GMM ESTIMATION AND UNIFORM SUBVECTOR INFERENCE WITH POSSIBLE IDENTIFICATION FAILURE

BY

Donald W.K. Andrews and Xu Cheng

COWLES FOUNDATION PAPER NO. 1414

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

2014

http://cowles.econ.yale.edu/

Econometric Theory, **30**, 2014, 287–333.
doi:10.1017/S0266466613000315

GMM ESTIMATION AND UNIFORM SUBVECTOR INFERENCE WITH POSSIBLE IDENTIFICATION FAILURE

Donald W.K. Andrews
Cowles Foundation for Research in Economics, Yale University

Xu Cheng
University of Pennsylvania

This paper determines the properties of standard generalized method of moments (GMM) estimators, tests, and confidence sets (CSs) in moment condition models in which some parameters are unidentified or weakly identified in part of the parameter space. The asymptotic distributions of GMM estimators are established under a full range of drifting sequences of true parameters and distributions. The asymptotic sizes (in a uniform sense) of standard GMM tests and CSs are established.

The paper also establishes the correct asymptotic sizes of "robust" GMM-based Wald, t, and quasi-likelihood ratio tests and CSs whose critical values are designed to yield robustness to identification problems.

The results of the paper are applied to a nonlinear regression model with endogeneity and a probit model with endogeneity and possibly weak instrumental variables.

1. INTRODUCTION

This paper gives a set of generalized method of moments (GMM) regularity conditions that are akin to the classic conditions in Hansen (1982) and Pakes and Pollard (1989). But, they allow for singularity of the GMM estimator's variance matrix due to the lack of identification of some parameters in part of the parameter space.[1] This paper is a sequel to Andrews and Cheng (2012a) (AC1). The latter paper provides results for general extremum estimators, t tests, and quasi-likelihood ratio (QLR) tests in the presence of possible weak identification under high-level assumptions. Here we provide more primitive conditions for GMM-based statistics by verifying the high-level assumptions of AC1. This paper provides results for Wald tests and confidence sets (CSs) that apply not only to GMM estimators but also to other extremum estimators covered by AC1. This paper also provides some results for minimum distance (MD) estimators,

Andrews gratefully acknowledges the research support of the National Science Foundation via grants SES-0751517 and SES-1058376. The authors thank a co-editor, two referees, Tim Armstrong, Xiaohong Chen, Sukjin Han, Yuichi Kitamura, Peter Phillips, Eric Renault, Frank Schorfheide, and Ed Vytlacil for helpful comments. Address correspondence to Xu Cheng, Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA, 19104. e-mail: xucheng@econ.upenn.edu.

tests, and CSs. Lastly, the paper analyzes two specific models that are not considered in AC1.

Under the conditions given, the asymptotic distributions of GMM estimators and Wald and QLR test statistics are established. The asymptotic sizes of standard GMM tests and CSs are established. In many cases, their asymptotic sizes are not correct. We show that Wald and QLR statistics combined with "identification robust" critical values have correct asymptotic sizes (in a uniform sense).

In contrast to standard GMM results in the literature, the results given here cover a full range of drifting sequences of true parameters and distributions. Such results are needed to establish the (uniform) asymptotic size properties of tests and CSs and to give good approximations to the finite-sample properties of estimators, tests, and CSs under weak identification. Nonsmooth sample moment conditions are allowed, as in Pakes and Pollard (1989) and Andrews (2002).

We consider moment condition models where the parameter θ 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. The parameters β, ζ, and π may be scalars or vectors. For example, this framework applies to the nonlinear regression model $Y_i = \beta \cdot h\left(X_{1,i}, \pi\right) + X_{2,i}'\zeta + U_i$ with endogenous variables $X_{1,i}$ or $X_{2,i}$ and instrumental variables (IVs) Z_i. Here lack of identification of π when $\beta = 0$ occurs because of nonlinearity. This framework also applies to the probit model with endogeneity: $y_i^* = Y_i\pi + X_i'\zeta_1^* + U_i^*$, where one observes $y_i = 1(y_i^* > 0)$, the endogenous variable Y_i, and the exogenous regressor vector X_i and the reduced form for Y_i is $Y_i = Z_i'\beta + X_i'\zeta_2 + V_i$. In this case, lack of identification of π occurs when $\beta = 0$ because the IVs are irrelevant.

We determine the asymptotic properties of GMM estimators and tests under drifting sequences of true parameters $\theta_n = (\beta_n, \zeta_n, \pi_n)$ for $n \geq 1$, where n indexes the sample size. The behavior of GMM estimators and tests depends on the magnitude of $||\beta_n||$. The asymptotic behavior of these statistics varies across three categories of sequences $\{\beta_n : n \geq 1\}$: Category I(a) $\beta_n = 0\ \forall n \geq 1$, π is unidentified; Category I(b) $\beta_n \neq 0$ and $n^{1/2}\beta_n \to b \in R^{d_\beta}$, π is weakly identified; Category II $\beta_n \to 0$ and $n^{1/2}||\beta_n|| \to \infty$, π is semistrongly identified; and Category III $\beta_n \to \beta_0 \neq 0$, π is strongly identified.

For Category I sequences, GMM estimators, tests, and CSs are shown to have nonstandard asymptotic properties. For Category II and III sequences, they are shown to have standard asymptotic properties such as normal and chi-squared distributions (under suitable assumptions). However, for Category II sequences, the rates of convergence of estimators of π are slower than $n^{1/2}$, and tests concerning π do not have power against $n^{-1/2}$-local alternatives. Furthermore, for Category II sequences, it is shown that Wald tests of certain (rather unusual) nonlinear hypotheses can have asymptotic null rejection probabilities equal to 1.0, rather than the desired nominal size $\alpha \in (0, 1)$, due to the different rates of convergence of $\widehat{\beta}_n$ and $\widehat{\pi}_n$. This can occur even though $\widehat{\beta}_n$ and $\widehat{\pi}_n$ are consistent and asymptotically normal. Conditions are provided under which the asymptotic null rejection probabilities of Wald tests equal their nominal size for Category II sequences.

See Armstrong, Hong, and Nekipelov (2012) for some related, but different, results.

Numerical results for the nonlinear regression model with endogeneity show that the GMM estimators of both β and π have highly nonnormal asymptotic and finite-sample ($n = 500$) distributions when π is unidentified or weakly identified. The asymptotics provide excellent approximations to the finite-sample distributions. Nominal 95% standard t confidence intervals (CIs) for β are found to have asymptotic size equal to 68% and finite-sample size of 72%. In contrast, nominal 95% standard QLR CIs for β have asymptotic and finite-sample size of 93%. There are no asymptotic size distortions for the standard t and QLR CIs for π, and the finite-sample sizes are close to the asymptotic values. However, the CIs for π are far from being similar asymptotically or in finite samples. The robust CIs for β have correct asymptotic size. Their finite-sample sizes are 91.5% for t CIs and 95% for QLR CIs for nominal 95% CIs.

To conclude, the numerical results show that (i) weak identification can have substantial effects on the properties of estimators and standard tests and CSs; (ii) the asymptotic results of the paper provide useful approximations to the finite-sample distributions of estimators and test statistics under weak identification and identification failure; and (iii) the robust tests and CSs improve the size properties of tests and CSs in finite-samples noticeably compared to standard tests and CSs.

Like the results in Hansen (1982), Pakes and Pollard (1989), and Andrews (2002), the present paper applies when the GMM criterion function has a stochastic quadratic approximation as a function of θ. This rules out a number of models of interest in which identification failure may appear, including regime switching models, mixture models, abrupt transition structural change models, and abrupt transition threshold autoregressive models.[2] This paper applies when the GMM criterion function does not depend on β when $\pi = 0$. This also rules out some models of interest, such as nonlinear regression models with endogeneity and (potentially) weak instruments.

Now, we discuss the literature related to this paper. The following papers are companions to this one: AC1, Andrews and Cheng (2012b) (AC1-SM), and Andrews and Cheng (2013a) (AC2). These papers provide related, complementary results to the present paper. AC1 provides results under high-level conditions and analyzes the ARMA(1,1) model in detail. AC1-SM provides proofs for AC1 and related results. AC2 provides primitive conditions and results for estimators and tests based on log likelihood criterion functions. It provides applications to a smooth transition threshold autoregressive (STAR) model and a nonlinear binary choice model.

Cheng (2008) establishes results for a nonlinear regression model with multiple sources of weak identification, whereas the present paper only considers a single source. However, the present paper applies to a much broader range of models.

Tests of $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ are tests in which a nuisance parameter π only appears under the alternative. Such tests have been considered in the literature since Davies (1977). The results of this paper cover tests of this sort, in addition to tests for a whole range of linear and nonlinear hypotheses that involve (β, ζ, π) and corresponding CSs.

The weak instrument literature is closely related to this paper. This is true especially of Stock and Wright (2000), Kleibergen (2005), and Guggenberger, Kleibergen, Mavroeidis, and Chen (2012). In comparison to Stock and Wright (2000), the present paper differs because it focuses on criterion functions that are indexed by a parameter β that determines the strength of identification. It also differs in that it considers subvector analysis. In contrast to Kleibergen (2005) and Guggenberger et al. (2012), the present paper does not focus on Lagrange multiplier statistics. Rather, it investigates the behavior of standard estimators and tests, in addition to robust tests based on Wald and QLR statistics. Other related papers from the weak IV literature include Nelson and Startz (1990), Dufour (1997), Staiger and Stock (1997), Kleibergen (2002), and Moreira (2003).

Antoine and Renault (2009, 2010) and Caner (2010) consider GMM estimation with IVs that lie in the semistrong category, using our terminology. Nelson and Startz (2007) and Ma and Nelson (2008) analyze models like those considered in this paper. They do not provide asymptotic results or robust tests and CSs of the sort given in this paper. Andrews and Mikusheva (2011) and Qu (2011) consider Lagrange multiplier tests in a maximum likelihood (ML) context where identification may fail, with emphasis on dynamic stochastic general equilibrium models. Andrews and Mikusheva (2012) consider subvector inference based on Anderson-Rubin-type MD statistics.

In likelihood scenarios, Lee and Chesher (1986) consider Lagrange multiplier 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 situation than considered here for two reasons. First, the present paper considers situations where identification fails at some parameter values in the parameter space (and this causes the GMM variance matrix to be singular at these parameter values). Second, this paper considers GMM-based statistics rather than likelihood-based statistics.

