \in r(\Theta)$, we define a restricted estimator $\widetilde{\theta}_n(v)$ of θ subject to the restriction that $r(\theta) = v$. By definition,

$$(4.8) \qquad \widetilde{\theta}_n(v) \in \Theta, \quad r(\widetilde{\theta}_n(v)) = v, \quad \text{and} \\ Q_n(\widetilde{\theta}_n(v)) = \inf_{\theta \in \Theta : r(\theta) = v} Q_n(\theta) + o(n^{-1}).$$

For testing $H_0 : r(\theta) = v$, the QLR test statistic is

$$(4.9) \qquad \mathrm{QLR}_n(v) = 2n\big(Q_n(\widetilde{\theta}_n(v)) - Q_n(\widehat{\theta}_n)\big)/\widehat{s}_n,$$

where $\widehat{s}_n$ is a real-valued scaling factor that is employed in some cases to yield a QLR statistic that has an asymptotic $\chi^2_{d_r}$ null distribution under strong identification; see Assumptions RQ2 and RQ3 below.

4.5. *QLR Assumptions*

If $r(\theta)$ includes restrictions on π (i.e., $d_{r_2} > 0$), then not all values $\pi \in \Pi$ are consistent with the restriction $r_2(\pi) = v_2$. For $v_2 \in r_2(\Theta)$, the set of π values that are consistent with $r_2(\pi) = v_2$ is denoted by

$$(4.10) \qquad \Pi_r(v_2) = \{\pi \in \Pi : r_2(\pi) = v_2 \text{ for some } \theta = (\psi, \pi) \in \Theta\}.$$

If $d_{r_2} = 0$, then by definition $\Pi_r(v_2) = \Pi \; \forall v_2 \in r_2(\Theta)$.

We assume that $r(\theta)$ satisfies the following assumption.

ASSUMPTION RQ1: (i) *$r(\theta)$ is continuously differentiable on Θ.*
(ii) *$r_\theta(\theta)$ is full row rank d_r $\forall \theta \in \Theta$.*
(iii) *$r(\theta)$ satisfies* (4.7).
(iv) *$d_H(\Pi_r(v_2), \Pi_r(v_{0,2})) \to 0$ as $v_2 \to v_{0,2}$ $\forall v_{0,2} \in r_2(\Theta^*)$.*
(v) *$Q(\psi, \pi; \gamma_0)$ is continuous in ψ at ψ_0 uniformly over $\pi \in \Pi$ (i.e., $\sup_{\pi \in \Pi} |Q(\psi, \pi; \gamma_0) - Q(\psi_0, \pi; \gamma_0)| \to 0$ as $\psi \to \psi_0$) $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$.*
(vi) *$Q(\theta; \gamma_0)$ is continuous in θ at θ_0 $\forall \gamma_0 \in \Gamma$ with $\beta_0 \neq 0$.*

In Assumption RQ1(iv), d_H denotes the Hausdorff distance. Assumption RQ1(i) and (ii) are standard. Assumption RQ1(iv) is easy to verify in most cases. Assumption RQ1(v) and (vi) are not restrictive.

Even under strong identification, it is known that the QLR statistic has an asymptotic $\chi^2_{d_r}$ null distribution only under additional assumptions to those used for Wald and Lagrange multiplier (LM) statistics.[42] The following assumptions correspond to these additional conditions.

[42]The reason is that the weight matrices of the Wald and LM statistics can be designed specifically to achieve an asymptotic $\chi^2_{d_r}$ null distribution, whereas with the QLR statistic, no weight matrix appears and at most one has a real-valued scaling factor $\widehat{s}_n$ with which to make adjustments.

ASSUMPTION RQ2: *Either* (i) $V(\gamma_0) = s(\gamma_0)J(\gamma_0)$ *for some nonrandom scalar constant* $s(\gamma_0)$ $\forall \gamma_0 \in \Gamma$, *or* (ii) $V(\gamma_0)$ *and* $J(\gamma_0)$ *are block diagonal* (*possibly after reordering their rows and columns*), *the restrictions* $r(\theta)$ *only involve parameters that correspond to one block of* $V(\gamma_0)$ *and* $J(\gamma_0)$—*call them* $V_{11}(\gamma_0)$ *and* $J_{11}(\gamma_0)$—*and for this block*, $V_{11}(\gamma_0) = s(\gamma_0)J_{11}(\gamma_0)$ *for some nonrandom scalar constant* $s(\gamma_0)$ $\forall \gamma_0 \in \Gamma$.

ASSUMPTION RQ3: *The scalar statistic* $\widehat{s}_n$ *satisfies* $\widehat{s}_n \to_p s(\gamma_0)$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *and under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$.

For example, Assumptions RQ2(i) and RQ3 hold with $s(\gamma_0) = \widehat{s}_n = 1$ for a correctly specified log-likelihood criterion function, a GMM criterion function with asymptotically optimal weight matrix, and an empirical likelihood criterion function. For a homoskedastic nonlinear regression model, Assumptions RQ2(i) and RQ3 hold with $s(\gamma_0)$ equal to the error variance σ^2 and $\widehat{s}_n$ equal to a consistent estimator of σ^2, such as the sample variance based on the residuals.

4.6. *Asymptotic Distribution of the QLR Statistic*

Now we determine the asymptotic distribution of the QLR statistic under the sequences $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ and $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ when the null hypotheses are true, that is, when $v = v_n = r(\theta_n)$ for $\gamma_n = (\theta_n, \phi_n)$ $\forall n \geq 1$. These results are needed to obtain asymptotic size results for QLR-based CS's. The results for the QLR statistic rely on results for the restricted estimator $\widetilde{\theta}_n(v_n)$. These results are complicated by the fact that not all values $\pi \in \Pi$ are necessarily consistent with the restrictions $r((\psi_n, \pi)) = v_n$. For brevity, results for the restricted estimators are stated in Supplemental Appendix B.

We use the notational simplifications

$$\text{QLR}_n = \text{QLR}_n(v_n) \quad \text{and} \quad \widetilde{\theta}_n = \widetilde{\theta}_n(v_n), \quad \text{where} \quad v_n = r(\theta_n) \quad \text{and} \quad \gamma_n = (\theta_n, \phi_n). \tag{4.11}$$

The matrix $r_\theta(\theta)$ of partial derivatives of $r(\theta)$ can be written as

$$r_\theta(\theta) = \frac{\partial}{\partial \theta'} r(\theta) = \begin{bmatrix} r_{1,\psi}(\psi) & 0_{d_{r_1} \times d_\pi} \\ 0_{d_{r_2} \times d_\psi} & r_{2,\pi}(\pi) \end{bmatrix}, \tag{4.12}$$

where $r_{1,\psi}(\psi) = (\partial/\partial\psi')r_1(\psi) \in R^{d_{r_1} \times d_\psi}$ and $r_{2,\pi}(\pi) = (\partial/\partial\pi')r_2(\pi) \in R^{d_{r_2} \times d_\pi}$.

For notational simplicity, let $\Pi_{r,0} = \Pi_r(v_{0,2})$, where $v_{0,2} = r_2(\pi_0)$ and $\gamma_0 = (\theta_0, \phi_0) \in \Gamma$. That is, $\Pi_{r,0}$ is the set of values π that are compatible with the restrictions on π when γ_0 is the true parameter value.

Next, we introduce the limit under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$ of the restricted concentrated criterion function after suitable normalization. For

$\pi \in \Pi$, define

$$(4.13)\quad \xi_r(\pi;\gamma_0,b) = \xi(\pi;\gamma_0,b) + \frac{1}{2}\tau(\pi;\gamma_0,b)' P_\psi(\pi;\gamma_0)' \times H(\pi;\gamma_0) P_\psi(\pi;\gamma_0)\tau(\pi;\gamma_0,b), \quad \text{where}$$
$$P_\psi(\pi;\gamma_0) = H^{-1}(\pi;\gamma_0) r_{1,\psi}(\psi_0)'(r_{1,\psi}(\psi_0) \times H^{-1}(\pi;\gamma_0) r_{1,\psi}(\psi_0)')^{-1} r_{1,\psi}(\psi_0)$$

and $\tau(\pi;\gamma_0,b)$ is defined in (3.9). The $d_\psi \times d_\psi$ matrix $P_\psi(\pi;\gamma_0)$ is an oblique projection matrix that projects onto the space spanned by the rows of $r_{1,\psi}(\psi_0)$.

The following result gives the asymptotic distribution of the QLR statistic under sequences $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$.

THEOREM 4.2: *Suppose Assumptions* A, B1–B3, C1–C5, RQ1, *and* RQ3 *hold*. *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $\|b\| < \infty$, $\mathrm{QLR}_n \to_d 2(\inf_{\pi \in \Pi_{r,0}} \xi_r(\pi;\gamma_0, b) - \inf_{\pi\in\Pi} \xi(\pi;\gamma_0,b))/s(\gamma_0)$.

COMMENTS: (i) Using Theorem 4.2, Figure 5 provides the asymptotic and finite-sample ($n = 250$) densities of the QLR statistic for tests concerning the MA parameter π in the ARMA(1,1) model for $\pi_0 = 0.4$ and $b = 0, -2, -4$, and -12. The black line in Figure 5 is the χ^2_1 density, which is the strong-

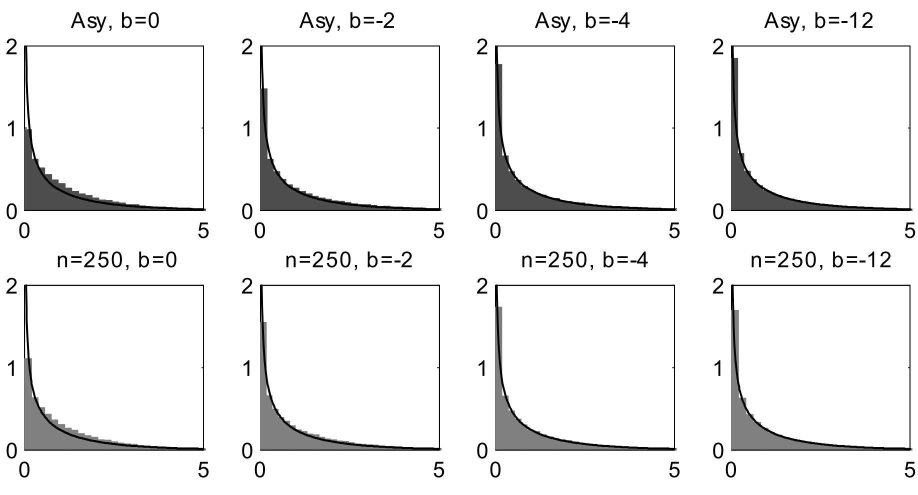

FIGURE 5.—Asymptotic and finite-sample ($n = 250$) densities of the QLR statistic for the MA parameter π in the ARMA(1,1) model when $\pi_0 = 0.4$ and with χ^2_1 density (black line).

identification asymptotic density of the QLR statistic. Figure 5 indicates that the QLR statistic is well approximated by a χ_1^2 distribution, even under weak identification. This suggests that the QLR statistic yields tests and CI's that are substantially less sensitive to weak identification than t-based tests and CI's are.

(ii) Figure 4(b) provides graphs of the 0.95 asymptotic quantiles of the QLR statistic for π as a function of $|b|$. For small to medium $|b|$ values, the graphs exceed the 0.95 quantile under strong identification (given by the horizontal black line). Thus, tests and CI's based on the standard critical values (from the χ_1^2 distribution) have incorrect asymptotic size. For the QLR statistic the exceedance is much smaller than for the $|t|$ statistic. For the QLR statistic, for $\pi_0 = 0.8$ and $b = 0$, the quantile is roughly 4.4, whereas for strong identification it is roughly 3.8.

(iii) The proof of Theorem 4.2 requires an extension of the arg max theorem (e.g., see Lemma 3.2.1 of van der Vaart and Wellner (1996, p. 286)) to the case where the maximum is taken over a sample-size dependent sequence of sets.[43] See Lemma 9.10 in Supplemental Appendix B. This lemma may be of use in other contexts.

(iv) Assumption RQ1(iii) rules out the case where any single restriction depends on both ψ and π, but, in some cases, a reparametrization can be used to obtain results for such restrictions. Suppose $d_\pi = d_\beta$. Consider restrictions of the form $r(\theta) = (r_1(\psi), \pi + \beta)$. In this case, the asymptotic distribution of the QLR statistic in Theorems 4.2 and 4.3 (below) is the same as its distribution when $r(\theta) = (r_1(\psi), \pi)$. We use this result in the ARMA(1, 1) example to obtain CI's for the AR parameter, which equals $\pi + \beta$.[44]

(v) The proof of Theorem 4.2 can be altered easily to yield some results for the QLR test under sequences of alternative hypothesis distributions, which yield asymptotic power results for QLR-based tests. Suppose the restrictions $r(\theta)$ depend only on π, that is, $d_{r_1} = 0$ and $r(\theta) = r_2(\pi)$. The sequence of true values of $r_2(\pi)$ satisfies $r_2(\pi_n) \to r_2(\pi_0) = v_{0,2}$ as $n \to \infty$. Now suppose the null hypothesis value of $r_2(\pi)$ is $v_{0,2}^{\text{null}}$, where $v_{0,2}^{\text{null}} \neq v_{0,2}$. Then the asymptotic distribution of QLR_n for this null hypothesis under the alternative hypothesis distributions $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ is given by the expression in Theorem 4.2, but with $\Pi_{r,0} = \Pi_r(v_{0,2})$ replaced by $\Pi_r(v_{0,2}^{\text{null}})$. This covers both local and fixed alternatives.