Sargan (1983), Phillips (1989), and Choi and Phillips (1992) establish finite-sample and asymptotic results for linear simultaneous equations models when some parameters are not identified. Shi and Phillips (2011) provide results for a nonlinear regression model with nonstationary regressors in which identification may fail.

The remainder of the paper is organized as follows. Section 2 defines the GMM estimators, criterion functions, tests, and confidence sets considered in the paper and specifies the drifting sequences of distributions that are considered. It also introduces the two examples that are considered in the paper. Section 3 states the assumptions employed. Section 4 provides the asymptotic results for the GMM estimators. Section 5 establishes the asymptotic distributions of Wald statistics under the null and under alternatives, determines the asymptotic size of standard Wald CSs, and introduces robust Wald tests and CSs, whose asymptotic size is equal to their nominal size. Section 6 considers QLR CSs based on the GMM criterion function. Section 7 provides numerical results for the nonlinear regression model with endogeneity.

Andrews and Cheng (2013b) provides five supplemental appendixes to this paper. Supplemental Appendix A verifies the assumptions of the paper for the probit model with endogeneity. Supplemental Appendix B provides proofs of the GMM estimation results given in Section 4. It also provides some results for MD estimators. Supplemental Appendix C provides proofs of the Wald test and CS results given in Section 5. Supplemental Appendix D provides some results used in the verification of the assumptions for the two examples. Supplemental Appendix E provides some additional numerical results for the nonlinear regression model with endogeneity.

All limits that follow are taken as $n \to \infty$. We 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.

We 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. We let $\Rightarrow$ denote weak convergence of a sequence of stochastic processes indexed by $\pi \in \Pi$ for some space Π. We employ the uniform metric d on the space $\mathcal{E}_v$ of R^v-valued functions on Π. See AC1-SM for more details regarding this.

2. ESTIMATOR, CRITERION FUNCTION, AND EXAMPLES

2.1. GMM Estimators

The GMM sample criterion function is

$$Q_n(\theta) = \overline{g}_n(\theta)' \mathcal{W}_n(\theta) \overline{g}_n(\theta)/2, \tag{2.1}$$

where $\overline{g}_n(\theta) : \Theta \to R^k$ is a vector of sample moment conditions and $\mathcal{W}_n(\theta) : \Theta \to R^{k \times k}$ is a symmetric random weight matrix.

The paper considers inference when θ is not identified (by the criterion function $Q_n(\theta)$) at some points in the parameter space. Lack of identification occurs when $Q_n(\theta)$ is flat with respect to some subvector of θ. To model this identification problem, θ is partitioned into three subvectors:

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

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 affect the identification of π. These conditions allow for a broad range of cases, including cases where reparametrization is used to transform a model into the framework considered here.

The true distribution of the observations $\{W_i : i \geq 1\}$ is denoted F_γ for some parameter $\gamma \in \Gamma$. We let P_γ and $\mathbb{E}_\gamma$ denote probability and expectation under F_γ. The parameter space Γ for the true parameter, referred to as the "true parameter space," is compact and is of the form

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

where 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 $i, m \geq 1$.[3] In the case of a moment condition model, the parameter ϕ indexes the part of the distribution of the observations that is not determined by the moment conditions, which typically is infinite dimensional.

By definition, the GMM estimator $\widehat{\theta}_n$ (approximately) minimizes $Q_n(\theta)$ over an "optimization parameter space" Θ:[4]

$$\widehat{\theta}_n \in \Theta \quad \text{and} \quad Q_n(\widehat{\theta}_n) = \inf_{\theta \in \Theta} Q_n(\theta) + o(n^{-1}). \tag{2.4}$$

We assume that the interior of Θ includes the true parameter space Θ^* (see Assumption B1 in Section 3.7). This ensures that the asymptotic distribution of $\widehat{\theta}_n$ is not affected by boundary restrictions for any sequence of true parameters in Θ^*. The focus of this paper is not on the effects of boundary restrictions.

Without loss of generality, the optimization parameter space Θ can be written as

$$\begin{aligned} \Theta &= \{\theta = (\psi, \pi) : \psi \in \Psi(\pi), \pi \in \Pi\}, \quad \text{where} \\ \Pi &= \{\pi : (\psi, \pi) \in \Theta \text{ for some } \psi\} \quad \text{and} \\ \Psi(\pi) &= \{\psi : (\psi, \pi) \in \Theta\} \quad \text{for } \pi \in \Pi. \end{aligned} \tag{2.5}$$

We allow $\Psi(\pi)$ to depend on π, and hence Θ need not be a product space between ψ and π.

The main focus of this paper is on GMM estimators, but the results also apply to MD estimators. However, the assumptions employed with MD estimators are not as primitive. The MD sample criterion function is defined exactly as the GMM criterion function is defined in (2.1) except that $\overline{g}_n(\theta)$ is not a vector of moment conditions but rather is the difference between an unrestricted estimator $\widehat{\xi}_n$ of a parameter ξ_0 and a vector of restrictions $h(\theta)$ on ξ_0. That is,

$$\overline{g}_n(\theta) = \widehat{\xi}_n - h(\theta), \quad \text{where } \xi_0 = h(\theta_0). \tag{2.6}$$

See Schorfheide (2011) for a discussion of MD estimation of dynamic stochastic general equilibrium models and weak identification problems in these models.

2.2. Example 1: Nonlinear Regression with Endogeneity

The first example is a nonlinear regression model with endogenous regressors estimated using IVs. The IVs are assumed to be strong. Potential identification failure in this model arises because of the nonlinearity in the regression function. Let $h(x,\pi) \in R$ be a function of x that is known up to the finite-dimensional parameter $\pi \in R^{d_\pi}$. The model is

$$Y_i = \beta \cdot h\left(X_{1,i},\pi\right) + X_{2,i}'\zeta + U_i \quad \text{and} \quad \mathbb{E}U_i Z_i = 0 \tag{2.7}$$

for $i = 1,...,n$, where $X_i = (X_{1,i}, X_{2,i}) \in R^{d_X}$, $X_{2,i} \in R^{d_{X_2}}$, $Z_i \in R^k$, and $k \geq d_{X_2} + d_\pi + 1$. The regressors X_i may be endogenous or exogenous. The function $h(x,\pi)$ is assumed to be twice continuously differentiable with respect to π. Let $h_\pi(x,\pi)$ and $h_{\pi\pi}(x,\pi)$ denote the first- and second-order partial derivatives of $h(x,\pi)$ with respect to π. For example, Areosa, McAleer, and Medeiros (2011) consider GMM estimation of smooth transition models with endogeneity (which are nonlinear regression models). In their case $h(x,\pi)$ involves the logistic function. They provide an empirical application of this model to inflation rate targeting in Brazil.

The GMM sample criterion function is

$$Q_n(\theta) = \overline{g}_n(\theta)'\mathcal{W}_n\overline{g}_n(\theta)/2, \quad \text{where } \overline{g}_n(\theta) = n^{-1}\sum_{i=1}^n U_i(\theta)Z_i,$$

$$U_i(\theta) = Y_i - \beta h\left(X_{1,i},\pi\right) - X_{2,i}'\zeta, \quad \text{and} \quad \mathcal{W}_n = \left(n^{-1}\sum_{i=1}^n Z_i Z_i'\right)^{-1}. \tag{2.8}$$

For simplicity, we use the optimal weight matrix under homoskedasticity. Alternatively, one can employ the optimal weight matrix under heteroskedasticity using a preliminary estimator $\overline{\theta}_n$. Provided $\mathcal{W}_n(\theta)$ and $\overline{\theta}_n$ satisfy the assumptions in Lemma 3.1 in Section 3.1, all results hold for this two-step estimator also. For example, the preliminary estimator $\overline{\theta}_n$ can be the estimator obtained under homoskedasticity, which is shown subsequently to satisfy the assumptions in Lemma 3.1.

When $\beta = 0$, $U_i(\theta)$ does not depend on π. In consequence, $Q_n(\theta)$ does not depend on π when $\beta = 0$.

Suppose the random variables $\{(X_i, Z_i, U_i) : i = 1,...,n\}$ are independent and identically distributed (i.i.d.) with distribution $\phi \in \Phi^*$, where Φ^* is a compact metric space with a metric d_Φ that induces weak convergence of (X_i, Z_i, U_i). In this example, the parameter of interest is $\theta = (\beta, \zeta, \pi)$ and the nuisance parameter is ϕ, which is infinite dimensional.

The true parameter space for θ is

$$\Theta^* = \mathcal{B}^* \times \mathcal{Z}^* \times \Pi^*, \quad \text{where } \mathcal{B}^* = \left[-b_1^*, b_2^*\right] \subset R, \tag{2.9}$$

$b_1^* \geq 0$, $b_2^* \geq 0$, b_1^* and b_2^* are not both equal to 0, $\mathcal{Z}^* \subset R^{d_\zeta}$ is compact, and $\Pi^* \subset R^{d_\pi}$ is compact.

Suppose $||h_{\pi\pi}(x,\pi_1) - h_{\pi\pi}(x,\pi_2)|| \leq M_{\pi\pi}(x)\delta \ \ \forall \pi_1, \pi_2 \in \Pi$ with $||\pi_1 - \pi_2|| \leq \delta$ for some nonstochastic function $M_{\pi\pi}(x) : \mathcal{X} \to R^+$ that satisfies the conditions in (2.11) later in this section, where δ is some positive constant and $\mathcal{X}$ denotes the union of the supports of $X_{1,i}$ over all $\phi \in \Phi^*$. Define

$$\begin{aligned} d_i(\pi) &= \left(h\left(X_{1,i},\pi\right), X_{2,i}, h_\pi\left(X_{1,i},\pi\right)\right) \in R^{d_{X_2}+d_\pi+1} \quad \text{and} \\ d^*_{\psi,i}(\pi_1,\pi_2) &= \left(h\left(X_{1,i},\pi_1\right), h\left(X_{1,i},\pi_2\right), X_{2,i}\right) \in R^{d_{X_2}+2}. \end{aligned} \tag{2.10}$$

Let $\mathbb{E}_\phi$ denote expectation under ϕ. For any $\theta^* \in \Theta^*$, the true parameter space for ϕ is

$$\begin{aligned} \Phi^*(\theta^*) = \Big\{ &\phi \in \Phi^* : \mathbb{E}_\phi U_i Z_i = 0,\ \mathbb{E}_\phi(U_i^2|X_i,Z_i) = \sigma^2(X_i,Z_i) > 0 \text{ a.s.},\ \mathbb{E}_\phi|U_i|^{4+\varepsilon} \leq C, \\ &\mathbb{E}_\phi \sup_{\pi\in\Pi}\left(||h\left(X_{1,i},\pi\right)||^{2+\varepsilon} + ||h_\pi\left(X_{1,i},\pi\right)||^{2+\varepsilon} + ||h_{\pi\pi}\left(X_{1,i},\pi\right)||^{1+\varepsilon}\right) \leq C, \\ &\mathbb{E}_\phi(\left\|X_{2,i}\right\|^{2+\varepsilon} + ||Z_i||^{2+\varepsilon} + M_{\pi\pi}(X_{1,i})) \leq C,\ \lambda_{\min}(\mathbb{E}_\phi Z_i Z_i') \geq \varepsilon, \\ &\mathbb{E}_\phi Z_i d^*_{\psi,i}(\pi_1,\pi_2)' \in R^{k\times(d_{X_2}+2)} \text{ has full column rank } \forall \pi_1,\pi_2 \in \Pi \text{ with } \pi_1 \neq \pi_2, \\ &\mathbb{E}_\phi Z_i d_i(\pi) \in R^{k\times(d_{X_2}+d_\pi+1)} \text{ has full column rank } \forall \pi \in \Pi \Big\}, \end{aligned} \tag{2.11}$$

for some constants $C < \infty$ and $\varepsilon > 0$. Note that in this example $\Phi^*(\theta^*)$ does not depend on θ^*.