(vi) The proof of Theorem 4.2 makes use of the approach of Chernoff (1954).

[43]The arg max/min theorem provides the asymptotic distribution of a maximizer/minimizer of a stochastic process that converges weakly to some limit process.

[44]See Section 9.4.4 of Supplemental Appendix B for more details.

Next, we give results for the QLR statistic under sequences $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$. Define

$$\text{(4.14)} \quad \lambda_{\text{QLR}}(\gamma_0) = G^*(\gamma_0)' J^{-1}(\gamma_0) P_\theta(\gamma_0)' J(\gamma_0) \times P_\theta(\gamma_0) J^{-1}(\gamma_0) G^*(\gamma_0), \quad \text{where}$$
$$P_\theta(\gamma_0) = J^{-1}(\gamma_0) r_\theta(\theta_0)' (r_\theta(\theta_0) J^{-1}(\gamma_0) r_\theta(\theta_0)')^{-1} r_\theta(\theta_0)$$

and $J(\gamma_0)$ and $G^*(\gamma_0)$ are defined in Assumptions D2 and D3. The matrix $P_\theta(\gamma_0)$ is an oblique projection matrix that projects onto the space spanned by the rows of $r_\theta(\theta_0)$.

THEOREM 4.3: *Suppose Assumptions* A, B1–B3, C1–C5, C7, C8, D1–D3, RQ1, *and* RQ3 *hold*. *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $\text{QLR}_n \to_d \lambda_{\text{QLR}}(\gamma_0)/s(\gamma_0)$.

COMMENT: When Assumption RQ2 holds, by Theorem 4.3 and some calculations, under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,

$$\text{(4.15)} \quad \text{QLR}_n \to_d \lambda_{\text{QLR}}(\gamma_0)/s(\gamma_0) \sim \chi^2_{d_r}.$$

4.7. *Asymptotic Size of Standard t and QLR Confidence Sets*

Now, we establish the asymptotic size of standard CS's obtained by inverting t and QLR statistics using Lemma 2.1 and Theorems 4.1–4.3. The standard nominal $1-\alpha$ symmetric two-sided t, upper one-sided t, lower one-sided t, and QLR CS's take the form in (2.4) with $\mathcal{T}_n(v) = |T_n(v)|$, $T_n(v)$, $-T_n(v)$, and $\text{QLR}_n(v)$, respectively, and $c_{n,1-\alpha}(v) = z_{1-\alpha/2}$, $z_{1-\alpha}$, $z_{1-\alpha}$, and $\chi^2_{d_r,1-\alpha}$, where $T_n(v)$ is defined in (4.2), $\text{QLR}_n(v)$ is defined in (4.9), $z_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal distribution, and $\chi^2_{d_r,1-\alpha}$ is the $1-\alpha$ quantile of the $\chi^2_{d_r}$ distribution.

For $h = (b, \gamma_0)$ with $\|b\| < \infty$ and H as in (2.8), define

$$\text{(4.16)} \quad T(h) = \begin{cases} T_\psi(\pi^*(\gamma_0, b); \gamma_0, b), & \text{if } d^*_\pi = 0, \\ T_\pi(\pi^*(\gamma_0, b); \gamma_0, b), & \text{if } d^*_\pi = 1, \end{cases}$$
$$\text{QLR}(h) = 2\Big(\inf_{\pi \in \Pi_{r,0}} \xi_r(\pi; \gamma_0, b) - \inf_{\pi \in \Pi} \xi(\pi; \gamma_0, b)\Big)\Big/ s(\gamma_0).$$

As defined, $T(h)$ is the asymptotic distribution of T_n under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ for $\|b\| < \infty$ given in Theorem 4.1(a) or (b) depending on the rank of $r_\pi(\theta)$, which is denoted by d^*_π. Only one of the cases applies for any particular parameter of interest $r(\theta)$ and it is known which applies. Here, $\text{QLR}(h)$ is the asymptotic distribution of QLR_n under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ for $\|b\| < \infty$ given in Theorem 4.2.

Let $c_{|t|,1-\alpha}(h)$, $c_{t,1-\alpha}(h)$, $c_{-t,1-\alpha}(h)$, and $c_{\text{QLR},1-\alpha}(h)$ denote the $1-\alpha$ quantiles of $|T(h)|$, $T(h)$, $-T(h)$, and $\text{QLR}(h)$, respectively, for $h \in H$.

As in (2.6), AsySz denotes the asymptotic size of a CS of nominal level $1-\alpha$. The asymptotic size results for the t and QLR CS's use the following distribution function (d.f.) continuity assumptions, which typically are not restrictive.

ASSUMPTION V3: *The d.f. of $T(h)$ is continuous at $z_{\alpha/2}$ and $z_{1-\alpha/2}$, z_α, and $z_{1-\alpha}$ $\forall h \in H$ in the two-sided, upper one-sided, and lower-sided cases, respectively.*

ASSUMPTION RQ4: *The d.f. of* QLR(h) *is continuous at* (i) $\chi^2_{d_r,1-\alpha}$ *and* (ii) $\sup_{h\in H} c_{\mathrm{QLR},1-\alpha}(h)$.

THEOREM 4.4: (a) *Suppose Assumptions* A, B1–B3, C1–C8, D1–D3, R, *and* V1–V3 *hold and* $d_r = 1$. *The standard nominal* $1-\alpha$ *symmetric two-sided, upper one-sided, and lower one-sided t CI's have* AsySz $= \min\{\inf_{h\in H} P(|T(h)| \le z_{1-\alpha/2}), 1-\alpha\}$, $\min\{\inf_{h\in H} P(T(h) \le z_{1-\alpha}), 1-\alpha\}$, *and* $\min\{\inf_{h\in H} P(-T(h) \le z_{1-\alpha}), 1-\alpha\}$, *respectively.*

(b) *Suppose Assumptions* A, B1–B3, C1–C5, C7, C8, D1–D3, RQ1–RQ3, *and* RQ4(i) *hold. Then the standard nominal* $1-\alpha$ *QLR CS has* AsySz $= \min\{\inf_{h\in H} P(\mathrm{QLR}(h) \le \chi^2_{d_r,1-\alpha}), 1-\alpha\}$.

COMMENTS: (i) Depending on the distributions of $\{T(h) : h \in H\}$ and $\{\mathrm{QLR}(h) : h \in H\}$, the t and QLR CS's have asymptotic sizes equal to or less than $1-\alpha$.

(ii) Figure 6 reports asymptotic and finite-sample coverage probabilities (CP's) of nominal 95% standard $|t|$ and QLR CI's (which employ normal and χ^2_1 critical values, respectively) for the MA parameter π in the ARMA(1,1)

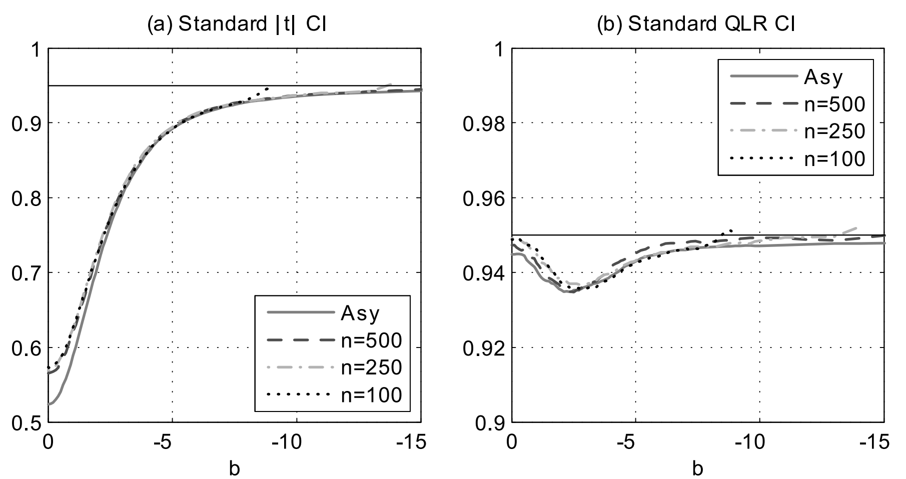

FIGURE 6.—Coverage probabilities of standard $|t|$ and QLR CI's for the MA parameter π in the ARMA(1,1) model when $\pi_0 = 0$.

model. The CP's are given as a function of b (≤ 0) for true $\pi_0 = 0.0$, for $n = 100$, 250, 500, and ∞ (i.e., asymptotic).[45] The CP's of the $|t|$ CI are very low for $|b|$ values less than 10. For $b = 0$, the asymptotic and finite-sample CP's are all below 0.60. Hence, the size of this nominal 95% CI is less than 0.60 asymptotically and in finite samples.[46] Figure 6 shows that the undercoverage of the standard QLR CI for π is much less severe than for the $|t|$ CI. The asymptotic size of the standard QLR CI is approximately 0.935, which is not much less than the nominal asymptotic size of 0.950. Note that the asymptotic CP's in Figure 6 provide a very good approximation to the finite-sample CP's.

In sum, the asymptotic results for tests and CS's vary over the three categories in Table I. For category I sequences, standard tests and CS's have asymptotic rejection/coverage probabilities that may differ, sometimes substantially, from their nominal level. In consequence, the asymptotic size of standard tests and CS's often is substantially different from the desired nominal size. For category II and III sequences, standard tests and CS's have the desired asymptotic rejection/coverage probability properties. For hypotheses or CS's that involve π, their power/noncoverage properties are standard for category II and III sequences.

5. ROBUST CONFIDENCE SETS

In this section, we construct robust CS's for $r(\theta)$ that have correct asymptotic size. A robust CS is obtained by inverting a test statistic, denoted here generically by $\mathcal{T}_n$, using a robust critical value that differs from a standard strong-identification critical value (such as a normal or $\chi^2_{d_r}$ quantile). The robust critical value can be data dependent. The test statistic $\mathcal{T}_n$ can be the t statistic defined in (4.3), the absolute value of the t statistic, the QLR statistic defined in (4.11), the Wald statistic analyzed in AC3, or some other statistic.

A robust critical value takes into account the fact that the test statistic, $\mathcal{T}_n$, has a nonstandard asymptotic distribution under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$. As a result, a larger critical value often is required under weak identification (i.e., $\|b\| < \infty$) than under semi-strong or strong identification (i.e., $\|b\| = \infty$).

[45]In Figures 6 and 7 below, the graphs for $n = 100$ are not given for all values of b because b is restricted by the parameter space. The same is true for the graphs for $n = 250$ in Figures 6 and 7(a) and (b). See Supplemental Appendix D for details. These parameter space restrictions are responsible for the wiggles that occur in some of the $n = 100$ and 250 graphs in Figures 6 and 7 near the right end of the graphs.

[46]More specifically, the asymptotic sizes of the nominal 95% standard $|t|$ and QLR CI's for π are computed to be 0.523 and 0.933, respectively. These results also apply to CI's for the AR parameter ρ. This is based on a grid of π_0 values with grid size 0.05 for $|\pi_0| \leq 0.60$ and grid size 0.025 for $0.625 \leq |\pi_0| \leq 0.825$.

A simple robust critical value is the "least favorable" (LF) critical value that is large enough for all identification categories. This yields a CS with correct asymptotic size, but one that typically is overly long and is not as informative as desirable when the model is strongly identified.

In consequence, we introduce data-dependent critical values that improve upon the LF critical value by using an identification-category-selection (ICS) procedure in the construction of the critical value. Two methods are considered: type 1 and type 2. The first is relatively simple. The second has preferable statistical properties, but is more intensive computationally.

We also introduce versions of these robust critical values that (i) impose the known null hypothesis value and (ii) plug in consistent estimators of consistently estimable nuisance parameters in the formulae for the robust critical values. We recommend employing combined null-imposed/plug-in versions of the robust critical values whenever possible because they yield the smallest critical values and still deliver asymptotically correct size. However, they may be more burdensome computationally than other versions of the robust critical values.

5.1. *Least Favorable Critical Values*

Let $\mathcal{T}(h)$ denote the asymptotic distribution of $\mathcal{T}_n$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, where $h = (b, \gamma_0) \in H$, and h and H are defined in (2.8). Let $c_{\mathcal{T},1-\alpha}(h)$ denote the $1-\alpha$ quantile of $\mathcal{T}(h)$ for $h \in H$. For example, when $\mathcal{T}_n$ is the two-sided t statistic $|T_n|$ of Section 4, then $\mathcal{T}(h)$ and $c_{\mathcal{T},1-\alpha}(h)$ equal $|T(h)|$ and $c_{|t|,1-\alpha}(h)$, respectively.

Under semi-strong and strong identification (i.e., $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$), $\mathcal{T}_n$ is assumed to have a standard asymptotic distribution, such as the standard normal or chi-squared distribution, as is typically the case. Let $c_{\mathcal{T},1-\alpha}(\infty)$ denote the $1-\alpha$ quantile of this distribution.

The LF critical value is

$$c^{\mathrm{LF}}_{\mathcal{T},1-\alpha} = \max\Big\{\sup_{h \in H} c_{\mathcal{T},1-\alpha}(h), c_{\mathcal{T},1-\alpha}(\infty)\Big\}. \tag{5.1}$$

The LF critical value can be improved (i.e., made smaller) by exploiting the knowledge of the null hypothesis value of $r(\theta)$. For example, if the null hypothesis specifies the value of π to be 3, then the supremum in (5.1) does not need to be taken over all $h \in H$, only over the h values for which $\pi = 3$. We call such a critical value a null-imposed (NI) LF critical value. Using a NI-LF critical value increases the computational burden because a different critical value is employed for each null hypothesis value.