2.3. Example 2: Probit Model with Endogeneity and Possibly Weak Instruments

The second example is a probit model with endogeneity and IVs that may be weak or irrelevant, which causes identification issues. Consider the following two-equation model with endogeneity of Y_i in the first equation:

$$\begin{aligned} y_i^* &= Y_i\pi + X_i'\zeta_1^* + U_i^* \quad \text{and} \\ Y_i &= Z_i'\beta + X_i'\zeta_2 + V_i, \end{aligned} \tag{2.12}$$

where $y_i^*, Y_i, U_i^*, V_i \in R$, $X_i \in R^{d_X}$, $Z_i \in R^{d_Z}$, and $\{(X_i, Z_i, U_i, V_i) : i = 1, ..., n\}$ are i.i.d. The outcome variable y_i^* of the first equation is not observed. Only the binary indicator $y_i = 1(y_i^* > 0)$ is observed, along with Y_i, X_i, and Z_i. That is, we observe $\{W_i = (y_i, Y_i, X_i, Z_i) : i = 1, ..., n\}$. Similar models with binary, truncated, or censored endogenous variables are considered in Amemiya (1974), Heckman (1978), Nelson and Olson (1978), Lee (1981), Smith and Blundell (1986), and Rivers and Vuong (1988), among others.

The reduced-form equations of the model are

$$\begin{aligned} y_i^* &= Z_i'\beta\pi + X_i'\zeta_1 + U_i \quad \text{and} \\ Y_i &= Z_i'\beta + X_i'\zeta_2 + V_i, \quad \text{where} \\ \zeta_1 &= \zeta_1^* + \pi\zeta_2 \quad \text{and} \quad U_i = U_i^* + \pi V_i. \end{aligned} \tag{2.13}$$

The variables (X_i, Z_i) are independent of the errors (U_i, V_i), and the errors (U_i, V_i) have a joint normal distribution with mean zero and covariance matrix Σ_{uv}, where

$$\Sigma_{uv} = \begin{pmatrix} 1 & \rho\sigma_v \\ \rho\sigma_v & \sigma_v^2 \end{pmatrix}. \tag{2.14}$$

The parameter of interest is $\theta = (\beta, \zeta, \pi)$, where $\zeta = (\zeta_1, \zeta_2)$.

In this model, weak identification of π occurs when β is close to 0. We analyze a GMM estimator of θ, and corresponding tests concerning functions of θ, in the presence of weak identification or lack of identification.

Let $L(\cdot)$ denote the distribution function of the standard normal distribution. Let $L'(x)$ and $L''(x)$ denote the first- and second-order derivatives of $L(x)$ with respect to x. We use the abbreviations

$$L_i(\theta) = L(Z_i'\beta\pi + X_i'\zeta_1),\ L_i'(\theta) = L'(Z_i'\beta\pi + X_i'\zeta_1), \quad \text{and}$$
$$L_i''(\theta) = L''(Z_i'\beta\pi + X_i'\zeta_1). \tag{2.15}$$

Now we specify the moment conditions for the GMM estimator. The log-likelihood function based on the first reduced-form equation in (2.13) and $y_i = 1(y_i^* > 0)$ is

$$\ell(\theta) = \sum_{i=1}^{n} \left[y_i \log(L_i(\theta)) + (1 - y_i) \log(1 - L_i(\theta)) \right]. \tag{2.16}$$

Let $a = \beta\pi$ and $a_0 = \beta_0\pi_0$. The log-likelihood function $\ell(\theta)$ depends on θ only through a and ζ_1. The expectation of the score function with respect to (a, ζ_1) yields the first set of moment conditions

$$\mathbb{E}_{\gamma_0} w_{1,i}(\theta_0)(y_i - L_i(\theta_0))\overline{Z}_i = 0, \quad \text{where}$$
$$w_{1,i}(\theta) = \frac{L_i'(\theta)}{L_i(\theta)(1 - L_i(\theta))} \quad \text{and}$$
$$\overline{Z}_i = (X_i, Z_i) \in R^{d_X + d_Z}. \tag{2.17}$$

The second reduced-form equation in (2.13) implies

$$\mathbb{E}_{\gamma_0} V_i(\theta_0)\overline{Z}_i = 0, \quad \text{where } V_i(\theta) = Y_i - Z_i'\beta - X_i'\zeta_2. \tag{2.18}$$

We consider a two-step GMM estimator of θ based on the moment conditions in (2.17) and (2.18). The resulting estimator has not appeared in the literature previously, but it is close to estimators in the papers referenced earlier; e.g., see Rivers and Vuong (1988). The GMM sample criterion function is

$$Q_n(\theta) = \overline{g}_n(\theta)'\mathcal{W}_n\overline{g}_n(\theta)/2, \quad \text{where}$$
$$\overline{g}_n(\theta) = n^{-1}\sum_{i=1}^{n} e_i(\theta) \otimes \overline{Z}_i \in R^{2(d_X + d_Z)} \quad \text{and} \quad e_i(\theta) = \begin{pmatrix} w_{1,i}(\theta)(y_i - L_i(\theta)) \\ Y_i - Z_i'\beta - X_i'\zeta_2 \end{pmatrix}. \tag{2.19}$$

In the first step, the weight matrix $\mathcal{W}_n$ is the identity matrix, yielding an estimator $\overline{\theta}_n$. In the second step, $\mathcal{W}_n$ is the optimal weight matrix that takes the form

$$\mathcal{W}_n = \mathcal{W}_n(\overline{\theta}_n), \quad \text{where } \mathcal{W}_n(\theta) = n^{-1}\sum_{i=1}^{n}\left(e_i(\theta)e_i(\theta)'\right)\otimes(\overline{Z}_i\overline{Z}_i'). \tag{2.20}$$

The optimization and true parameter spaces Θ and Θ^* are $\Theta = \mathrm{X}_{j=1}^{k}[-b_{L,j}, b_{H,j}] \times \mathcal{Z} \times \Pi$ and $\Theta^* = \mathrm{X}_{j=1}^{k}[-b_{L,j}^*, b_{H,j}^*] \times \mathcal{Z}^* \times \Pi^*$, where $b_{L,j}, b_{H,j}, b_{L,j}^*, b_{H,j}^* \in R$, $0 \le b_{L,j}^* < b_{L,j}$, $0 \le b_{H,j}^* < b_{H,j}$, $b_{L,j}^*, b_{H,j}^*$ are not both 0, for $j = 1, ..., k$, $\mathcal{Z}^* \subset int(\mathcal{Z}) \subset R^{2d_X}$, $\Pi^* \subset int(\Pi) \subset R$, $\mathcal{Z}^*, \mathcal{Z}, \Pi^*$, and Π are compact.[5]

Define $\overline{w}_{1,i} = \sup_{\theta\in\Theta}|w_{1,i}(\theta)|$ and $\overline{w}_{2,i} = \sup_{\theta\in\Theta}|w_{2,i}(\theta)|$, where $w_{2,i}(\theta) = L_i''(\theta)/(L_i(\theta)(1-L_i(\theta)))$.

The nuisance parameter ϕ is defined by $\phi = (\rho, \sigma_v, F) \in \Phi^*$, where F is the distribution of (X_i, Z_i) and Φ^* is a compact metric space with a metric d_Φ that induces weak convergence of (X_i, Z_i). We use P_ϕ and $\mathbb{E}_\phi$ to denote probability and expectation under ϕ, respectively, for random quantities that depend only on (X_i, Z_i). For any $\theta^* \in \Theta^*$, the true parameter space for ϕ is

$$\Phi(\theta^*) = \Big\{\phi = (\rho, \sigma_v, F) \in \Phi : |\rho| < 1, \sigma_v \ge \varepsilon,\ P_\phi(\overline{Z}_i'c = 0) < 1 \quad \text{for any} \quad c \neq 0,$$
$$\mathbb{E}_\phi\left(||\overline{Z}_i||^{4+\varepsilon} + \overline{w}_{1,i}^{4+\varepsilon} + \overline{w}_{2,i}^{2+\varepsilon}\right) \le C\Big\}, \tag{2.21}$$

for some $C < \infty$ and $\varepsilon > 0$. Note that in this example, $\Phi(\theta^*)$ does not depend on θ^*.

The verification of the assumptions of this paper for this example is given in Supplemental Appendix A.

2.4. Confidence Sets and Tests

We return now to the general framework. We are interested in the effect of lack of identification or weak identification on the GMM estimator $\widehat{\theta}_n$. Also, we are interested in its effects on CSs 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. A nominal $1-\alpha$ CS for $r(\theta)$ is

$$CS_n = \{v : \mathcal{T}_n(v) \le c_{n,1-\alpha}(v)\}, \tag{2.22}$$

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 $H_0 : r(\theta) = v$. The critical values considered in this paper may depend on the null value v of $r(\theta)$ and also on the data. The coverage probability of a CS for $r(\theta)$ is

$$P_\gamma(r(\theta) \in CS_n) = P_\gamma(\mathcal{T}_n(r(\theta)) \le c_{n,1-\alpha}(r(\theta))), \tag{2.23}$$

where $P_\gamma(\cdot)$ denotes probability when γ is the true value.

We are interested in the finite-sample size of the CS, which is the smallest finite-sample coverage probability of the CS over the parameter space. It is approximated by the asymptotic size, which is defined to be

$$AsySz = \liminf_{n\to\infty} \inf_{\gamma\in\Gamma} P_\gamma\left(r(\theta)\in CS_n\right). \tag{2.24}$$

For a test, we are interested in its null rejection probabilities and in particular its maximum null rejection probability, which is the size of the test. A test's asymptotic size is an approximation to the latter. The null rejection probabilities and asymptotic size of a test are given by

$$P_\gamma\left(\mathcal{T}_n(v) > c_{n,1-\alpha}(v)\right) \quad \text{for } \gamma=(\theta,\phi)\in\Gamma \quad \text{with } r(\theta)=v \quad \text{and}$$
$$AsySz = \limsup_{n\to\infty} \sup_{\gamma\in\Gamma: r(\theta)=v} P_\gamma\left(\mathcal{T}_n(v) > c_{n,1-\alpha}(v)\right). \tag{2.25}$$

2.5. Drifting Sequences of Distributions

To determine the asymptotic size of a CS or test, 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. The reason is that the value of γ at which the finite-sample size of a CS or test is attained may vary with the sample size. Similarly, to investigate the finite-sample behavior of the GMM estimator under weak identification, we need to consider its asymptotic behavior under drifting sequences of true distributions—as in Stock and Wright (2000).

Results in Andrews and Guggenberger (2009, 2010) and Andrews, Cheng, and Guggenberger (2009) show that the asymptotic sizes of CSs and tests are determined by certain drifting sequences of distributions. In this paper, the following sequences $\{\gamma_n\}$ are key:

$$\begin{aligned}
\Gamma(\gamma_0) &= \{\{\gamma_n\in\Gamma : n\geq 1\} : \gamma_n\to\gamma_0\in\Gamma\},\\
\Gamma(\gamma_0,0,b) &= \left\{\{\gamma_n\}\in\Gamma(\gamma_0) : \beta_0=0 \quad \text{and} \quad n^{1/2}\beta_n\to b\in(R\cup\{\pm\infty\})^{d_\beta}\right\}, \quad \text{and}\\
\Gamma(\gamma_0,\infty,\omega_0) &= \left\{\{\gamma_n\}\in\Gamma(\gamma_0) : n^{1/2}\|\beta_n\|\to\infty \quad \text{and} \quad \beta_n/\|\beta_n\|\to\omega_0\in R^{d_\beta}\right\},
\end{aligned} \tag{2.26}$$

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

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 following terminology: "under $\{\gamma_n\}\in\Gamma(\gamma_0)$" means "when the true parameters are $\{\gamma_n\}\in\Gamma(\gamma_0)$ for any $\gamma_0\in\Gamma$"; "under $\{\gamma_n\}\in\Gamma(\gamma_0,0,b)$" means "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 $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$" means "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 $||\omega_0|| = 1$."

3. ASSUMPTIONS

This section provides relatively primitive sufficient conditions for GMM estimators.

3.1. Assumption GMM1

The first assumption specifies the basic identification problem. It also provides conditions that are used to determine the probability limit of the GMM estimator, when it exists, under all categories of drifting sequences of distributions.

Assumption GMM1.

(i) If $\beta = 0$, $\overline{g}_n(\theta)$ and $\mathcal{W}_n(\theta)$ do not depend on π, $\forall \theta \in \Theta$, $\forall n \geq 1$, for any true parameter $\gamma^* \in \Gamma$.

(ii) Under $\{\gamma_n\} \in \Gamma(\gamma_0)$, $\sup_{\theta \in \Theta} ||\overline{g}_n(\theta) - g_0(\theta; \gamma_0)|| \rightarrow_p 0$ and $\sup_{\theta \in \Theta} ||\mathcal{W}_n(\theta) - \mathcal{W}(\theta; \gamma_0)|| \rightarrow_p 0$ for some nonrandom functions $g_0(\theta; \gamma_0) : \Theta \times \Gamma \rightarrow R^k$ and $\mathcal{W}(\theta; \gamma_0) : \Theta \times \Gamma \rightarrow R^{k \times k}$.

(iii) When $\beta_0 = 0$, $g_0(\psi, \pi; \gamma_0) = 0$ if and only if $\psi = \psi_0$, $\forall \pi \in \Pi$, $\forall \gamma_0 \in \Gamma$.

(iv) When $\beta_0 \neq 0$, $g_0(\theta; \gamma_0) = 0$ if and only if $\theta = \theta_0$, $\forall \gamma_0 \in \Gamma$.

(v) $g_0(\theta; \gamma_0)$ is continuously differentiable in θ on Θ, with its partial derivatives with respect to θ and ψ denoted by $g_\theta(\theta; \gamma_0) \in R^{k \times d_\theta}$ and $g_\psi(\theta; \gamma_0) \in R^{k \times d_\psi}$, respectively.

(vi) $\mathcal{W}(\theta; \gamma_0)$ is continuous in θ on Θ $\forall \gamma_0 \in \Gamma$.

(vii) $0 < \lambda_{\min}(\mathcal{W}(\psi_0, \pi; \gamma_0)) \leq \lambda_{\max}(\mathcal{W}(\psi_0, \pi; \gamma_0)) < \infty$, $\forall \pi \in \Pi$, $\forall \gamma_0 \in \Gamma$.

(viii) $\lambda_{\min}(g_\psi(\psi_0, \pi; \gamma_0)' \mathcal{W}(\psi_0, \pi; \gamma_0) g_\psi(\psi_0, \pi; \gamma_0)) > 0$, $\forall \pi \in \Pi$, $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$.

(ix) $\Psi(\pi)$ is compact $\forall \pi \in \Pi$, and Π and Θ are compact.

(x) $\forall \varepsilon > 0$, $\exists \delta > 0$ such that $d_H(\Psi(\pi_1), \Psi(\pi_2)) < \varepsilon$ $\forall \pi_1, \pi_2 \in \Pi$ with $\|\pi_1 - \pi_2\| < \delta$, where $d_H(\cdot)$ is the Hausdorff metric.

Assumption GMM1(i) is the key condition that concerns the lack of identification (by the moment functions) when $\beta = 0$. Assumptions GMM1(ii)–(x) are mostly fairly standard GMM regularity conditions, but with some adjustments due to the lack of identification of π when $\beta = 0$, e.g., see Assumption GMM1(iii). Note that Assumption GMM1(viii) involves the derivative matrix of $g_0(\theta; \gamma_0)$ with respect to ψ only, not $\theta = (\psi, \pi)$. In consequence, this assumption is not restrictive.

The weight matrix $\mathcal{W}_n(\theta)$ depends on θ only when a continuous updating GMM estimator is considered. For a two-step estimator, $\mathcal{W}_n(\theta)$ depends on a preliminary estimator $\overline{\theta}_n$ but does not depend on θ. Let $\mathcal{W}_n(\overline{\theta}_n)$ be the weight matrix for a two-step estimator. (This is a slight abuse of notation because in (2.1) $\mathcal{W}_n(\theta)$ and $\overline{g}_n(\theta)$ are indexed by the same θ, whereas here they are different.)

For the weight matrix of a two-step estimator to satisfy Assumption GMM1(ii), we need

$$\mathcal{W}_n\left(\overline{\theta}_n\right) \to_p \mathcal{W}(\theta_0; \gamma_0) \tag{3.1}$$

for some nonrandom matrix $\mathcal{W}(\theta_0; \gamma_0)$ under $\{\gamma_n\} \in \Gamma(\gamma_0)$. This is not an innocuous assumption in the weak identification scenario because the preliminary estimator $\overline{\theta}_n$ may be inconsistent. Lemma 3.1 shows that (3.1) holds despite the inconsistency of $\overline{\pi}_n$ that occurs under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$, where $\overline{\theta}_n = (\overline{\psi}_n, \overline{\pi}_n)$.

LEMMA 3.1. *Suppose* $\overline{\theta}_n = (\overline{\psi}_n, \overline{\pi}_n)$ *is an estimator of* θ *such that*

(a) $\overline{\theta}_n \to_p \theta_0$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0)$, $\forall \gamma_0 \in \Gamma$ *with* $\beta_0 \neq 0$,
(b) $\overline{\psi}_n \to_p \psi_0$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0)$, $\forall \gamma_0 \in \Gamma$ *with* $\beta_0 = 0$,
(c) $\mathcal{W}_n(\theta)$ *satisfies Assumptions GMM1(i), GMM1(ii), and GMM1(vi), and*
(d) Π *is compact.*
Then, $\mathcal{W}_n(\overline{\theta}_n) \to_p \mathcal{W}(\theta_0; \gamma_0)$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0)\ \forall \gamma_0 \in \Gamma$.

Remarks.

1. Lemma 3.1 allows for inconsistency of $\overline{\pi}_n$, i.e., $\overline{\pi}_n - \pi_n \neq o_p(1)$, under $\{\gamma_n\} \in \Gamma(\gamma_0)$ with $\beta_0 = 0$. Inconsistency occurs under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$; see Theorem 4.1(a) in Section 4.
2. Typically, the preliminary estimator $\overline{\theta}_n$ is obtained by minimizing $Q_n(\theta)$ in (2.1) with a weight matrix $\mathcal{W}_n(\theta)$ that does not depend on θ or any estimator of θ. In such cases, the properties of $\overline{\theta}_n$ assumed in Lemma 3.1 hold provided Assumption GMM1 holds with the specified weight matrix.[6]

Example 1 (cont.)
For this example, the key quantities in Assumption GMM1 are

$$\begin{aligned}
&g_0(\theta; \gamma_0) = \mathbb{E}_{\phi_0}(\beta_0 h(X_{1,i}, \pi_0) - \beta h(X_{1,i}, \pi) + X_{2,i}'(\zeta_0 - \zeta))Z_i,\\
&\mathcal{W}(\theta; \gamma_0) = \mathcal{W}(\gamma_0) = \left(\mathbb{E}_{\phi_0} Z_i Z_i'\right)^{-1},\\
&g_\psi(\theta; \gamma_0) = -\mathbb{E}_{\phi_0} Z_i d_{\psi,i}(\pi)', \quad \text{and} \quad g_\theta(\theta; \gamma_0) = -\mathbb{E}_{\phi_0} Z_i d_{\theta,i}(\pi)', \quad \text{where}\\
&\quad d_{\psi,i}(\pi) = (h(X_{1,i}, \pi), X_{2,i}) \in R^{d_{X_2}+1} \quad \text{and}\\
&\quad d_{\theta,i}(\pi) = (h(X_{1,i}, \pi), X_{2,i}, \beta h_\pi(X_{1,i}, \pi)) \in R^{d_{X_2}+d_\pi+1}.
\end{aligned} \tag{3.2}$$

Assumption GMM1(i) holds by the form of $\overline{g}_n(\theta)$ and $\mathcal{W}_n$ in (2.8) and the fact that $U_i(\theta)$ does not depend on π when $\beta = 0$. Assumption GMM1(ii) holds by the uniform law of large numbers (LLN) given in Lemma 12.1 in Supplemental Appendix D under the conditions in (2.11).