To be precise, let

$$H(v) = \{h = (b, \gamma_0) \in H : \|b\| < \infty, r(\theta_0) = v\}, \tag{5.2}$$

where $\gamma_0 = (\theta_0, \phi_0)$. By definition, $H(v)$ is the subset of H that is consistent with the null hypothesis $H_0 : r(\theta_0) = v$, where θ_0 denotes the true value. The NI-LF critical value, denoted $c^{\text{NI-LF}}_{\mathcal{T},1-\alpha}(v)$, is defined by replacing H with $H(v)$ in (5.1) when the null hypothesis value is $r(\theta_0) = v$. Note that v takes values in the set $V_r = \{v_0 : r(\theta_0) = v_0 \text{ for some } h = (b, \gamma_0) \in H\}$.[47]

When part of γ is unknown under H_0 but can be consistently estimated, then a *plug-in* LF (or plug-in NI-LF) critical value can be used that has the correct size asymptotically and is smaller than the LF (or NI-LF) critical value. The plug-in critical value replaces elements of γ with consistent estimators in the formulae in (5.1), and the supremum over H (or $H(v)$) is reduced to a supremum over the resulting subset of H, denoted $\widehat{H}_n$, for which the consistent estimators appear in each vector γ. For example, if ζ is consistently estimated by $\widehat{\zeta}_n$, then H is replaced by $\widehat{H}_n = \{h = (b, \gamma) \in H : \gamma = (\beta, \widehat{\zeta}_n, \pi, \phi)\}$ or $H(v)$ is replaced by $H(v) \cap \widehat{H}_n$. Note that the parameter b is not consistently estimable, so it cannot be replaced by a consistent estimator.

5.2. *Data-Dependent Robust Critical Values: Type 1*

Here we improve on the LF critical value by employing an identification-category-selection (ICS) procedure that uses the data to determine whether b is finite. If b is deemed to be finite (i.e., π is weakly identified (or unidentified)), then the LF critical value is used; otherwise, the standard asymptotic critical value is used. This ICS critical value is closely related to a method suggested in Andrews (1999, Sec. 6.4, 2000, Sec. 4) for boundary problems, and to the generalized moment selection critical value method used in Andrews and Soares (2010) and some other papers for inference in partially identified models based on moment inequalities. It also is related to, but quite distinct from, the approach in Forchini and Hillier (2003).[48]

The ICS procedure chooses between the identification categories $\mathcal{IC}_0 : \|b\| < \infty$ and $\mathcal{IC}_1 : \|b\| = \infty$. The (unrestricted) statistic used for identification-category selection is

$$A_n = (n\widehat{\beta}_n' \widehat{\Sigma}^{-1}_{\beta\beta,n} \widehat{\beta}_n / d_\beta)^{1/2}, \tag{5.3}$$

where $\widehat{\Sigma}_{\beta\beta,n}$ is the upper left $d_\beta \times d_\beta$ block of $\widehat{\Sigma}_n$ and $\widehat{\Sigma}_n$ is the estimator of the covariance matrix defined in (4.1). We use A_n to assess the strength of identification.

[47] When $r(\theta) = \beta$ and the null hypothesis imposes that $\beta = v$, the parameter b can be imposed to equal $n^{1/2}v$. In this case, $H(v) = H_n(v) = \{h = (b, \gamma_0) \in H : b = n^{1/2}v\}$. The asymptotic size results given below for NI-LF CI's and robust CI's with NI critical values hold in this case.

[48] Forchini and Hillier (2003) advocate carrying out inference conditional on a test statistic that measures the strength of identification. They focus on estimation. Here we consider tests and inference that is unconditional.

Alternatively, one can use a null-imposed ICS (NI-ICS) statistic. For the restriction $r(\theta_n) = v_n$, the NI-ICS statistic is $A_n(v_n) = (n\widetilde{\beta}_n'\widetilde{\Sigma}_{\beta\beta,n}^{-1}\widetilde{\beta}_n/d_\beta)^{1/2}$, where $\widetilde{\beta}_n$ is the restricted estimator of β (subject to $r(\theta) = v_n$) and $\widetilde{\Sigma}_{\beta\beta,n}$ is an estimator of its asymptotic variance. Specifically, we take $\widetilde{\Sigma}_{\beta\beta,n}$ to be the upper left $d_\beta \times d_\beta$ block of $\widetilde{\Sigma}_n$, where $\widetilde{\Sigma}_n = \widetilde{P}_n^{\perp}\widetilde{J}_n^{-1}\widetilde{V}_n\widetilde{J}_n^{-1}\widetilde{P}_n^{\perp\prime}$, $\widetilde{J}_n = \widehat{J}_n(\widetilde{\theta}_n)$, $\widetilde{V}_n = \widehat{V}_n(\widetilde{\theta}_n)$, $\widetilde{P}_n^{\perp} = I_{d_\theta} - \widetilde{P}_n$, $\widetilde{P}_n = \widetilde{J}_n^{-1}r_\theta(\widetilde{\theta}_n)'(r_\theta(\widetilde{\theta}_n)\widetilde{J}_n^{-1}r_\theta(\widetilde{\theta}_n)')^{-1}r_\theta(\widetilde{\theta}_n)$, and $\widehat{J}_n(\theta)$ and $\widehat{V}_n(\theta)$ are as in Assumption V1. This form for $\widetilde{\Sigma}_{\beta\beta,n}$ is based on the asymptotic results for the restricted estimator $\widetilde{\theta}_n$ given in Supplemental Appendix B. The NI-ICS statistic has better ICS properties under the null hypothesis than the unrestricted ICS statistic because it exploits the restrictions, but it is misspecified under the alternative. Hence, the preference for one ICS statistic over the other may depend on the model of interest.

Let $\{\kappa_n : n \geq 1\}$ be a sequence of constants, that is, tuning parameters, that diverges to infinity as $n \to \infty$. One selects $\mathcal{IC}_0$ if $A_n \leq \kappa_n$ and one selects $\mathcal{IC}_1$ otherwise. Under $\mathcal{IC}_0$, A_n is $O_p(1)$. Hence, one consistently selects $\mathcal{IC}_0$ provided κ_n diverges to infinity. We make the following assumption:

ASSUMPTION K: (i) $\kappa_n \to \infty$ *and* (ii) $\kappa_n/n^{1/2} \to 0$.

For example, $\kappa_n = (\ln n)^{1/2}$, which is analogous to the Bayesian information criterion (BIC) penalty term, satisfies Assumption K.

Using the ICS procedure described above, the type 1 robust CS with nominal level $1-\alpha$ is obtained by inverting a test based on $\mathcal{T}_n$ with critical value $\widetilde{c}_{\mathcal{T},1-\alpha,n}$ defined by

$$(5.4)\qquad \widetilde{c}_{\mathcal{T},1-\alpha,n} = \begin{cases} c^{\mathrm{LF}}_{\mathcal{T},1-\alpha}, & \text{if } A_n \leq \kappa_n, \\ c_{\mathcal{T},1-\alpha}(\infty), & \text{if } A_n > \kappa_n. \end{cases}$$

The type 1 robust critical value $\widetilde{c}_{\mathcal{T},1-\alpha,n}$ can be improved by employing NI and/or plug-in versions of it. They are defined by replacing H with $H(v)$, $\widehat{H}_n$, or $H(v) \cap \widehat{H}_n$, as in Section 5.1. The type 1 NI robust critical value is denoted $\widetilde{c}_{\mathcal{T},1-\alpha,n}(v)$ for $v \in V_r$.

5.3. *Data-Dependent Robust Critical Values: Type 2*

Next, we consider a type 2 robust critical value that does not require the tuning parameter κ_n to diverge to infinity as $n \to \infty$. In consequence, asymptotic size-correction factors Δ_1 and Δ_2 can be introduced. These size-correction factors are designed to improve the asymptotic approximations. The type 2 robust critical value also provides a continuous transition from a weak-identification critical value to a strong-identification critical value using a transition function

$s(x)$. This robust critical value is akin to the method employed in Andrews and Barwick (2012) for moment inequality models.

Let $s(x)$ be a continuous function on $[0,\infty)$ that satisfies (i) $0 \le s(x) \le 1$, (ii) $s(x)$ is nonincreasing in x, (iii) $s(0) = 1$, and (iv) $s(x) \to 0$ as $x \to \infty$. Examples of transition functions include $s(x) = \exp(-c \cdot x)$ for some $c > 0$ and $s(x) = (1 + c \cdot x)^{-1}$ for some $c > 0$.[49,50] In the ARMA example, we use the function $s(x) = \exp(-x/2)$.

The type 2 robust critical value is

$$\widehat{c}_{\mathcal{T},1-\alpha,n} = \begin{cases} c_B, & \text{if } A_n \le \kappa, \\ c_S + [c_B - c_S] \cdot s(A_n - \kappa), & \text{if } A_n > \kappa, \end{cases} \quad \text{where} \tag{5.5}$$

$$c_B = c^{\mathrm{LF}}_{\mathcal{T},1-\alpha} + \Delta_1, \quad c_S = c_{\mathcal{T},1-\alpha}(\infty) + \Delta_2,$$

and $\Delta_1 \ge 0$ and $\Delta_2 \ge 0$ are defined below. Here, "B" denotes big and "S" denotes small. When $A_n \le \kappa$, $\widehat{c}_{\mathcal{T},1-\alpha,n}$ equals the LF critical value $c^{\mathrm{LF}}_{\mathcal{T},1-\alpha}$ plus a size-correction factor Δ_1. When $A_n > \kappa$, $\widehat{c}_{\mathcal{T},1-\alpha,n}$ is a convex combination of $c^{\mathrm{LF}}_{\mathcal{T},1-\alpha} + \Delta_1$ and $c_{\mathcal{T},1-\alpha}(\infty) + \Delta_2$, where Δ_2 is another size-correction factor and the weight given to the standard critical value $c_{\mathcal{T},1-\alpha}(\infty)$ increases with the strength of identification, as measured by $A_n - \kappa$.

The unrestricted ICS statistic A_n satisfies $A_n \to_d A(h)$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$, where

$$A(h) = (\tau_\beta(\pi^*; \gamma_0, b)' \Sigma^{-1}_{\beta\beta}(\pi^*; \gamma_0) \tau_\beta(\pi^*; \gamma_0, b)/d_\beta)^{1/2}, \tag{5.6}$$

where π^* abbreviates $\pi^*(\gamma_0, b)$, and $\tau_\beta(\pi; \gamma_0, b)$ and $\Sigma_{\beta\beta}(\pi; \gamma_0)$ are defined in (3.10) and (4.4), respectively.[51–53]

For any Δ_1 and Δ_2, under $\gamma_n \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$, the asymptotic null rejection probability (NRP) of a test based on the statistic $\mathcal{T}_n$ and the robust

[49]The asymptotic size results given in Theorem 5.1 below also hold for the abrupt transition function $s(x) = 1 - 1(x > 0)$, which is discontinuous at $x = 0$, provided one adds the assumption that $P(A(h) = \kappa) = 0 \ \forall h \in H$, where $A(h)$ is defined in (5.6) below. The latter condition is satisfied in most examples.

[50]If $c^{\mathrm{LF}}_{\mathcal{T},1-\alpha} = \infty$, one should take $s(x)$ to equal 0 for x sufficiently large and define $\infty \times 0$ in (5.5) to equal 0. Then the critical value $\widehat{c}_{\mathcal{T},1-\alpha,n}$ is infinite if A_n is small and is finite if A_n is sufficiently large.

[51]The convergence in distribution follows from Theorem 3.1(a) and Assumption V1.

[52]In the vector β case, $\Sigma_{\beta\beta}(\pi; \gamma_0)$ is replaced with $\Sigma_{\beta\beta}(\pi, \omega^*(\pi; \gamma_0, b); \gamma_0)$ in (5.6), where $\Sigma_{\beta\beta}(\pi, \omega; \gamma_0)$ is defined in (8.1) and $\omega^*(\pi; \gamma_0, b)$ is defined in (8.2) in Supplemental Appendix A. When the type 2 robust critical value is considered in the vector β case, h is defined to include $\omega_0 \in R^{d_\beta}$ with $\|\omega_0\| = 1$ as an element, that is, $h = (b, \gamma_0, \omega_0)$ and $H = \{h = (b, \gamma_0, \omega_0) : \|b\| < \infty, \gamma_0 \in \Gamma \text{ with } \beta_0 = 0, \|\omega_0\| = 1\}$.