To verify Assumption GMM1(iii), we write

$$g_0(\psi,\pi;\gamma_0) - g_0(\psi_0,\pi;\gamma_0) = \mathbb{E}_{\phi_0}(-\beta h(X_{1,i},\pi) + X_{2,i}'(\zeta_0-\zeta))Z_i = \left[\mathbb{E}_{\phi_0} Z_i d_{\psi,i}(\pi)'\right]\Delta, \quad \textbf{(3.3)}$$

where $\Delta = (-\beta, \zeta_0 - \zeta) \in R^{d_{X_2}+1}$. We need to show that when $\beta_0 = 0$ the quantity in (3.3) does not equal zero $\forall \psi \neq \psi_0$ and $\forall \pi \in \Pi$. This holds because $d_{\psi,i}(\pi)$ is a subvector of $d^*_{\psi,i}(\pi_1,\pi_2)$ and $\mathbb{E}_\phi Z_i d^*_{\psi,i}(\pi_1,\pi_2)'$ has full column rank $\forall \pi_1,\pi_2 \in \Pi$ with $\pi_1 \neq \pi_2$ by (2.11).

To verify Assumption GMM1(iv), we write

$$g_0(\theta;\gamma_0) - g_0(\theta_0;\gamma_0) = \mathbb{E}_{\phi_0}(\beta_0 h(X_{1,i},\pi_0) - \beta h(X_{1,i},\pi) + X_{2,i}'(\zeta_0-\zeta))Z_i = \left[\mathbb{E}_{\phi_0} Z_i d^*_{\psi,i}(\pi_0,\pi)'\right]c, \quad \textbf{(3.4)}$$

where $c = (\beta_0, -\beta, \zeta_0 - \zeta) \in R^{d_{X_2}+2}$. We need to show that when $\beta_0 \neq 0$ the quantity in (3.4) does not equal zero when $\theta \neq \theta_0$. This holds when $\pi \neq \pi_0$ because $\mathbb{E}_{\phi_0} Z_i d^*_{\psi,i}(\pi_0,\pi)'$ has full column rank for $\pi \neq \pi_0$ by (2.11). When $\pi = \pi_0$,

$$g_0(\theta;\gamma_0) - g_0(\theta_0;\gamma_0) = g_0(\psi,\pi_0;\gamma_0) - g_0(\psi_0,\pi_0;\gamma_0) = \left[\mathbb{E}_{\phi_0} Z_i d_{\psi,i}(\pi_0)'\right]\Delta_1, \quad \textbf{(3.5)}$$

where $\Delta_1 = (\beta_0 - \beta, \zeta_0 - \zeta) \in R^{d_{X_2}+1}$. The quantity in (3.5) does not equal zero for $\psi \neq \psi_0$ because $\mathbb{E}_{\phi_0} Z_i d_{\psi,i}(\pi_0)'$ has full column rank. This completes the verification of Assumption GMM1(iv).

Assumption GMM1(v) holds by the assumption that $h(x,\pi)$ is twice continuously differentiable with respect to π and the moment conditions in (2.11). Assumption GMM1(vi) holds automatically because $\mathcal{W}(\theta;\gamma_0) = (\mathbb{E}_{\phi_0} Z_i Z_i')^{-1}$ does not depend on θ. Assumption GMM1(vii) holds because $\mathbb{E}_{\phi_0} Z_i Z_i' \in R^{k\times k}$ is positive definite $\forall \gamma_0 \in \Gamma$. Assumption GMM1(viii) holds because $\mathcal{W}(\psi_0,\pi;\gamma_0) = \mathbb{E}_{\phi_0} Z_i Z_i'$ is positive definite and $g_\psi(\psi_0,\pi;\gamma_0)$ has full rank by the conditions in (2.11). Assumption GMM1(ix) holds because $\Theta = \mathcal{B} \times \mathcal{Z} \times \Pi$, and $\mathcal{B}$, $\mathcal{Z}$, Π, and $\Psi = \mathcal{B} \times \mathcal{Z}$ are all compact. Assumption GMM1(x) holds automatically because Ψ does not depend on π.

For brevity, the verifications of Assumptions GMM1 and GMM2–GMM5, which follow, for the probit model with endogeneity are given in Supplemental Appendix A.

3.2. Assumption GMM2

The next assumption, Assumption GMM2, is used when verifying that the GMM criterion function satisfies a quadratic approximation with respect to ψ when $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ and with respect to θ when $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$. In the former case, the expansion is around the value

$$\psi_{0,n} = (0, \zeta_n), \quad \textbf{(3.6)}$$

rather than around the true value $\psi_n = (\beta_n, \zeta_n)$. The reason for expanding around $\psi_{0,n}$ is that the first term in the expansion of $Q_n(\psi, \pi)$ does not depend on π when $\psi = \psi_{0,n}$ by Assumption GMM1(i).

Under $\{\gamma_n\} \in \Gamma(\gamma_0)$, define the centered sample moment conditions by

$$\widetilde{g}_n(\theta; \gamma_0) = \overline{g}_n(\theta) - g_0(\theta; \gamma_0). \tag{3.7}$$

We define a matrix $B(\beta)$ that is used to normalize the (generalized) first-derivative matrix of the sample moments $\overline{g}_n(\theta)$ so that it is full rank asymptotically. Let $B(\beta)$ be the $d_\theta \times d_\theta$ diagonal matrix defined by

$$B(\beta) = \text{Diag}\left\{1'_{d_\psi}, \iota(\beta)1'_{d_\pi}\right\}, \tag{3.8}$$

where $\iota(\beta) = \beta$ if β is a scalar and $\iota(\beta) = ||\beta||$ if β is a vector.[7]

Assumption GMM2.

(i) Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, $\sup_{\psi \in \Psi(\pi): ||\psi - \psi_{0,n}|| \le \delta_n} ||\widetilde{g}_n(\psi, \pi; \gamma_0) - \widetilde{g}_n(\psi_{0,n}, \pi; \gamma_0)|| / (n^{-1/2} + ||\psi - \psi_{0,n}||) = o_{p\pi}(1)$ for all constants $\delta_n \to 0$.

(ii) Under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $\sup_{\theta \in \Theta_n(\delta_n)} ||\widetilde{g}_n(\theta; \gamma_0) - \widetilde{g}_n(\theta_n; \gamma_0)|| / (n^{-1/2} + ||B(\beta_n)(\theta - \theta_n)||) = o_p(1)$ for all constants $\delta_n \to 0$, where $\Theta_n(\delta_n) = \{\theta \in \Theta : \|\psi - \psi_n\| \le \delta_n \|\beta_n\| \text{ and } \|\pi - \pi_n\| \le \delta_n\}$.

When $\overline{g}_n(\theta)$ is continuously differentiable in θ, Assumption GMM2 is easy to verify. In this case, Assumption GMM2*, which follows, is a set of sufficient conditions for Assumption GMM2.

Assumption GMM2 allows for nonsmooth sample moment conditions. It is analogous to Assumption GMM2(iv) of Andrews (2002), which in turn is shown to be equivalent to condition (iii) of Theorem 3.3 of Pakes and Pollard (1989). In contrast to these conditions in the literature, Assumption GMM2 applies under drifting sequences of true parameters and provides conditions that allow for weak identification. Nevertheless, Assumption GMM2 can be verified by methods used in Pakes and Pollard (1989) and Andrews (2002).

Assumption GMM2*.

(i) $\overline{g}_n(\theta)$ is continuously differentiable in θ on Θ $\forall n \ge 1$.

(ii) Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, $\sup_{\theta \in \Theta: ||\psi - \psi_{0,n}|| \le \delta_n} \left\|(\partial/\partial\psi')\overline{g}_n(\theta) - g_\psi(\theta; \gamma_0)\right\| = o_p(1)$ for all constants $\delta_n \to 0$.

(iii) Under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $\sup_{\theta \in \Theta_n(\delta_n)} \left\|\left((\partial/\partial\theta')\overline{g}_n(\theta) - g_\theta(\theta; \gamma_0)\right) B^{-1}(\beta_n)\right\| = o_p(1)$ for all constants $\delta_n \to 0$.

When $\overline{g}_n(\theta)$ takes the form of a sample average, Assumption GMM2* can be verified by a uniform LLN and the switch of $\mathbb{E}$ and ∂ under some regularity conditions.

LEMMA 3.2. *Assumption GMM2* implies Assumption GMM2.*

Example 1 (cont.)
We verify Assumption GMM2 in this example using the sufficient condition Assumption GMM2*. The key quantities in Assumption GMM2* are

$$\frac{\partial}{\partial \psi'}\overline{g}_n(\theta) = n^{-1}\sum_{i=1}^{n} Z_i d_{\psi,i}(\pi)' \quad \text{and} \quad \frac{\partial}{\partial \theta'}\overline{g}_n(\theta) = n^{-1}\sum_{i=1}^{n} Z_i d_{\theta,i}(\pi)'. \tag{3.9}$$

Assumption GMM2*(i) holds with the partial derivatives given in (3.9). Assumption GMM2*(ii) holds by the uniform LLN given in Lemma 12.1 in Supplemental Appendix D under the conditions in (2.11). Assumption GMM2*(iii) holds by this uniform LLN and $\beta/\beta_n = 1 + o(1)$ for $\theta \in \Theta_n(\delta_n)$. □

3.3. Assumption GMM3

Under Assumptions GMM1 and GMM2, Assumption GMM3 which follows, is used when establishing the asymptotic distribution of the GMM estimator under weak and semistrong identification, i.e., when $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$.

Define the $k \times d_\beta$ matrix of partial derivatives of the average population moment function with respect to the true β value, β^*, to be

$$K_{n,g}(\theta;\gamma^*) = n^{-1}\sum_{i=1}^{n} \frac{\partial}{\partial \beta^{*\prime}} \mathbb{E}_{\gamma^*} g(W_i,\theta), \tag{3.10}$$

where $\gamma^* = (\beta^*, \zeta^*, \pi^*, \phi^*)$. The domain of the function $K_{n,g}(\theta;\gamma^*)$ is $\Theta_\delta \times \Gamma_0$, where $\Theta_\delta = \{\theta \in \Theta : ||\beta|| < \delta\}$ and $\Gamma_0 = \{\gamma_a = (a\beta, \zeta, \pi, \phi) \in \Gamma : \gamma = (\beta, \zeta, \pi, \phi) \in \Gamma$ with $||\beta|| < \delta$ and $a \in [0,1]\}$ for some $\delta > 0$.[8]

Assumption GMM3.