[53]Analogously, the NI-ICS statistic $A_n(v_n)$ satisfies $A_n(v_n) \to_d A(h, v_0)$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$, where, for brevity, $A(h, v_0)$ is defined in Comment (iii) to Theorem 9.1 in Supplemental Appendix B. When the NI-ICS statistic is employed, $A(h)$ is replaced with $A(h, v_0)$ in all formulae that follow.

critical value $\widehat{c}_{\mathcal{T},1-\alpha,n}$ is

$$\begin{aligned}(5.7)\quad & \mathrm{NRP}(\Delta_1, \Delta_2; h) \\ & = P(\mathcal{T}(h) > c_B \,\&\, A(h) \le \kappa) + P(\mathcal{T}(h) > c_A(h) \,\&\, A(h) > \kappa) \\ & = P(\mathcal{T}(h) > c_B) + P\big(\mathcal{T}(h) \in (c_A(h), c_B] \,\&\, A(h) > \kappa\big), \quad \text{where} \\ & c_A(h) = c_S + (c_B - c_S) \cdot s(A(h) - \kappa).\end{aligned}$$

The constants Δ_1 and Δ_2 are chosen such that $\mathrm{NRP}(\Delta_1, \Delta_2; h) \le \alpha \ \forall h \in H$. In particular, we define

$$\begin{aligned}(5.8)\quad & \Delta_1 = \sup_{h \in H_1} \Delta_1(h), \quad \text{where} \\ & \Delta_1(h) \ge 0 \quad \text{solves} \quad \mathrm{NRP}(\Delta_1(h), 0; h) = \alpha \quad \text{or} \\ & \Delta_1(h) = 0 \quad \text{if} \quad \mathrm{NRP}(0, 0; h) < \alpha, \\ & H_1 = \{(b, \gamma_0) : (b, \gamma_0) \in H \,\&\, \|b\| \le \|b_{\max}\| + D\}, \\ & \Delta_2 = \sup_{h \in H} \Delta_2(h), \quad \text{where} \\ & \Delta_2(h) \quad \text{solves} \quad \mathrm{NRP}(\Delta_1, \Delta_2(h); h) = \alpha \quad \text{or} \\ & \Delta_2(h) = 0 \quad \text{if} \quad \mathrm{NRP}(\Delta_1, 0; h) < \alpha.\end{aligned}$$

By definition $b_{\max}$ is such that $c_{\mathcal{T},1-\alpha}(h)$ is maximized over $h \in H$ at $h_{\max} = (b_{\max}, \gamma_{\max}) \in H$ for some $\gamma_{\max} \in \Gamma$ and D is a nonnegative constant, such as 1.[54,55] As defined, Δ_1 and Δ_2 can be computed sequentially, which is computationally convenient.

The adjustment via Δ_1 size-corrects for b values that are at or near $b_{\max}$. Size correction is needed here because the ICS statistic A_n is larger than κ with a positive probability asymptotically, even under sequences of true parameters for which the LF critical value is needed to achieve correct asymptotic size.

The adjustment via Δ_2 size-corrects for relatively large values of b. Size correction may be needed here to handle the difference between the ideal critical value for the given value of b and the robust critical value that is determined by

[54]When $\mathrm{NRP}(0, 0; h) > \alpha$, a unique solution $\Delta_1(h)$ typically exists because $\mathrm{NRP}(\Delta_1, 0; h)$ is always nonincreasing in Δ_1 and is typically strictly decreasing and continuous in Δ_1. If no exact solution to $\mathrm{NRP}(\Delta_1(h), 0; h) = \alpha$ exists, then $\Delta_1(h)$ is taken to be any value for which $\mathrm{NRP}(\Delta_1(h), 0; h) \le \alpha$ and $\Delta_1(h) \ge 0$ is as small as possible. Analogous comments apply to the equation $\mathrm{NRP}(\Delta_1, \Delta_2(h); h) = \alpha$ and the definition of $\Delta_2(h)$.

[55]When the LF critical value is achieved at $\|b\| = \infty$, that is, $c_{\mathcal{T},1-\alpha}(\infty) \ge \sup_{h \in H} c_{\mathcal{T},1-\alpha}(h)$, the standard asymptotic critical value $c_{\mathcal{T},1-\alpha}(\infty)$ yields a test or CI with correct asymptotic size, and constants Δ_1 and Δ_2 are not needed. Hence, here we consider the case where $\|b_{\max}\| < \infty$. If $\sup_{h \in H} c_{\mathcal{T},1-\alpha}(h)$ is not attained at any point $h_{\max}$, then $b_{\max}$ can be taken to be any point such that $c_{\mathcal{T},1-\alpha}(h_{\max})$ is arbitrarily close to $\sup_{h \in H} c_{\mathcal{T},1-\alpha}(h)$ for some $h_{\max} = (b_{\max}, \gamma_{\max}) \in H$.

the transition function $s(A_n - \kappa)$. Typically, this discrepancy is small and only a small adjustment Δ_2 is needed.

Given the definitions of Δ_1 and Δ_2, the rejection probability is close to the nominal level α when h is close to $h_{\max}$ (due to the adjustment with Δ_1) and when $\|b\|$ is large (due to the adjustment with Δ_2).

The type 2 robust critical value can be improved by employing NI and/or plug-in versions of it, denoted by $\widehat{c}_{\mathcal{T},1-\alpha,n}(v)$, as in Section 5.1; see Supplemental Appendix A for details.

For any given value of κ, the type 2 robust CS has correct asymptotic size due to the choice of Δ_1 and Δ_2. In consequence, we choose κ based on the false coverage probabilities (FCP's) of the robust CS. When $d_r = 1$, an FCP of a CI for $r(\theta)$ is the probability that the CI includes a value different from the true value $r(\theta)$. Small FCP's are closely linked to short CI's; see Pratt (1961).

The method we use to choose κ is to minimize the average asymptotic FCP of the robust CS at a chosen set of points.[56] We are interested in a robust CS for $r(\theta)$. Let $\mathcal{K}$ denote the set of κ values from which we select. First, for given $h \in H$, we choose a null value $v_{H_0}(h)$ that differs from the true value $v_0 = r(\theta_0)$ (where $h = (b, \gamma_0)$ and $\gamma_0 = (\theta_0, \phi_0)$). The null value $v_{H_0}(h)$ is selected such that the robust CS, based on a reasonable choice of κ such as $\kappa = 1.5$ or 2, has a FCP that is in a range of interest such as close to 0.50.[57,58] Second, we compute the FCP of the value $v_{H_0}(h)$ for each robust CS with $\kappa \in \mathcal{K}$. Third, we repeat the first two steps for each $h \in \mathcal{H}$, where $\mathcal{H}$ is a representative subset of H.[59] The optimal choice of κ is the value that minimizes over $\mathcal{K}$ the average FCP at $v_{H_0}(h)$ over $h \in \mathcal{H}$.

5.4. *Asymptotic Size of Robust t and QLR CS's*

In this section, we show that the LF and data-dependent robust CS's defined above have correct asymptotic size when $\mathcal{T}_n$ equals the t statistic, the absolute value of the t statistic, or the QLR statistic. Analogous results for robust Wald CS's are given in AC3.

[56]For t and Wald CS's, asymptotic FCP's follow from the results in this paper and AC3. For QLR CI's, asymptotic FCP's follow from the results of this paper only for restrictions involving π or $\pi + \beta$; see Comments (iv) and (v) to Theorem 4.2. For other restrictions, one can use a large finite-sample size when determining κ.

[57]For reasonable choices, the value of κ used to obtain $v_{H_0}(h)$ typically has very little effect on the final comparison across different values of κ. For example, this is true in the ARMA(1, 1) example considered below.

[58]When b is close to 0, the FCP may be larger than 0.50 for all admissible v due to weak identification. In such cases, $v_{H_0}(h)$ is taken to be the admissible value that minimizes the FCP for the selected value of κ that is being used to obtain $v_{H_0}(h)$.

[59]When $r(\theta) = \pi$ or $r(\theta) = \pi + \beta$, we do not include h values in $\mathcal{H}$ for which $b = 0$ because when $b = 0$, there is no information about π and it is not necessarily desirable to have a small FCP.

The asymptotic size results of this section rely on the following d.f. continuity conditions, which are not restrictive in most examples.

ASSUMPTION LF: (i) *The d.f. of* $\mathcal{T}(h)$ *is continuous at* $c_{\mathcal{T},1-\alpha}(h)$ $\forall h \in H$.
(ii) *If* $c^{LF}_{\mathcal{T},1-\alpha} > c_{\mathcal{T},1-\alpha}(\infty)$, $c^{LF}_{\mathcal{T},1-\alpha}$ *is attained at some* $h_{\max} \in H$.

ASSUMPTION NI-LF: (i) *The d.f. of* $\mathcal{T}(h)$ *is continuous at* $c_{\mathcal{T},1-\alpha}(h)$ $\forall h \in H(v)$, $\forall v \in V_r$.
(ii) *For some* $v \in V_r$, $c^{LF}_{\mathcal{T},1-\alpha}(v) = c_{\mathcal{T},1-\alpha}(\infty)$ *or* $c^{LF}_{\mathcal{T},1-\alpha}(v)$ *is attained at some* $h_{\max} \in H$.

For $h \in H$, define $\widehat{c}_{\mathcal{T},1-\alpha}(h)$ as $\widehat{c}_{\mathcal{T},1-\alpha,n}$ is defined in (5.5), but with $A(h)$ in place of A_n. The distribution of $\widehat{c}_{\mathcal{T},1-\alpha}(h)$ is the asymptotic distribution of $\widehat{c}_{\mathcal{T},1-\alpha,n}$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ for $\|b\| < \infty$.

ASSUMPTION ROB2: (i) $P(\mathcal{T}(h) = \widehat{c}_{\mathcal{T},1-\alpha}(h)) = 0$ $\forall h \in H$.
(ii) *If* $\Delta_2 > 0$, $\mathrm{NRP}(\Delta_1, \Delta_2; h^*) = \alpha$ *for some point* $h^* \in H$, *where* Δ_1 *and* Δ_2 *are defined in* (5.8).

The NI asymptotic quantile $\widehat{c}_{\mathcal{T},1-\alpha}(h, v)$ and Assumption NI-Rob2 are defined analogously to $\widehat{c}_{\mathcal{T},1-\alpha}(h)$ and Assumption Rob2; see Supplemental Appendix A for details.

For $\mathcal{T}_n$ equal to $|T_n|$, T_n, $-T_n$, or QLR_n, we have $\mathcal{T}(h)$ equal to $|T(h)|$, $T(h)$, $-T(h)$, or $\mathrm{QLR}(h)$, respectively, the quantile $c_{\mathcal{T},1-\alpha}(h)$ equal to $c_{|t|,1-\alpha}(h)$, $c_{t,1-\alpha}(h)$, $c_{-t,1-\alpha}(h)$, or $c_{\mathrm{QLR},1-\alpha}(h)$ defined just below (4.16), the quantile $c_{\mathcal{T},1-\alpha}(\infty)$ equal to $z_{1-\alpha/2}$, $z_{1-\alpha}$, $z_{1-\alpha}$, or $\chi^2_{d_r,1-\alpha}$, and the quantities $c^{\mathrm{LF}}_{\mathcal{T},1-\alpha}$, $c^{\mathrm{LF}}_{\mathcal{T},1-\alpha}(v)$, $\widetilde{c}_{\mathcal{T},1-\alpha,n}$, $\widetilde{c}_{\mathcal{T},1-\alpha,n}(v)$, $\widehat{c}_{\mathcal{T},1-\alpha,n}$, $\widehat{c}_{\mathcal{T},1-\alpha,n}(v)$, $\widehat{c}_{\mathcal{T},1-\alpha}(h)$, and $\widehat{c}_{\mathcal{T},1-\alpha}(h, v)$ defined as above with $\mathcal{T} = |t|$, t, $-t$, or QLR, respectively.

THEOREM 5.1: (a) *Suppose Assumptions* A, B1–B3, C1–C8, D1–D3, R, V1, *and* V2 *hold and* $d_r = 1$. *Then the nominal* $1-\alpha$ *symmetric two-sided*, *upper one-sided*, *and lower one-sided robust t CI's all have* $\mathrm{AsySz} = 1-\alpha$ *when based on the following critical values*: (i) *LF*, (ii) *NI-LF*, (iii) *type* 1 *robust*, (iv) *type* 1 *robust with NI critical values*, (v) *type* 2 *robust*, *and* (vi) *type* 2 *robust with NI critical values*, *provided the following additional Assumptions hold*, *respectively*: (i) LF, (ii) NI-LF, (iii) K *and* V3, (iv) K *and* V3, (v) Rob2, *and* (vi) NI-Rob2, *where* $\mathcal{T}(h)$ *in Assumptions* LF, NI-LF, Rob2, *and* NI-Rob2 *is equal to* $|T(h)|$, $T(h)$, *and* $-T(h)$ *in the two-sided*, *upper one-sided*, *and lower-sided cases*, *respectively*.
(b) *Suppose Assumptions* A, B1–B3, C1–C5, C7, C8, D1–D3, RQ1–RQ3, *and* RQ4(i) *hold*. *Then the nominal* $1-\alpha$ *QLR CS has* $\mathrm{AsySz} = 1-\alpha$ *when based on the following critical values*: (i) *LF*, (ii) *NI-LF*, (iii) *type* 1 *robust*, (iv) *type* 1 *robust with NI critical values*, (v) *type* 2 *robust*, *and* (vi) *type* 2 *robust with NI critical values*, *provided the following additional Assumptions hold*, *respectively*: (i) LF, (ii) NI-LF, (iii) K, RQ4, V1, *and* V2, (iv) K, RQ4, V1, *and* V2, (v) C6, Rob2,

V1, *and* V2, *and* (vi) C6, NI-Rob2, V1, *and* V2, *where* $\mathcal{T}(h)$ *in Assumptions* LF, NI-LF, Rob2, *and* NI-Rob2 *is equal to* QLR(h).

COMMENTS: (i) Plug-in versions of the robust CI's considered in Theorem 5.1 also have asymptotically correct size under continuity assumptions on $c_{\mathcal{T},1-\alpha}(h)$ that typically are not restrictive. For brevity, we do not provide formal results here. Theorem 5.1 also applies to robust tests that employ the NI-ICS statistic $A_n(v_n)$ in place of A_n.

(ii) If part (ii) of Assumptions LF, NI-LF, Rob2, or NI-Rob2 does not hold, then the corresponding part of Theorem 5.1(a) or (b) still holds, but with AsySz $\geq 1-\alpha$. For example, Assumption LF(ii) fails in the unusual case that $c^{\mathrm{LF}}_{\mathcal{T},1-\alpha}=\infty$ and Assumption NI-LF(ii) fails if $c^{LF}_{\mathcal{T},1-\alpha}(v)=\infty\ \forall v\in V_r$.

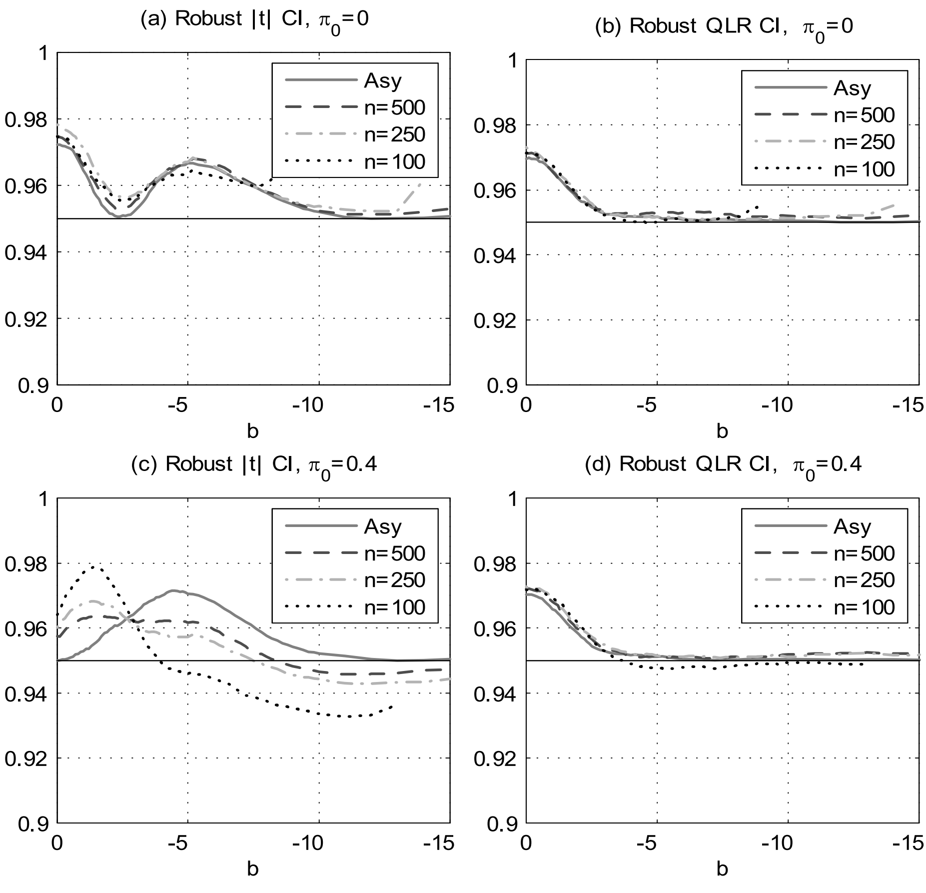

FIGURE 7.—Coverage probabilities of robust $|t|$ and QLR CI's for the MA parameter π in the ARMA$(1,1)$ model when $\pi_0=0$ and $\pi_0=0.4$, $\kappa=1.5$, and $s(x)=\exp(-x/2)$.

(iii) Figure 7 reports the asymptotic and finite-sample CP's of type 2 robust $|t|$ and QLR CI's for the MA parameter π in the ARMA$(1, 1)$ model as a function of b (≤ 0) for $\pi_0 = 0.0$ and 0.4. The type 2 robust CI's use NI critical values and the (unrestricted) ICS statistic A_n. They employ the transition function $s(x) = \exp(-x/2)$ and the constants $\kappa = 1.5$ and $D = 1$. The choices of $s(x)$ and D were determined via some experimentation to be good choices in terms of yielding CP's that are relatively close to the nominal size 0.95 across different values of b. Given $s(x)$ and D, the choice of κ was determined using the method described at the end of Section 5.3 based on minimizing average FCP's. The details are given in Supplemental Appendix D. It turns out that a wide range of κ values yields similar average FCP's, so the particular choice of $\kappa = 1.5$ is not at all crucial.[60,61]

Figure 7(a) and (b) shows that the CP's of both the $|t|$ and QLR CI's are greater than or equal to 0.95 for all b when $\pi_0 = 0.0$. However, the QLR CI is closer to being similar, both asymptotically and in finite samples. Only for $|b| \leq 3$ are its CP's greater than 0.95. The asymptotic approximations provided by Theorem 5.1 perform very well in Figure 7(a) and (b).

The results in Figure 7(d) for the QLR CI for $\pi_0 = 0.4$ are quite similar to those in Figure 7(b) for $\pi_0 = 0.0$. For the $|t|$ CI in Figure 7(c) for $\pi_0 = 0.4$, however, there is a greater discrepancy between the asymptotic and finite-sample results than when $\pi_0 = 0.0$. In addition, there is some undercoverage. For $n = 100$, the CP's of $|t|$ CI are as low as 0.93 for some b values. However, the magnitude of the undercoverage of the robust $|t|$ CI is very small compared to that of the standard $|t|$ CI.

5.5. *Asymptotic Power Comparisons for Robust QLR Tests*

In this section, we compare the power of type 2 robust QLR tests to the CLR test of Moreira (2003) in the linear IV regression model. The CLR test is approximately asymptotically optimal under weak and strong identification in the classes of invariant similar and invariant tests; see Andrews, Moreira, and Stock (2006, 2008) and Chernozhukov, Hansen, and Jansson (2009). This is the only model covered by the general results of this paper for which an asymptotically optimal test exists under weak identification, as far as we are aware. Hence, this is a good benchmark model to consider.

[60]This is shown in several tables in Supplemental Appendix D. The reason for similar average FCP's across different κ values is that if κ is changed, the constants Δ_1 and Δ_2 change in a manner that substantially offsets the effect of the change in κ. This occurs because, for any given κ, the constants Δ_1 and Δ_2 must yield a CI with the desired size.

[61]The value $\kappa = 1.5$ is used for all CI's considered, whether they are $|t|$ or QLR-based and whether they are for π or ρ. This value of κ minimizes the average FCP measured to two significant digits for all cases considered; see the tables in Supplemental Appendix D.

In short, we find that the type 2 robust test based on the NI-ICS statistic has power that is essentially equal to that of the CLR test. Hence, this robust test has approximately asymptotically optimal power. The type 2 robust test based on the unrestricted ICS statistic generally has lower power than the CLR test.

The structural model we consider is

$$y_{1,i} = y_{2,i}\pi + u_i^* \quad \text{and} \quad y_{2,i} = Z_i'\beta + v_i^*, \tag{5.9}$$

where $(u_i^*, v_i^*)' \sim N(0, \Upsilon^*)$ for a positive definite (p.d.) 2×2 matrix Υ^*, (u_i^*, v_i^*) and Z_i are independent, $\{(Z_i', u_i^*, v_i^*)' : i = 1, \ldots, n\}$ are i.i.d., $y_{1,i}, y_{2,i}, u_i^*, v_i^* \in R$, $Z_i \in R^k$, $\pi \in R$, and $\beta \in R^k$.[62,63] The reduced-form equations are

$$y_{1,i} = \pi \cdot Z_i'\beta + u_i \quad \text{and} \quad y_{2,i} = Z_i'\beta + v_i, \tag{5.10}$$

where $u_i = u_i^* + v_i^*\pi$, $v_i = v_i^*$, and $(u_i, v_i)' \sim N(0, \Upsilon)$.

Let $\zeta = \text{vech}(\Upsilon^{-1}) \in R^3$. The log-likelihood function for $\theta = (\beta, \zeta, \pi)$ (multiplied by $-n^{-1}$ and ignoring a constant) is

$$Q_n(\theta) = \frac{1}{2}\log|\Upsilon| + \frac{1}{2}n^{-1}\sum_{i=1}^{n}\varepsilon_i(\beta, \pi)'\Upsilon^{-1}\varepsilon_i(\beta, \pi), \quad \text{where} \tag{5.11}$$

$$\varepsilon_i(\beta, \pi) = (y_{1,i} - \pi \cdot Z_i'\beta, y_{2,i} - Z_i'\beta)' \in R^2.$$

Assumption A holds because $Q_n(\theta)$ does not depend on π when $\beta = 0$.

For brevity, Supplemental Appendix F provides the details of the parameter space, the quantities that appear in the assumptions and asymptotic distributions, formulae for the asymptotic power calculations (see (13.18), (13.21), and (13.23)), and verification of the assumptions for this model.

We now report asymptotic power comparisons for tests concerning the structural parameter π. We consider a type 2 robust QLR test that uses an NI-ICS statistic and one that uses an unrestricted ICS statistic. We compare them to the CLR test, as well as the LM test of Kleibergen (2002) and Moreira

[62] Using the notation of this paper, in which β determines the strength of identification of π, the parameters (β, π) in (5.9) are reversed from the usual notation used in the IV regression literature.

[63] For simplicity, we consider a model without exogenous variables X_i in either equation because they do not affect the asymptotic power comparisons. As is well known, such variables can be projected out and the results given here apply with Z_i being viewed as the projection residual; for example, see Section 2 of Andrews, Moreira, and Stock (2006) with a population projection in place of a sample projection. Provided X_i includes an intercept, this yields Z_i to have mean zero. Also for simplicity and because they do not affect the power comparisons, we assume the errors are normally distributed. The results can be extended to nonnormal finite variance errors, provided (u_i^*, v_i^*) is symmetrically distributed or the instruments have mean zero. By the discussion above, the latter is not restrictive.

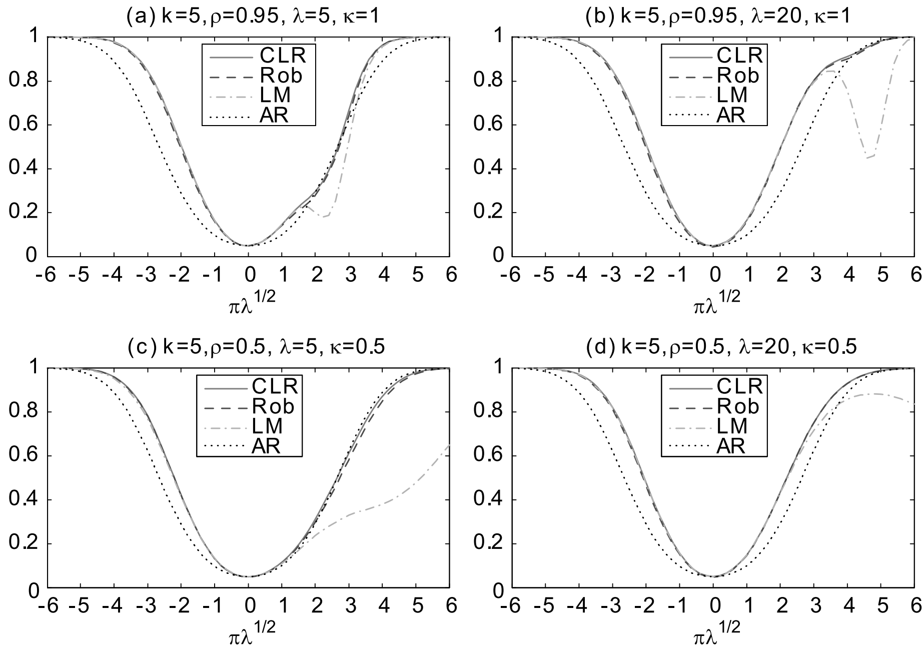

FIGURE 8.—Power functions for the CLR, robust QLR, LM, and AR tests for the structural parameter π in the linear IV model: $k = 5$, $\rho = 0.95$, 0.5, $\lambda = 5, 20$. The ICS statistic for the robust QLR test is the null-imposed Wald statistic.

(2009) and the well known Anderson–Rubin (AR) test. We report results for the same parameter configurations as in Andrews, Moreira, and Stock (2006). The asymptotic power of the tests just depends on $\lambda = b'b$, where $b = \lim_{n\to\infty} n^{1/2}\beta_n \in R^k$ indexes the strength of the IV's, the number of IV's k, the correlation between the reduced-form errors ρ, and $\pi - \pi_{H_0}$, where π denotes the true value of π and π_{H_0} is the null value of π, which we set to 0 w.l.o.g. The significance level of the tests is 5%. All results are based on 50,000 simulation repetitions. See Supplemental Appendix F for further details concerning the numerical work.

Figure 8 provides results for $\lambda = 5, 20$, $k = 5$, $\rho = 0.95$, 0.5, and $\pi\lambda^{1/2} \in [-6, 6]$. Figure 8 shows that the power of the robust QLR test that uses the NI-ICS statistic is essentially the same as that of the CLR test.

Figures in Supplemental Appendix D show a number of related results. First, the conclusion based on Figure 8 for $k = 5$ also holds for $k = 2, 10$. Second, the robust QLR test with NI-ICS statistic is close to being asymptotically similar. Third, the robust QLR test with unrestricted ICS statistic has power below that of the CLR test, more so when $\rho = 0.95$ than when $\rho = 0.5$. Fourth, the standard QLR CI for π exhibits substantial size distortions. For nominal level

95%, its asymptotic size varies between 0.6 and 0.9, depending on the parameter configurations.