(i) $\overline{g}_n(\theta)$ takes the form $\overline{g}_n(\theta) = n^{-1}\sum_{i=1}^{n} g(W_i,\theta)$ for some function $g(W_i,\theta) \in R^k$ $\forall \theta \in \Theta$.

(ii) $\mathbb{E}_{\gamma^*} g(W_i, \psi^*, \pi) = 0$ $\forall \pi \in \Pi$, $\forall i \geq 1$ when the true parameter is γ^* $\forall \gamma^* = (\psi^*, \pi^*, \phi^*) \in \Gamma$ with $\beta^* = 0$.

(iii) Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, $n^{-1/2}\sum_{i=1}^{n}(g(W_i, \psi_{0,n}, \pi_n) - \mathbb{E}_{\gamma_n} g(W_i, \psi_{0,n}, \pi_n)) \rightarrow_d N(0, \Omega_g(\gamma_0))$ for some k by k matrix $\Omega_g(\gamma_0)$.

(iv) (a) $K_{n,g}(\theta;\gamma^*)$ exists $\forall(\theta,\gamma^*) \in \Theta_\delta \times \Gamma_0$, $\forall n \geq 1$. (b) For some nonstochastic $k \times d_\beta$ matrix-valued function $K_g(\psi_0, \pi; \gamma_0)$, $K_{n,g}(\psi_n, \pi; \widetilde{\gamma}_n) \rightarrow K_g(\psi_0, \pi; \gamma_0)$ uniformly over $\pi \in \Pi$ for all nonstochastic sequences $\{\psi_n\}$ and $\{\widetilde{\gamma}_n\}$ such that $\widetilde{\gamma}_n \in \Gamma$, $\widetilde{\gamma}_n \rightarrow \gamma_0 = (0, \zeta_0, \pi_0, \phi_0)$ for some $\gamma_0 \in \Gamma$, $(\psi_n, \pi) \in \Theta$, and $\psi_n \rightarrow \psi_0 = (0, \zeta_0)$. (c) $K_g(\psi_0, \pi; \gamma_0)$ is continuous on Π $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$.

(v) $\forall \omega_0 \in R^{d_\beta}$ with $||\omega_0|| = 1$, $K_g(\psi_0, \pi; \gamma_0)\omega_0 = g_\psi(\psi_0, \pi; \gamma_0)S$ for some $S \in R^{d_\psi}$ if and only if $\pi = \pi_0$.

(vi) Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, $n^{-1}\sum_{i=1}^{n}(\partial/\partial\psi')\mathbb{E}_{\gamma_n} g(W_i, \psi, \pi)|_{(\psi,\pi)=\theta_n} \to g_\psi(\theta_0; \gamma_0)$.

Assumption GMM3(iii) can be verified using a triangular array central limit theorem (CLT). Although Assumption GMM3(iv) is somewhat complicated, it is not restrictive; see the verification of it in the two examples. A set of primitive sufficient conditions for Assumption GMM3(iv) is given in Appendix A of AC1-SM.[9]

In Assumption GMM3(v), the equality holds for $\pi = \pi_0$ with $S = -[I_{d_\beta} : 0_{d_\beta \times d_\zeta}]'\omega_0$ by Lemma 9.3 in AC1-SM under the assumptions therein. For any $\pi \neq \pi_0$, Assumption GMM3(v) requires that any linear combination of the columns of $K_g(\psi_0, \pi; \gamma_0)$ cannot be in the column space of $g_\psi(\psi_0, \pi; \gamma_0)$.

With identically distributed observations, Assumption GMM3(vi) can be verified by the exchange of $\mathbb{E}$ and ∂ under suitable regularity conditions.

Example 1 (cont.)
For this example, the key quantities in Assumption GMM3 are

$$\begin{aligned} g(W_i,\theta) &= (Y_i - \beta h(X_{1,i},\pi) - X_{2,i}'\zeta)Z_i, \\ \Omega_g(\gamma_0) &= \mathbb{E}_{\phi_0} U_i^2 Z_i Z_i', \quad \text{and} \\ K_{g,n}(\theta,\gamma^*) &= K_g(\theta,\gamma^*) = \mathbb{E}_{\phi^*} h(X_{1,i},\pi^*)Z_i. \end{aligned} \tag{3.11}$$

Assumption GMM3(i) holds with $g(W_i,\theta)$ in (3.11). To verify Assumption GMM3(ii), we have

$$\mathbb{E}_{\phi^*} g(W_i,\theta) = \mathbb{E}_{\phi^*}\left(U_i + \beta^* h(X_{1,i},\pi^*) - \beta h(X_{1,i},\pi) + X_{2,i}'(\zeta^* - \zeta)\right) Z_i. \tag{3.12}$$

When $\beta = \beta^* = 0$ and $\zeta = \zeta^*$, $\mathbb{E}_{\phi^*} g(W_i,\theta) = 0 \; \forall \pi \in \Pi$.

Next, we show that Assumption GMM3(iii) holds with $\Omega_g(\gamma_0)$ in (3.11). Define

$$\begin{aligned} G_{g,n}(\pi_n) &= n^{-1/2}\sum_{i=1}^{n}\left(g(W_i,\psi_{0,n},\pi_n) - \mathbb{E}_{\phi_n} g(W_i,\psi_{0,n},\pi_n)\right) \\ &= n^{-1/2}\sum_{i=1}^{n} U_i Z_i + \beta_n\left[n^{-1/2}\sum_{i=1}^{n}\left(h(X_i,\pi_n)Z_i - \mathbb{E}_{\phi_n} h(X_i,\pi_n)Z_i\right)\right]. \end{aligned} \tag{3.13}$$

By the CLT for triangular arrays of rowwise i.i.d. random variables given in Lemma 12.3 in Supplemental Appendix C, $n^{-1/2}\sum_{i=1}^{n} U_i Z_i \to_d N(0, \Omega_g(\gamma_0))$. The second term on the right-hand side of the second equality in (3.13) is $o_p(1)$ because $\beta_n \to 0$ and $n^{-1/2}\sum_{i=1}^{n}(h(X_i,\pi_n)Z_i - \mathbb{E}_{\phi_n} h(X_i,\pi_n)Z_i) = O_p(1)$ by the CLT in Lemma 12.3 in Supplemental Appendix C. Hence, $G_{g,n}(\pi_n) \to_d N(0, \Omega_g(\gamma_0))$.

Next, we show that Assumption GMM3(iv) holds with $K_{g,n}(\theta,\gamma^*)$ and $K_g(\theta,\gamma^*)$ in (3.11). Assumption GMM3(iv)(a) is implied by (3.12) and the moment conditions in (2.11). The convergence in Assumption GMM3(iv)(b) holds because $\phi_n \to \phi_0$ induces weak convergence of (X_i, Z_i) by the definition of the metric on Φ^* and $\mathbb{E}_\phi \sup_{\pi\in\Pi} ||h(X_{1,i},\pi)Z_i||^{1+\delta} \leq C$ for some $\delta > 0$ and $C < \infty$ by the conditions in (2.11). The convergence holds uniformly over $\pi \in \Pi$ by Lemma 12.1 in Supplemental Appendix D because Π is compact and $\mathbb{E}_{\phi^*} \sup_{\pi\in\Pi} ||h_\pi(X_{1,i},\pi)|| \cdot ||Z_i|| \leq C$ for some $C < \infty$. Assumption GMM3(iv)(c) holds because Π is compact, $h(x,\pi)$ is continuous in π, and $\mathbb{E}_{\phi^*} \sup_{\pi\in\Pi} ||h(X_{1,i},\pi)|| \cdot ||Z_i|| \leq C$ for some $C < \infty$ by the conditions in (2.11). This completes the verification of Assumption GMM3(iv).

To verify Assumption GMM3(v), note that for $S \in R^{d_{X_2}+1}$ we have

$$\begin{aligned} &K_g(\psi_0,\pi;\gamma_0)\omega_0 - g_\psi(\psi_0,\pi;\gamma_0)S \\ &\quad = \mathbb{E}_{\phi_0} Z_i h(X_{1,i},\pi_0)\omega_0 + \mathbb{E}_{\phi_0} Z_i d_{\psi,i}(\pi)'S \\ &\quad = \mathbb{E}_{\phi_0} Z_i d^*_{\psi,i}(\pi_0,\pi)'\Delta_2, \quad \text{where} \quad \Delta_2 = (\omega_0, S) \neq 0_{d_\zeta+2}. \end{aligned} \tag{3.14}$$

Because $\mathbb{E}_{\phi_0} Z_i d^*_{\psi,i}(\pi_0,\pi)'$ has full column rank for all $\pi \neq \pi_0$ by (2.11), $K_g(\psi_0,\pi;\gamma_0)\omega_0 \neq g_\psi(\psi_0,\pi;\gamma_0)S$ for any $\pi \neq \pi_0$. When $\pi = \pi_0$, $K_g(\psi_0,\pi;\gamma_0)\omega_0 = g_\psi(\psi_0,\pi;\gamma_0)S$ if $S = (-\omega_0, 0_{d_\zeta})$ $(\in R^{d_\zeta+1})$. This completes the verification of Assumption GMM3 for this example.

3.4. Assumption GMM4

To obtain the asymptotic distribution of $\widehat{\pi}_n$ when $\beta_n = O(n^{-1/2})$ via the continuous mapping theorem, we use Assumption GMM4 stated subsequently.

Under Assumptions GMM1(i) and GMM1(ii), $\mathcal{W}(\psi_0,\pi;\gamma_0)$ does not depend on π when $\beta_0 = 0$. For simplicity, let $\mathcal{W}(\psi_0;\gamma_0)$ abbreviate $\mathcal{W}(\psi_0,\pi;\gamma_0)$ when $\beta_0 = 0$.