6. ARMA EXAMPLE

In this section, we provide asymptotic results for the ARMA(1, 1) model specified in (1.1). It has been known for many years that common MA and AR roots lead to identification failure in the ARMA(1, 1) model in the important scenario where the series is white noise; see Ansley and Newbold (1980). Results for testing the null hypothesis of white noise in an ARMA(1, 1) model are provided by Hannan (1982) and Andrews and Ploberger (1996). However, no papers provide an asymptotic analysis of standard estimators, CI's, or tests for any other null hypothesis (such as tests concerning the MA or AR parameter) that deal with the identification issue. We do so here. We also provide identification-robust CI's.[64]

6.1. *Key Quantities*

We now specify the key quantities that arise in the ARMA model. More specifically, these quantities arise in Assumptions B1–B3, C1–C8, D1–D3, V1, and V2 and in the form of the asymptotic distributions. For brevity, these assumptions are verified in Supplemental Appendix C.

The (conditional) log-likelihood function $Q_n(\theta)$ is specified in (1.2). The conditioning value ε_0 is asymptotically negligible, so for simplicity (and w.l.o.g. for the asymptotic results), we set $\varepsilon_0 = Y_0$ in the log likelihood. See Supplemental Appendix C for details regarding the calculation of $Q_n(\theta)$. Let ϕ_0 denote the distribution of $\zeta_0^{-1/2}\varepsilon_t$. For notational simplicity, we sometimes write the true and generic AR parameters as $\rho_0 = \pi_0 + \beta_0$ and $\rho = \pi + \beta$, respectively.

The optimization and true parameter spaces are

$$\begin{aligned}(6.1)\qquad \Theta &= \{\theta = (\beta, \zeta, \pi)' : \beta \in [\rho_L - \pi, \rho_U - \pi], \\ &\qquad \zeta \in [\zeta_L, \zeta_U], \pi \in \Pi = [\pi_L, \pi_U]\}, \\ \Theta^* &= \{\theta = (\beta, \zeta, \pi)' : \beta \in [\rho_L^* - \pi, \rho_U^* - \pi], \\ &\qquad \zeta \in [\zeta_L^*, \zeta_U^*], \pi \in [\pi_L^*, \pi_U^*]\},\end{aligned}$$

[64]The results for this example can be extended to the case where the mean of the strictly stationary time series Y_t is μ_0. In this case, (1.1) holds with Y_t and Y_{t-1} replaced with $Y_t - \mu_0$ and $Y_{t-1} - \mu_0$, respectively. The mean μ_0 can be estimated by ML, in which case Y_t is replaced by $Y_t - \mu$ in the criterion function and the criterion function is minimized w.r.t. μ as well as the other parameters, or μ_0 can be estimated by $\overline{Y}_n = n^{-1}\sum_{t=1}^n Y_t$, in which case Y_t is replaced with $Y_t - \overline{Y}_n$ in the criterion function. In either case, the asymptotic results concerning (β, ζ, π) are the same whether or not μ_0 is estimated, due to the block diagonality of the information matrix between μ and (β, ζ, π).

where $-1 < \rho_L < \pi_L < \pi_U < \rho_U < 1$, $0 < \zeta_L < \zeta_U < \infty$, $\pi_L < \pi_L^* < \pi_U^* < \pi_U$, $\rho_L < \rho_L^* < \pi_L^* < \pi_U^* < \rho_U^* < \rho_U$, and $\zeta_L < \zeta_L^* < \zeta_U^* < \zeta_U$. By the definition of Θ, the autoregressive parameter $\rho = \pi + \beta$ lies in $[\rho_L, \rho_U]$.[65]

Let ξ_t denote the normalized innovation $\zeta^{-1/2}\varepsilon_t$, which has mean 0 and variance 1. The true parameter space for $\gamma = (\theta, \phi)$ is

$$\text{(6.2)} \qquad \Gamma = \{\gamma = (\theta, \phi) : \theta \in \Theta^*, \phi \in \Phi^*\},$$

where Φ^* is some compact subset of Φ w.r.t. the metric d_Φ and

$$\Phi = \left\{\phi : E_\phi \xi_t = 0, E_\phi \xi_t^2 = 1, E_\phi(\xi_t^2 - 1)^2 \geq \delta_1, E_\phi|\xi_t|^{4+\delta_2} \leq K\right\}$$

for some constants $\delta_1, \delta_2 > 0$ and $0 < K < \infty$, where d_Φ is some metric on the space of distributions on R that induces weak convergence. With these definitions of Θ, Θ^*, and Γ, Assumptions B1 and B2 hold; see Supplemental Appendix C.

In the ARMA example, the function $Q(\theta; \gamma_0)$ in Assumption B3(i) is

$$\text{(6.3)} \qquad Q(\theta; \gamma_0) = E_{\gamma_0}\rho_t(\theta), \quad \text{where}$$

$$\rho_t(\theta) = \frac{1}{2}\log\zeta + \frac{1}{2\zeta}\left(Y_t - \beta\sum_{j=0}^{\infty}\pi^j Y_{t-j-1}\right)^2.$$

The generalized derivatives of $Q_n(\theta)$ w.r.t. ψ, which appear in Assumption C1, are the ordinary first and second partial derivatives of the approximation $Q_n^\infty(\theta)$ to $Q_n(\theta)$. Here, $Q_n^\infty(\theta)$ is defined by

$$\text{(6.4)} \qquad Q_n^\infty(\theta) = \frac{1}{2}\log\zeta + \frac{1}{2\zeta}n^{-1}\sum_{t=1}^{n}\left(Y_t - \beta\sum_{j=0}^{\infty}\pi^j Y_{t-j-1}\right)^2,$$

where the sum over j runs to ∞, rather than to $t-1$.

Assumption C1 is verified with

$$\text{(6.5)} \qquad D_\psi Q_n(\theta) = n^{-1}\sum_{t=1}^{n}\rho_{\psi,t}(\theta) = \begin{pmatrix}\rho_{\beta,t}(\theta)\\ \rho_{\zeta,t}(\theta)\end{pmatrix}, \quad \text{where}$$

$$\rho_{\beta,t}(\theta) = -\zeta^{-1}\left(Y_t - \beta\sum_{j=0}^{\infty}\pi^j Y_{t-j-1}\right)\sum_{k=0}^{\infty}\pi^k Y_{t-k-1},$$

$$\rho_{\zeta,t}(\theta) = -\frac{1}{2}\zeta^{-2}\left(\left(Y_t - \beta\sum_{j=0}^{\infty}\pi^j Y_{t-j-1}\right)^2 - \zeta\right).$$

[65]The conditions $\rho_L < \pi_L$ and $\pi_U < \rho_U$ imply that β can take values in a neighborhood of zero for any value of $\pi \in \Pi$.

Assumption C2(i) holds in this example with $m(W_i, \theta) = \rho_{\psi,t}(\theta)$. Assumption C2(ii) holds because, for all $\gamma^* \in \Gamma$ with $\beta^* = 0$, $E_{\gamma^*}\rho_{\beta,t}(\psi^*, \pi) = -\zeta^{*-1}E_{\gamma^*}\varepsilon_t \sum_{j=0}^{\infty} \pi^j Y_{t-j-1} = 0$ and $E_{\gamma^*}\rho_{\zeta,t}(\psi^*, \pi) = -(1/2)\zeta^{*-2}(E_{\gamma^*}\varepsilon_t^2 - \zeta^*) = 0$ using (6.2) and the definitions of $\rho_{\beta,t}(\theta)$ and $\rho_{\zeta,t}(\theta)$ in (6.5).

The empirical process $\{G_n(\pi) : \pi \in \Pi\}$ in Assumption C3 is

$$G_n(\pi) = n^{-1/2}\sum_{t=1}^{n}\begin{pmatrix} -\zeta_n^{-1} Y_t \sum_{k=0}^{\infty} \pi^k Y_{t-k-1} \\ -(1/2)\zeta_n^{-2}(Y_t^2 - \zeta_n) \end{pmatrix} - \begin{pmatrix} -E_{\gamma_n}\zeta_n^{-1} Y_t \sum_{k=0}^{\infty} \pi^k Y_{t-k-1} \\ -E_{\gamma_n}(1/2)\zeta_n^{-2}(Y_t^2 - \zeta_n) \end{pmatrix}. \tag{6.6}$$

The limit process $\{G(\pi; \gamma_0) : \pi \in \Pi\}$ in Assumption C3 is the Gaussian process

$$G(\pi; \gamma_0) = \begin{pmatrix} \sum_{j=0}^{\infty} \pi^j Z_j \\ (1/2)\zeta_0^{-2}\big(E_{\gamma_0}(\varepsilon_t^2 - \zeta_0)^2\big)^{1/2} Z \end{pmatrix}, \tag{6.7}$$

where $Z, Z_0, Z_1, \ldots$ are independent standard normal random variables. The mean of $G(\pi; \gamma_0)$ is zero. The covariance kernel of $G(\pi; \gamma_0)$ is $\Omega(\pi_1, \pi_2; \gamma_0) = \text{Diag}\{(1 - \pi_1\pi_2)^{-1}, (1/4)\zeta_0^{-4}E_{\gamma_0}(\varepsilon_t^2 - \zeta_0)^2\} \in R^{2\times 2}$. The convergence in Assumption C3 is established using the method in Andrews and Ploberger (1996).

The matrices $D_{\psi\psi}Q_n(\psi_{0,n}, \pi)$ and $H(\pi; \gamma_0)$ in Assumption C4 are, for $\gamma_0 \in \Gamma$ with $\beta_0 = 0$,

$$D_{\psi\psi}Q_n(\psi_{0,n}, \pi) = n^{-1}\sum_{t=1}^{n}\begin{bmatrix} \rho_{\beta\beta,t}(\psi_{0,n}, \pi) & \rho_{\beta\zeta,t}(\psi_{0,n}, \pi) \\ \rho_{\beta\zeta,t}(\psi_{0,n}, \pi) & \rho_{\zeta\zeta,t}(\psi_{0,n}, \pi) \end{bmatrix}, \quad \text{where} \tag{6.8}$$

$$\rho_{\beta\beta,t}(\psi_{0,n}, \pi) = \zeta_n^{-1}\left(\sum_{j=0}^{\infty} \pi^j Y_{t-j-1}\right)^2,$$

$$\rho_{\beta\zeta,t}(\psi_{0,n}, \pi) = \zeta_n^{-2} Y_t \sum_{k=0}^{\infty} \pi^k Y_{t-k-1},$$

$$\rho_{\zeta\zeta,t}(\psi_{0,n}, \pi) = -(1/2)\zeta_n^{-2} + \zeta_n^{-3} Y_t^2,$$

$$H(\pi; \gamma_0) = E_{\gamma_0}\rho_{\psi\psi,t}(\psi_0, \pi) = \begin{bmatrix} (1 - \pi^2)^{-1} & 0 \\ 0 & (2\zeta_0^2)^{-1} \end{bmatrix}.$$

The matrix $H(\pi; \gamma_0)$ satisfies Assumption C4(ii) because $\inf_{\pi\in\Pi}(1 - \pi^2)^{-1} = (1 - \max^2\{|\pi_L|, |\pi_U|\})^{-1} > 0$.

The matrix $K_n(\theta;\gamma_0)$, which appears in Assumption C5(i), is complicated and, hence, for brevity, is given in (10.34), (10.36), and (10.38) in Supplemental Appendix C. Its limit, $K(\pi;\gamma_0)$, which appears in Assumption C5, is much simpler and is given by

$$K(\pi;\gamma_0)=\begin{pmatrix}-(1-\pi_0\pi)^{-1}\\ 0\end{pmatrix}. \tag{6.9}$$

Combining (6.7)–(6.9), the stochastic process $\xi(\pi;\gamma_0,b)$ is

$$\begin{aligned}\xi(\pi;\gamma_0,b)=&-\frac{1}{2}\left(G(\pi;\gamma_0)+\begin{pmatrix}-b/(1-\pi_0\pi)\\ 0\end{pmatrix}\right)'\begin{bmatrix}1-\pi^2 & 0\\ 0 & 2\zeta_0^2\end{bmatrix}\\ &\times\left(G(\pi;\gamma_0)+\begin{pmatrix}-b/(1-\pi_0\pi)\\ 0\end{pmatrix}\right).\end{aligned} \tag{6.10}$$

Assumption C6 is verified in this example using Assumption C6** and Lemma 8.5 given in Supplemental Appendix A.

In the ARMA example, the function $\eta(\pi;\gamma_0,\omega_0)$ in Assumption C7 is

$$\eta(\pi;\gamma_0,\omega_0)=-\frac{1-\pi^2}{2(1-\pi_0\pi)^2}. \tag{6.11}$$

It is uniquely minimized at $\pi=\pi_0$, as required by Assumption C7, because its derivative w.r.t. π is $(\pi-\pi_0)/(1-\pi_0\pi)^3$, which is zero for $\pi=\pi_0$, strictly negative for $\pi<\pi_0$, and strictly positive for $\pi>\pi_0$.

For brevity, the quantity $(\partial/\partial\psi')E_{\gamma_n}D_\psi Q_n(\psi,\pi_n)|_{\psi=\psi_n}$ in Assumption C8 and the verification of Assumption C8 is given in Supplemental Appendix C.

The matrix $B(\beta)$ for the ARMA example is $B(\beta)=\text{Diag}\{1,1,\beta\}\in R^{3\times 3}$. The generalized derivatives of $Q_n(\theta)$ w.r.t. θ that appear in Assumption D1 are the ordinary first and second partial derivatives of $Q_n^\infty(\theta)$, defined in (6.4). The first derivatives are

$$\begin{aligned}DQ_n(\theta)&=n^{-1}\sum_{t=1}^n\rho_{\theta,t}(\theta)\\ &=n^{-1}\sum_{t=1}^n(\rho_{\beta,t}(\theta),\rho_{\zeta,t}(\theta),\rho_{\pi,t}(\theta))',\quad\text{where}\\ \rho_{\pi,t}(\theta)&=-\zeta^{-1}\left(Y_t-\beta\sum_{j=0}^\infty\pi^jY_{t-j-1}\right)\beta\sum_{k=0}^\infty k\pi^{k-1}Y_{t-k-1},\end{aligned} \tag{6.12}$$

and $\rho_{\beta,t}(\theta)$ and $\rho_{\zeta,t}(\theta)$ are given in (6.5). For brevity, the second derivatives are given in (10.11)–(10.13) of Supplemental Appendix C.

Assumption D2 holds in this example with

$$
\begin{aligned}
(6.13)\qquad J(\gamma_0) = \mathrm{Diag}\Bigg\{ & \zeta_0^{-1} E_{\gamma_0}\Bigg(\sum_{j=0}^{\infty} \pi_0^j Y_{t-j-1}\Bigg)^2, (2\zeta_0^2)^{-1}, \\
& \zeta_0^{-1} E_{\gamma_0}\Bigg(\sum_{j=0}^{\infty} j\pi_0^{j-1} Y_{t-j-1}\Bigg)^2 \Bigg\} \\
& + \Bigg(\zeta_0^{-1} E_{\gamma_0}\Bigg(\sum_{j=0}^{\infty} \pi_0^j Y_{t-j-1}\Bigg) \sum_{k=0}^{\infty} k\pi_0^{k-1} Y_{t-k-1}\Bigg) \times \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.
\end{aligned}
$$

To verify Assumption D3(i) in this example, we have

$$
\begin{aligned}
(6.14)\qquad n^{1/2} B^{-1}(\beta_n) D Q_n(\theta_n) &= -n^{-1/2} \sum_{t=1}^{n} \begin{pmatrix} \zeta_n^{-1} \varepsilon_t \sum_{k=0}^{\infty} \pi_n^k Y_{t-k-1} \\ (1/2)\zeta_n^{-2}(\varepsilon_t^2 - \zeta_n) \\ \zeta_n^{-1} \varepsilon_t \sum_{k=0}^{\infty} k\pi_n^{k-1} Y_{t-k-1} \end{pmatrix} \\
&\to_d N(0, V(\gamma_0)),
\end{aligned}
$$

where the equality holds by the definitions in (6.5) and (6.12), and the convergence in distribution holds by a triangular array martingale difference CLT.

The matrix

$$
\begin{aligned}
(6.15)\qquad V(\gamma_0) = \mathrm{Diag}\Bigg\{ & \zeta_0^{-1} E_{\gamma_0}\Bigg(\sum_{j=0}^{\infty} \pi_0^j Y_{t-j-1}\Bigg)^2, \frac{E_{\gamma_0}(\varepsilon_t^2 - \zeta_0)^2}{4\zeta_0^4}, \\
& \zeta_0^{-1} E_{\gamma_0}\Bigg(\sum_{j=0}^{\infty} j\pi_0^{j-1} Y_{t-j-1}\Bigg)^2 \Bigg\} \\
& + \Bigg(\zeta_0^{-1} E_{\gamma_0}\Bigg(\sum_{j=0}^{\infty} \pi_0^j Y_{t-j-1}\Bigg) \sum_{k=0}^{\infty} k\pi_0^{k-1} Y_{t-k-1}\Bigg) \times \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.
\end{aligned}
$$

Note that $J(\gamma_0) = V(\gamma_0)$ if $(2\zeta_0^2)^{-1} = (4\zeta_0^4)^{-1} E_{\gamma_0}(\varepsilon_t^2 - \zeta_0)^2$, which holds when ε_t has a normal distribution.[66]

[66]The verification of the conditions needed for the CLT, the derivation of the form of $V(\gamma_0)$, and the verification of Assumption D3(ii) are given in Supplemental Appendix C.

In this example, $\tau_\beta(\pi;\gamma_0,b)$ of (3.10) equals $-(1-\pi^2)(\sum_{j=0}^{\infty}\pi^j Z_j-(1-\pi_0\pi)^{-1}b)$.

We estimate $J(\gamma_0)$ and $V(\gamma_0)$ by $\widehat{J}_n=\widehat{J}_n(\widehat{\theta}_n)$ and $\widehat{V}_n=\widehat{V}_n(\widehat{\theta}_n)$, respectively, where[67]

$$\begin{aligned}(6.16)\qquad \widehat{J}_n(\theta) = \mathrm{Diag}\Bigg\{&\zeta^{-1}n^{-1}\sum_{t=1}^{n}\Bigg(\sum_{j=0}^{t-1}\pi^j Y_{t-j-1}\Bigg)^2, (2\zeta^2)^{-1},\\ &\zeta^{-1}n^{-1}\sum_{t=1}^{n}\Bigg(\sum_{j=0}^{t-1}j\pi^{j-1}Y_{t-j-1}\Bigg)^2\Bigg\}\\ &+\Bigg(\zeta^{-1}n^{-1}\sum_{t=1}^{n}\Bigg(\sum_{j=0}^{t-1}\pi^j Y_{t-j-1}\Bigg)\sum_{k=0}^{t-1}k\pi^{k-1}Y_{t-k-1}\Bigg)\\ &\times\begin{bmatrix}0&0&1\\0&0&0\\1&0&0\end{bmatrix}\end{aligned}$$

and $\widehat{V}_n(\theta)=\widehat{J}_n(\theta)$ but with its $(2,2)$ element, $(2\zeta^2)^{-1}$, replaced by

$$(6.17)\qquad (4\zeta^2)^{-1}n^{-1}\sum_{t=1}^{n}\Bigg(\Bigg(Y_t-\beta\sum_{j=0}^{t-1}\pi^j Y_{t-j-1}\Bigg)^2-\zeta\Bigg)^2.$$

For brevity, the quantities $J(\theta;\gamma_0)$ and $V(\theta;\gamma_0)$ in Assumption V1 (scalar β) are given in (10.57) and (10.58) of Supplemental Appendix C.

The asymptotic null distribution of the t statistic for tests concerning the MA parameter π is determined by Theorem 4.1(b). Under $\{\gamma_n\}\in\Gamma(\gamma_0,0,b)$ with $|b|<\infty$, it is the distribution of

$$(6.18)\qquad T_\pi(\pi^*;\gamma_0,b)=\frac{\left|\sum_{j=0}^{\infty}(\pi^*)^j Z_j-(1-\pi_0\pi^*)^{-1}b\right|(1-(\pi^*)^2)(\pi^*-\pi_0)}{(\Sigma_{\pi\pi}(\pi^*)_{22})^{1/2}},$$

[67]For hypotheses and CI's that involve only β and/or π, the $(2,2)$ elements of $\widehat{J}_n$ and $\widehat{V}_n$ are not needed. In such cases, the matrices $\widehat{J}_n$ and $\widehat{V}_n$ with their second rows and columns deleted are the same. For Assumptions V1 and V2 to hold for the quantity in (6.17), more moments need to be assumed on ε_t. Specifically, in Φ (defined in (6.2)), the condition $E_\phi|\xi_t|^{4+\delta_2}\le K$ needs to be replaced by $E_\phi|\xi_t|^{8+\delta_2}\le K$ for the proof to go through. This condition is only needed for hypotheses and CI's that involve the innovation variance ζ.

where π^* abbreviates $\pi^*(\gamma_0, b)$, $\{Z_j : j \geq 0\}$ are i.i.d. $N(0,1)$ random variables,

$$\pi^*(\gamma_0, b) = \arg\min_{\pi \in \Pi} -\frac{1}{2}\left(\sum_{j=0}^{\infty} \pi^j Z_j - (1-\pi_0\pi)^{-1} b\right)^2 (1-\pi^2), \tag{6.19}$$

$$\Sigma_{\pi\pi}(\pi) = \begin{bmatrix} \sum_{j=0}^{\infty} \pi^{2j} & \sum_{j=0}^{\infty} j\pi^{2j-1} \\ \sum_{j=0}^{\infty} j\pi^{2j-1} & \sum_{j=0}^{\infty} j^2\pi^{2j-2} \end{bmatrix}^{-1},$$

and $\Sigma_{\pi\pi}(\pi)_{22}$ denotes the $(2,2)$ element of $\Sigma_{\pi\pi}(\pi)$.[68] The limit distribution in (6.18) only depends on b and π_0. Under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, the t statistic for the MA parameter π has a $N(0,1)$ asymptotic null distribution by Theorem 4.1(c).

We consider QLR tests and CS's involving functions of (β, π), not ζ. In consequence, the key Assumption RQ2(ii) for the QLR statistic holds.[69] It holds because $V(\gamma_0)$ and $J(\gamma_0)$ are block diagonal (after reordering their rows and columns) between the (β, π) and ζ parameters, and the blocks of $V(\gamma_0)$ and $J(\gamma_0)$ that correspond to the (β, π) parameters are equal; see (6.13) and (6.15). In consequence, $\widehat{s}_n = 1$ in this example and the standard critical value is $\chi^2_{d_r,1-\alpha}$.

By Theorem 4.2, for a test concerning the MA parameter π, the asymptotic null distribution of the QLR statistic under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $|b| < \infty$ is the distribution of

$$\begin{aligned} &2\Big(\xi(\pi_0; \gamma_0, b) - \inf_{\pi \in \Pi} \xi(\pi; \gamma_0, b)\Big) \\ &\quad = -\left(\sum_{j=0}^{\infty} \pi_0^j Z_j - (1-\pi_0^2)^{-1} b\right)^2 (1-\pi_0^2) \\ &\qquad + \inf_{\pi \in \Pi}\left(\sum_{j=0}^{\infty} \pi^j Z_j - (1-\pi_0\pi)^{-1} b\right)^2 (1-\pi^2). \end{aligned} \tag{6.20}$$

This limit distribution only depends on b and π_0.[70] Under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, the QLR statistic has a χ^2_1 asymptotic null distribution by Theorem 4.3 and (4.15).

[68]The first equality in (6.19) holds using the expression for $\xi(\pi; \gamma_0, b)$ in this example given in (6.10) plus simplifications based on (6.7)–(6.9).

[69]This assumption is needed for the QLR statistic to have a $\chi^2_{d_r}$ asymptotic null distribution under strong identification.

[70]The equality in (6.20) uses the simplifications in (6.19).

6.2. *AR Parameter*

The estimator $\widehat{\rho}_n = \widehat{\pi}_n + \widehat{\beta}_n$ of the AR parameter has the same asymptotic distribution as the estimator for the MA estimator $\widehat{\pi}_n$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$. This holds because $\widehat{\rho}_n = \widehat{\pi}_n + o_p(1)$ when $\|b\| < \infty$ and $\|\beta_n\|^{-1}(\widehat{\rho}_n - \rho_n) = \|\beta_n\|^{-1}(\widehat{\pi}_n - \pi_n) + o_p(1)$ when $\|b\| = \infty$. In consequence, the t statistics for ρ and π have the same asymptotic null distribution under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$. Furthermore, they have the same $N(0, 1)$ asymptotic null distribution under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$. For tests concerning the AR parameter ρ, the QLR statistic has the same asymptotic null distribution as given above for tests concerning the MA parameter π. This holds by Comment (iv) to Theorem 4.2 and Section 9.4.4 of Supplemental Appendix B. Hence, the asymptotic size properties of each test and CI considered here is the same for both ρ and π.

6.3. *Numerical Results*

Figures 1–7 above provide a variety of asymptotic and finite-sample numerical results for the ARMA$(1, 1)$ model. Additional numerical results are reported in Supplemental Appendix D. These include (i) analogous figures to the figures given above but for $\pi_0 = 0.0$ and 0.7, rather than $\pi_0 = 0.4$, (ii) analogous figures to those above but for the AR parameter $\rho = \pi + \beta$, rather than the MA parameter π, (iii) tables of asymptotic and finite-sample coverage probabilities for $|t|$ and QLR CI's for π and ρ, and (iv) tables giving FCP results for NI-LF and type 2 robust CI's for π and ρ. Generally speaking, the results for (i) and (ii) are similar to the results reported above. For brevity, details concerning the numerical results are provided in Supplemental Appendix D. Table S-I in Supplemental Appendix D provides the $c^{\mathrm{LF}}_{\mathcal{T},1-\alpha}(v)$, $\Delta_1(v)$, and $\Delta_2(v)$ values necessary to compute the type 2 NI robust critical values for the $|t|$ and QLR test statistics for computing CI's for the MA and AR parameters.