The following quantities arise in the asymptotic distributions of $\widehat{\theta}_n$ and various test statistics when $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ and $||b|| < \infty$. Define

$$\begin{aligned} \Omega(\pi_1,\pi_2;\gamma_0) &= g_\psi(\psi_0,\pi_1;\gamma_0)'\mathcal{W}(\psi_0;\gamma_0)\Omega_g(\gamma_0)\mathcal{W}(\psi_0;\gamma_0)g_\psi(\psi_0,\pi_2;\gamma_0), \\ H(\pi;\gamma_0) &= g_\psi(\psi_0,\pi;\gamma_0)'\mathcal{W}(\psi_0;\gamma_0)g_\psi(\psi_0,\pi;\gamma_0), \quad \text{and} \\ K(\psi_0,\pi;\gamma_0) &= g_\psi(\psi_0,\pi;\gamma_0)'\mathcal{W}(\psi_0;\gamma_0)K_g(\psi_0,\pi;\gamma_0). \end{aligned} \tag{3.15}$$

Let $G(\cdot;\gamma_0)$ denote a mean zero Gaussian process indexed by $\pi \in \Pi$ with bounded continuous sample paths and covariance kernel $\Omega(\pi_1,\pi_2;\gamma_0)$ for $\pi_1,\pi_2 \in \Pi$.

Next, we define a "weighted noncentral chi-square" process $\{\xi(\pi;\gamma_0,b) : \pi \in \Pi\}$ that arises in the asymptotic distributions. Let

$$\xi(\pi;\gamma_0,b) = -\frac{1}{2}(G(\pi;\gamma_0) + K(\pi;\gamma_0)b)'H^{-1}(\pi;\gamma_0)(G(\pi;\gamma_0) + K(\pi;\gamma_0)b). \tag{3.16}$$

Under Assumptions GMM1–GMM3, $\{\xi(\pi;\gamma_0,b) : \pi \in \Pi\}$ has bounded continuous sample paths almost surely (a.s.).

Assumption GMM4. 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 may depend 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 GMM4, $\pi^*(\gamma_0,b)$ is random.

Next, we provide a sufficient condition for Assumption GMM4. We partition $g_\psi(\theta;\gamma_0)\in R^{k\times d_\psi}$ as

$$g_\psi(\theta;\gamma_0)=\left[g_\beta(\theta;\gamma_0):g_\zeta(\theta;\gamma_0)\right], \tag{3.17}$$

where $g_\beta(\theta;\gamma_0)\in R^{k\times d_\beta}$ and $g_\zeta(\theta;\gamma_0)\in R^{k\times d_\zeta}$. When $\beta_0=0$, $g_\zeta(\psi_0,\pi;\gamma_0)$ does not depend on π by Assumptions GMM1(i) and GMM3(ii) and is denoted by $g_\zeta(\psi_0;\gamma_0)$ for simplicity. When $d_\beta=1$ and $\beta_0=0$, define

$$g^*_\psi(\psi_0,\pi_1,\pi_2;\gamma_0)=[g_\beta(\psi_0,\pi_1;\gamma_0):g_\beta(\psi_0,\pi_2;\gamma_0):g_\zeta(\psi_0;\gamma_0)]\in R^{k\times(d_\zeta+2)}. \tag{3.18}$$

Assumption GMM4*.

(i) $d_\beta=1$ (e.g., β is a scalar).

(ii) $g^*_\psi(\psi_0,\pi_1,\pi_2;\gamma_0)$ has full column rank, $\forall\pi_1,\pi_2\in\Pi$ with $\pi_1\neq\pi_2$, $\forall\gamma_0\in\Gamma$ with $\beta_0=0$.

(iii) $\Omega_g(\gamma_0)$ is positive definite, $\forall\gamma_0\in\Gamma$ with $\beta_0=0$.

LEMMA 3.3. *Assumptions GMM1–GMM3 and GMM4* imply Assumption GMM4.*

Example 1 (cont.)
We verify Assumption GMM4 in this example using the sufficient condition Assumption GMM4*. The key quantity in Assumption GMM4* is

$$\begin{aligned} g^*_\psi(\psi_0,\pi_1,\pi_2;\gamma_0) &= -\mathbb{E}_{\phi_0}Z_i(h(X_{1,i},\pi_1),h(X_{1,i},\pi_2),X'_{2,i}) \\ &= -\mathbb{E}_{\phi_0}Z_i d^*_{\psi,i}(\pi_1,\pi_2). \end{aligned} \tag{3.19}$$

Assumption GMM4*(i) holds automatically. Assumption GMM4*(ii) holds because $\mathbb{E}_{\phi_0}Z_i d^*_{\psi,i}(\pi_1,\pi_2)$ has full column rank $\forall\pi_1,\pi_2\in\Pi$ with $\pi_1\neq\pi_2$ by (2.11). Assumption GMM4*(iii) holds with $\Omega_g(\gamma_0)=\mathbb{E}_{\phi_0}U_i^2Z_iZ_i'$ because $\mathbb{E}_{\phi_0}Z_iZ_i'$ is positive definite and $\mathbb{E}(U_i^2|Z_i)>0$ a.s. This completes the verification of Assumption GMM4.

3.5. Assumption GMM5

Under Assumptions GMM1 and GMM2, Assumption GMM5 is used in what follows to establish the asymptotic distribution of the GMM estimator under semistrong and strong identification, i.e., when $\{\gamma_n\}\in\Gamma(\gamma_0,\infty,\omega_0)$.

Assumption GMM5. Under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,

(i) $n^{1/2}\overline{g}_n(\theta_n) \to_d N(0, V_g(\gamma_0))$ for some symmetric and positive definite $d_\theta \times d_\theta$ matrix $V_g(\gamma_0)$,

(ii) for all constants $\delta_n \to 0$, $\sup_{\theta \in \Theta_n(\delta_n)} ||(g_\theta(\theta;\gamma_0) - g_\theta(\theta_n;\gamma_0))B^{-1}(\beta_n)|| = o(1)$, and

(iii) $g_\theta(\theta_n;\gamma_0)B^{-1}(\beta_n) \to J_g(\gamma_0)$ for some matrix $J_g(\gamma_0) \in R^{k \times d_\theta}$ with full column rank.[10]

Now, we define two key quantities that arise in the asymptotic distribution of the estimator $\widehat{\theta}_n$ when $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$. Let

$$V(\gamma_0) = J_g(\gamma_0)'\mathcal{W}(\theta_0;\gamma_0)V_g(\gamma_0)\mathcal{W}(\theta_0;\gamma_0)J_g(\gamma_0) \quad \text{and}$$
$$J(\gamma_0) = J_g(\gamma_0)'\mathcal{W}(\theta_0;\gamma_0)J_g(\gamma_0). \tag{3.20}$$

Let $G^*(\gamma_0) \sim N(0_{d_\theta}, V(\gamma_0))$ for $\gamma_0 \in \Gamma$.

Example 1 (cont.)
The key quantities in Assumption GMM5 for this example are

$$V_g(\gamma_0) = \mathbb{E}_{\phi_0} U_i^2 Z_i Z_i' \quad \text{and} \quad J_g(\gamma_0) = -\mathbb{E}_{\phi_0} Z_i d_i(\pi_0)'. \tag{3.21}$$

Assumption GMM5(i) holds by the CLT for triangular arrays of rowwise i.i.d. random variables given in Lemma 12.3 in Supplemental Appendix C. Assumption GMM5(ii) holds with $g_\theta(\theta;\gamma_0)$ defined as in (3.2) because $\beta_n/\beta = 1 + o(1)$ for $\theta \in \Theta_n(\delta_n)$ and $g_\theta(\theta;\gamma_0)B^{-1}(\beta) = -\mathbb{E}_{\phi_0} Z_i d_i(\pi)'$ is continuous in π uniformly over $\pi \in \Pi$, which in turn holds by the moment conditions in (2.11) and the compactness of Π.

Assumption GMM5(iii) holds because

$$g_\theta(\theta_n;\gamma_n)B^{-1}(\beta_n) = -\mathbb{E}_{\phi_n} Z_i d_i(\pi_n)' \to -\mathbb{E}_{\phi_0} Z_i d_i(\pi_0)', \tag{3.22}$$

where the convergence holds because (i) $\mathbb{E}_{\phi_n} Z_i d_i(\pi)' \to \mathbb{E}_{\phi_0} Z_i d_i(\pi)$ uniformly over $\pi \in \Pi$ by arguments analogous to those used in the verification of Assumption GMM3(iv)(b) and (ii) $\pi_n \to \pi_0$. The matrix $J_g(\gamma_0)$ has full column rank by (2.11). This completes the verification of Assumption GMM5.

3.6. Minimum Distance Estimators

Assumptions GMM1, GMM2, GMM4, and GMM5 apply equally well to the MD estimator as to the GMM estimator. Only Assumption GMM3 does not apply to the MD estimator. In place of part of Assumption GMM3, we employ the following assumption for MD estimators.

Assumption MD. Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, $n^{1/2}\overline{g}_n(\psi_{0,n}, \pi_n) = O_p(1)$.

3.7. Parameter Space Assumptions

Next, we specify conditions on the parameter spaces Θ and Γ.

Define $\Theta^*_\delta = \{\theta \in \Theta^* : ||\beta|| < \delta\}$, where Θ^* is the true parameter space for θ; see (2.3). The optimization parameter space Θ satisfies the following conditions.

Assumption B1.

(i) $int(\Theta) \supset \Theta^*$.

(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 (2.5).

(iii) Π is compact.

Because the optimization parameter space is user selected, Assumption B1 can be made to hold by the choice of Θ.

The true parameter space Γ satisfies the following conditions.

Assumption B2.

(i) Γ is compact and (2.3) holds.

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

(iii) $\forall \gamma = (\beta, \zeta, \pi, \phi) \in \Gamma$ with $0 < ||\beta|| < \delta$ for some $\delta > 0$, $\gamma_a = (a\beta, \zeta, \pi, \phi) \in \Gamma$ $\forall a \in [0, 1]$.

Assumption B2(ii) guarantees that Γ is not empty and that there are elements γ of Γ whose β values are nonzero but are arbitrarily close to 0, which is the region of the true parameter space where near lack of identification occurs. Assumption B2(iii) ensures that Γ is compatible with the existence of the partial derivatives that arise in (3.10) and Assumption GMM3.

Example 1 (cont.)

Given the definitions in (2.9)–(2.11), the true parameter space Γ is of the form in (2.3). Thus, Assumption B2(i) holds. Assumption B2(ii) follows from the form of $\mathcal{B}^*$ given in (2.9). Assumption B2(iii) follows from the form of $\mathcal{B}^*$ and the fact that Θ^* is a product space and $\Phi^*(\theta^*)$ does not depend on β^*. Hence, the true parameter space Γ satisfies Assumption B2.