REFERENCES

ANDERSON, T. W., AND H. RUBIN (1949): "Estimators of the Parameters of a Single Equation in a Complete Set of Stochastic Equations," *Annals of Mathematical Statistics*, 21, 570–582. [2159]

ANDREWS, D. W. K. (1993): "Tests for Parameter Instability and Structural Change With Unknown Change Point," *Econometrica*, 61, 821–856. [2167]

——— (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, Vol. IV, ed. by R. F. Engle and D. McFadden. Amsterdam: North-Holland. [2170]

——— (1999): "Estimation When a Parameter Is on a Boundary," *Econometrica*, 67, 1341–1383. [2190]

——— (2000): "Inconsistency of the Bootstrap When a Parameter Is on the Boundary of the Parameter Space," *Econometrica*, 68, 399–405. [2190]

——— (2001): "Testing When a Parameter Is on the Boundary of the Maintained Hypothesis," *Econometrica*, 69, 683–734. [2157]

ANDREWS, D. W. K., AND P. BARWICK (2012): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *Econometrica*, 80 (forthcoming). [2192]

ANDREWS, D. W. K., AND X. CHENG (2011a): "Maximum Likelihood Estimation and Uniform Inference With Sporadic Identification Failure," Cowles Foundation Discussion Paper No. 1824, Yale University. [2157]

——— (2011b): "GMM Estimation and Uniform Subvector Inference With Possible Identification Failure," Cowles Foundation Discussion Paper No. 1828, Yale University. [2157]

——— (2012): "Supplement to 'Estimation and Inference With Weak, Semi-Strong, and Strong Identification'," *Econometrica Supplemental Material*, 80, http://www.econometricsociety.org/ecta/Supmat/9456_miscellaneous.pdf; http://www.econometricsociety.org/ecta/Supmat/9456_data_and_programs.zip. [2157,2160]

ANDREWS, D. W. K., AND P. GUGGENBERGER (2009): "Validity of Subsampling and 'Plug-in Asymptotic' Inference for Parameters Defined by Moment Inequalities," *Econometric Theory*, 25, 669–709. [2163]

——— (2010): "Asymptotic Size and a Problem With Subsampling and With the m out of n Bootstrap," *Econometric Theory*, 26, 426–468. [2163,2165]

ANDREWS, D. W. K., AND C. J. MCDERMOTT (1995): "Nonlinear Econometric Models With Deterministically Trending Variables," *Review of Economic Studies*, 62, 343–360. [2157]

ANDREWS, D. W. K., AND W. PLOBERGER (1994): "Optimal Tests When a Nuisance Parameter Is Present Only Under the Alternative," *Econometrica*, 62, 1383–1414. [2158]

——— (1995): "Admissibility of the Likelihood Ratio Test When a Nuisance Parameter Is Present Only Under the Alternative," *The Annals of Statistics*, 23, 1609–1629. [2158]

——— (1996): "Testing for Serial Correlation Against an ARMA$(1,1)$ Process," *Journal of the American Statistical Association*, 91, 1331–1342. [2200,2202]

ANDREWS, D. W. K., AND G. SOARES (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157. [2190]

ANDREWS, D. W. K., AND J. H. STOCK (2007): "Inference With Weak Instruments," in *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, Vol. III, ed. by R. Blundell, W. K. Newey, and T. Persson. Cambridge, U.K.: Cambridge University Press. [2158]

ANDREWS, D. W. K., X. CHENG, AND P. GUGGENBERGER (2009): *Generic Results for Establishing the Asymptotic Size of Confidence Intervals and Tests*, Cowles Foundation Discussion Paper No. 1813, Yale University. [2163]

ANDREWS, D. W. K., X. LIU, AND W. PLOBERGER (1998): "Tests for White Noise Against Alternatives With Both Seasonal and Non-Seasonal Serial Correlation," *Biometrika*, 85, 727–740. [2157]

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. [2156,2197-2199]

——— (2008): "Efficient Two-Sided Nonsimilar Invariant Tests in IV Regression With Weak Instruments," *Journal of Econometrics*, 146, 241–254. [2156,2197]

ANDREWS, I., AND A. MIKUSHEVA (2011): "Maximum Likelihood Inference in Weakly Identified DSGE Models," Unpublished Manuscript, Department of Economics, MIT. [2159]

ANSLEY, C. F., AND P. NEWBOLD (1980): "Finite Sample Properties of Estimators for Autoregressive Moving Average Models," *Journal of Econometrics*, 13, 159–183. [2155,2200]

ANTOINE, B., AND E. RENAULT (2009): "Efficient GMM With Nearly-Weak Instruments," *Journal of Econometrics*, 12, S135–S171. [2159,2163]

——— (2010): "Efficient Inference With Poor Instruments, a General Framework," in *Handbook of Empirical Economics and Finance*, ed. by D. Giles and A. Ullah. Oxford, U.K.: Taylor and Francis, Chapter 2, 29–70. [2159,2163]

BAI, J. (1997): "Estimation of a Change Point in Multiple Regression Models," *Review of Economics and Statistics*, 79, 551–563. [2158]

CANER, M. (2010): "Testing, Estimation in GMM and CUE With Nearly Weak Identification," *Econometric Reviews*, 29, 330–363. [2159]

CHAN, K. S. (1993): "Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Autoregressive Model," *The Annals of Statistics*, 21, 520–533. [2158]

CHAN, K. S., AND R. S. TSAY (1998): "Limiting Properties of the Least Squares Estimator of a Continuous Threshold Autoregressive Model," *Biometrika*, 85, 413–426. [2157,2169]

CHENG, X. (2008): "Robust Confidence Intervals in Nonlinear Regression Under Weak Identification," Unpublished Working Paper, Department of Economics, Yale University. [2158,2160]

CHERNOFF, H. (1954): "On the Distribution of the Likelihood Ratio," *Annals of Mathematical Statistics*, 25, 573–578. [2185]

CHERNOZHUKOV, V., C. HANSEN, AND M. JANSSON (2009): "Admissible Invariant Similar Tests for Instrumental Variables Regression," *Econometric Theory*, 25, 808–818. [2197]

CHO, J. S., I. ISHIDA, AND H. WHITE (2011): "Revisiting Tests for Neglected Nonlinearity Using Artificial Neural Networks," *Neural Computation*, 23, 1133–1186. [2158]

CHOI, I., AND P. C. B. PHILLIPS (1992): "Asymptotic and Finite Sample Distribution Theory for IV Estimators and Tests in Partially Identified Structural Equations," *Journal of Econometrics*, 51, 113–150. [2159]

DAVIES, R. B. (1977): "Hypothesis Testing When a Nuisance Parameter Is Present Only Under the Alternative," *Biometrika*, 64, 247–254. [2158]

——— (1987): "Hypothesis Testing When a Nuisance Parameter Is Present Only Under the Alternative," *Biometrika*, 74, 33–43. [2158]

DRTON, M. (2009): "Likelihood Ratio Tests and Singularities," *The Annals of Statistics*, 37, 979–1012. [2158]

DUFOUR, J.-M. (1997): "Impossibility Theorems in Econometrics With Applications to Structural and Dynamic Models," *Econometrica*, 65, 1365–1387. [2158,2159]

ELLIOTT, G., AND U. K. MÜLLER (2007): "Confidence Sets for the Date of a Single Break in Linear Time Series Regressions," *Journal of Econometrics*, 141, 1196–1218. [2156,2158]

——— (2008): "Pre and Post Break Parameter Inference," Unpublished Manuscript, Department of Economics, Princeton University. [2156,2158]

ELLIOTT, G., U. MÜLLER, AND M. WATSON (2011): "Nearly Optimal Tests When a Nuisance Parameter Is Present Under the Null Hypothesis," Unpublished Manuscript, Department of Economics, Princeton University. [2156]

FORCHINI, G., AND G. HILLIER (2003): "Conditional Inference for Possibly Unidentified Structural Equations," *Econometric Theory*, 19, 707–743. [2190]

GHYSELS, E., A. SINKO, AND R. VALKANOV (2007): "MIDAS Regressions: Further Results and New Directions," *Econometric Reviews*, 26, 53–90. [2157]

GLESER, L. J., AND J. T. HWANG (1987): "The Nonexistence of $100(1-\alpha)$ Confidence Sets of Finite Expected Diameter in Errors in Variables and Related Models," *The Annals of Statistics*, 15, 1351–1362. [2159]

HAHN, J., AND G. KUERSTEINER (2002): "Discontinuities of the Weak Instruments Limiting Distribution," *Economic Letters*, 75, 325–331. [2163]

HAN, S. (2009): "Identification and Inference in a Bivariate Probit Model With Weak Instruments," Unpublished Manuscript, Department of Economics, Yale University. [2157]

HANNAN, E. J. (1982): "Testing for Autocorrelation and Akaike's Criterion," *Journal of Applied Probability*, Special Volume 19A: Essays in Statistical Science, ed. by J. Gani and E. J. Hannan, 403–412. [2200]

HANSEN, B. E. (1996): "Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis," *Econometrica*, 64, 413–430. [2158]

——— (2000): "Sample Splitting and Threshold Estimation," *Econometrica*, 68, 575–603. [2158]

KLEIBERGEN, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70, 1781–1803. [2158,2198]

——— (2005): "Testing Parameters in GMM Without Assuming That They Are Identified," *Econometrica*, 73, 1103–1123. [2158]

LEE, L.-F. (1981): "Simultaneous Equations Models With Discrete and Ordered Variables," in *Econometric Analysis of Discrete Data*, ed. by C. F. Manski and D. McFadden. Cambridge, MA: MIT Press, Chapter 2, 346–364. [2157]

LEE, L.-F., AND A. CHESHER (1986): "Specification Testing When Score Test Statistics Are Identically Zero," *Journal of Econometrics*, 31, 121–149. [2159]
LIU, X., AND Y. SHAO (2003): "Asymptotics for Likelihood Ratio Tests Under Loss of Identifiability," *The Annals of Statistics*, 31, 807–832. [2158]
MA, J., AND C. R. NELSON (2008): "Valid Inference for a Class of Models Where Standard Inference Performs Poorly; Including Nonlinear Regression, ARMA, GARCH, and Unobserved Components," Unpublished Manuscript, Department of Economics, University of Washington. [2155,2159]
MAGNUSSON, L. M. (2010): "Inference in Limited Dependent Variable Models Robust to Weak Identification," *Econometrics Journal*, 13, S56–S79. [2157]
MOREIRA, M. J. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71, 1027–1048. [2156,2158,2197]
——— (2009): "Tests With Correct Size When Instruments Can Be Arbitrarily Weak," *Journal of Econometrics*, 152, 131–140. [2198,2199]
NELSON, C. R., AND R. STARTZ (1990): "Some Further Results on the Exact Small Sample Properties of the Instrumental Variables Estimator," *Econometrica*, 58, 967–976. [2158]
——— (2007): "The Zero-Information-Limit Condition and Spurious Inference in Weakly Identified Models," *Journal of Econometrics*, 138, 47–62. [2155,2159]
NELSON, F., AND L. OLSON (1978): "Specification and Estimation of a Simultaneous-Equation Model With Limited Dependent Variables," *International Economic Review*, 19, 695–709. [2157]
NEWEY, W. K., AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. IV, Ch. 36, ed. by R. F. Engle and D. McFadden. Amsterdam: North-Holland. [2167]
PARK, J. Y., AND P. C. B. PHILLIPS (1999): "Asymptotics for Nonlinear Transformations of Integrated Time Series," *Econometric Theory*, 15, 269–298. [2160]
——— (2001): "Nonlinear Regressions With Integrated Time Series," *Econometrica*, 69, 117–161. [2160]
PHILLIPS, P. C. B. (1989): "Partially Identified Econometric Models," *Econometric Theory*, 5, 181–240. [2159]
PICARD, D. (1985): "Testing and Estimating Change-Point in Time Series," *Advances in Applied Probability*, 17, 841–867. [2158]
POLLARD, D. (1984): *Convergence of Stochastic Processes*. New York: Springer-Verlag. [2170]
——— (1990): *Empirical Processes: Theory and Applications*. CBMS Conference Series in Statistics and Probability, Vol. 2. Hayward, CA: Institute of Mathematical Statistics. [2170]
PRATT, J. W. (1961): "Length of Confidence Intervals," *Journal of the American Statistical Association*, 56, 549–567. [2194]
QU, Z. (2011): "Inference and Specification Testing in DSGE Models With Possible Weak Identification," Unpublished Manuscript, Department of Economics, Boston University. [2159]
QU, Z., AND P. PERRON (2007): "Estimating and Testing Structural Changes in Multivariate Regressions," *Econometrica*, 75, 459–502. [2158]
RIVERS, D., AND Q. H. VUONG (1988): "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," *Journal of Econometrics*, 39, 347–366. [2157]
ROTNITZKY, A., D. R. COX, M. BOTTAI, AND J. ROBINS (2000): "Likelihood-Based Inference With Singular Information Matrix," *Bernoulli*, 6, 243–284. [2159]
SARGAN, J. D. (1983): "Identification and Lack of Identification," *Econometrica*, 51, 1605–1633. [2154,2159]
SHI, X., AND P. C. B. PHILLIPS (2011): "Nonlinear Cointegrating Regression Under Weak Identification," *Econometric Theory*, 28, 1–39. [2160]
STAIGER, D., AND J. H. STOCK (1997): "Instrumental Variables Regression With Weak Instruments," *Econometrica*, 65, 557–586. [2158,2159,2163]
STOCK, J. H., AND J. H. WRIGHT (2000): "GMM With Weak Instruments," *Econometrica*, 68, 1055–1096. [2158,2163]

TRIPATHI, G. T. (1999): "A Matrix Extension of the Cauchy–Schwarz Inequality," *Economics Letters*, 63, 1–3. [2174]

VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*. New York: Springer. [2170,2185]

Cowles Foundation, Yale University, 30 Hillhouse Ave., New Haven, CT 06520-8281, U.S.A.; donald.andrews@yale.edu

and

Dept. of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, U.S.A.; xu.cheng@sas.upenn.edu.

Manuscript received July, 2010; final revision received January, 2012.