The optimization parameter space Θ takes the form

$$\Theta = \mathcal{B} \times \mathcal{Z} \times \Pi, \quad \text{where } \mathcal{B} = [-b_1, b_2] \subset R, \tag{3.23}$$

$b_1 > b_1^*$, $b_2 > b_2^*$, $\mathcal{Z} \subset R^{d_\zeta}$ is compact, $\Pi \subset R^{d_\pi}$ is compact, $\mathcal{Z}^* \subset int(\mathcal{Z})$, and $\mathcal{B}^* \subset int(\mathcal{B})$. Given these conditions, Assumptions B1(i) and B1(iii) follow immediately. Assumption B1(ii) holds by taking $\delta < \min\{b_1^*, b_2^*\}$ and $\mathcal{Z}^0 = int(\mathcal{Z})$.

4. GMM ESTIMATION RESULTS

This section provides the asymptotic results of the paper for the GMM estimator $\widehat{\theta}_n$. Define a concentrated GMM estimator $\widehat{\psi}_n(\pi)$ $(\in \Psi(\pi))$ of ψ for given $\pi \in \Pi$ by

$$Q_n(\widehat{\psi}_n(\pi), \pi) = \inf_{\psi \in \Psi(\pi)} Q_n(\psi, \pi) + o(n^{-1}). \tag{4.1}$$

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

$$Q_n^c(\widehat{\pi}_n) = \inf_{\pi\in\Pi} Q_n^c(\pi) + o(n^{-1}). \tag{4.2}$$

We assume that the GMM estimator $\widehat{\theta}_n$ in (2.4) can be written as $\widehat{\theta}_n = (\widehat{\psi}_n(\widehat{\pi}_n),\widehat{\pi}_n)$. Note that if (4.1) and (4.2) hold and $\widehat{\theta}_n = (\widehat{\psi}_n(\widehat{\pi}_n),\widehat{\pi}_n)$, then (2.4) automatically holds.

For $\gamma_n = (\beta_n,\zeta_n,\pi_n,\phi_n) \in \Gamma$, let $Q_{0,n} = Q_n(\psi_{0,n},\pi)$, where $\psi_{0,n} = (0,\zeta_n)$. Note that $Q_{0,n}$ does not depend on π by Assumption GMM1(i).

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

$$\tau(\pi;\gamma_0,b) = -H^{-1}(\pi;\gamma_0)(G(\pi;\gamma_0) + K(\pi;\gamma_0)b) - (b,0_{d_\zeta}), \tag{4.3}$$

where $(b,0_{d_\zeta}) \in R^{d_\psi}$. Note that, by (3.16) and (4.3), $\xi(\pi;\gamma_0,b) = -1/2\,(\tau(\pi;\gamma_0,b) + (b,0_{d_\zeta}))' H(\pi;\gamma_0)(\tau(\pi;\gamma_0,b) + (b,0_{d_\zeta}))$. Let

$$\pi^*(\gamma_0,b) = \arg\min_{\pi\in\Pi} \xi(\pi;\gamma_0,b). \tag{4.4}$$

THEOREM 4.1. *Suppose Assumptions GMM1–GMM4, B1, and B2 hold. Under* $\{\gamma_n\} \in \Gamma(\gamma_0,0,b)$ *with* $||b|| < \infty$,

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

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

Remarks.

1. The results of Theorem 4.1 and Theorem 4.2 which following are the same as those in Theorems 3.1 and 3.2 of AC1, but they are obtained under more primitive conditions, which are designed for GMM estimators.
2. Define the Gaussian process $\{\tau_\beta(\pi;\gamma_0,b) : \pi \in \Pi\}$ by

   $$\tau_\beta(\pi;\gamma_0,b) = S_\beta \tau(\pi;\gamma_0,b) + b, \tag{4.5}$$

   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 the Wald and t statistics later in this paper.
3. Assumption GMM4 is not needed for Theorem 4.1(b).

THEOREM 4.2. *Suppose Assumptions GMM1–GMM5, B1, and B2 hold. Under* $\{\gamma_n\} \in \Gamma(\gamma_0,\infty,\omega_0)$,

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

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

Remark. The results of Theorems 4.1 and 4.2 hold for MD estimators under the assumptions listed in Supplemental Appendix B.

5. WALD CONFIDENCE SETS AND TESTS

In this section, we consider a CS for a function $r(\theta)$ of θ by inverting a Wald test of the hypotheses $H_0 : r(\theta) = v$ for $v \in r(\Theta)$. We also consider Wald tests of H_0. We establish the asymptotic distributions of the Wald statistic under drifting sequences of null and alternative distributions that cover the entire range of strengths of identification. We determine the asymptotic size of standard Wald CSs. We introduce robust Wald CSs whose asymptotic size is guaranteed to equal their nominal size. The results in this section apply not just to Wald statistics based on GMM estimators but to Wald tests based on any of the estimators considered in AC1 and AC2 also.

5.1. Wald Statistics

The Wald statistics are defined as follows. Let

$$\Sigma(\gamma_0) = J^{-1}(\gamma_0)' V(\gamma_0) J^{-1}(\gamma_0) \quad \text{and} \quad \widehat{\Sigma}_n = \widehat{J}_n^{-1} \widehat{V}_n \widehat{J}_n^{-1}, \tag{5.1}$$

where $\widehat{J}_n$ and $\widehat{V}_n$ are estimators of $J(\gamma_0)$ and $V(\gamma_0)$. The Wald statistic takes the form

$$W_n(v) = n(r(\widehat{\theta}_n) - v)'(r_\theta(\widehat{\theta}_n) B^{-1}(\widehat{\beta}_n) \widehat{\Sigma}_n B^{-1}(\widehat{\beta}_n) r_\theta(\widehat{\theta}_n)')^{-1}(r(\widehat{\theta}_n) - v), \tag{5.2}$$

where $r_\theta(\theta) = (\partial/\partial\theta') r(\theta) \in R^{d_r \times d_\theta}$.

When $d_r = 1$, the t statistic takes the form

$$T_n(v) = \frac{n^{1/2}(r(\widehat{\theta}_n) - v)}{(r_\theta(\widehat{\theta}_n) B^{-1}(\widehat{\beta}_n) \widehat{\Sigma}_n B^{-1}(\widehat{\beta}_n) r_\theta(\widehat{\theta}_n)')^{1/2}}. \tag{5.3}$$

Although these definitions of the Wald and t statistics involve $B^{-1}(\widehat{\beta}_n)$, they are the same as the standard definitions used in practice. By Theorem 4.2(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 Wald statistics, 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 Wald statistics is used under all sequences of true parameters $\gamma_n \in \Gamma(\gamma_0)$.

In the results that follow (except in Section 5.6), we consider the behavior of the Wald statistics 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 following notational simplification:

$$W_n = W_n(v_n), \quad \text{where } v_n = r(\theta_n). \tag{5.4}$$

5.2. Rotation

To obtain the asymptotic distribution of the Wald statistic we consider a rotation of $r(\widehat{\theta}_n)$ and $r_\theta(\widehat{\theta}_n)$ by a matrix $A(\widehat{\theta}_n)$. The rotation is designed to separate the effects of the randomness in $\widehat{\psi}_n$ and $\widehat{\pi}_n$, which have different rates of convergence for some sequences $\{\gamma_n\}$. Similar rotations are carried out in the analysis of partially identified models in Sargan (1983) and Phillips (1989), in the nonstationary time series literature (e.g., see Park and Phillips, 1988), and in the GMM analysis in Antoine and Renault (2009, 2010).

We partition $r_\theta(\theta)$ conformably with $\theta = (\psi, \pi)$:

$$r_\theta(\theta) = [r_\psi(\theta) : r_\pi(\theta)]. \tag{5.5}$$

Suppose rank $(r_\pi(\theta)) = d_\pi^*$ $(\leq \min(d_r, d_\pi))$ $\forall \theta \in \Theta_\delta$ for some $\delta > 0$. (This is Assumption R1(iii) in Section 5.3). For $\theta \in \Theta_\delta$, let $A(\theta) = [A_1(\theta)' : A_2(\theta)']' \in O(d_r)$, where the rows of $A_1(\theta) \in R^{(d_r - d_\pi^*) \times d_r}$ span the null space of $r_\pi(\theta)'$, the rows of $A_2(\theta) \in R^{d_\pi^* \times d_r}$ span the column space of $r_\pi(\theta)$, and $O(d_r)$ stands for the orthogonal group of degree d_r over the real space. Hence,

$$A(\theta) r_\pi(\theta) = \begin{bmatrix} A_1(\theta) r_\pi(\theta) \\ A_2(\theta) r_\pi(\theta) \end{bmatrix} = \begin{bmatrix} 0_{(d_r - d_\pi^*) \times d_\pi} \\ r_\pi^*(\theta) \end{bmatrix}, \tag{5.6}$$

where $r_\pi^*(\theta) \in R^{d_\pi^* \times d_\pi}$ has full row rank d_π^*. For simplicity, hereafter we write the 0 matrix as 0 when there is no confusion about its dimension.

With the $A(\theta)$ rotation, the derivative matrix $r_\theta(\theta)$ becomes

$$r_\theta^A(\theta) = A(\theta) r_\theta(\theta) = \begin{bmatrix} r_\psi^*(\theta) & 0 \\ r_\psi^0(\theta) & r_\pi^*(\theta) \end{bmatrix}, \tag{5.7}$$

where the $(d_r - d_\pi^*) \times d_\psi$ matrix $r_\psi^*(\theta)$ has full row rank $d_r - d_\pi^*$. When $d_\pi^* = d_r$, $A_1(\theta)$ and $[r_\psi^*(\theta) : 0]$ disappear. When $d_\pi^* = 0$, $A_2(\theta)$ and $[r_\psi^0(\theta) : r_\pi^*(\theta)]$ disappear.

The effect of randomness in $\widehat{\pi}_n$ on $r(\widehat{\theta}_n)$ is concentrated in the full rank matrix $r_\pi^*(\widehat{\theta}_n)$ because the upper right corner of $r_\theta^A(\widehat{\theta}_n)$ is 0. The effect of randomness in $\widehat{\psi}_n$ on $r(\widehat{\theta}_n)$ is incorporated in both $r_\psi^*(\widehat{\theta}_n)$ and $r_\psi^0(\widehat{\theta}_n)$.

Using the rotation by $A(\widehat{\theta}_n)$, the Wald statistic in (5.2) can be written as

$$W_n = n(r(\widehat{\theta}_n) - v)' A(\widehat{\theta}_n)' (r_\theta^A(\widehat{\theta}_n) B^{-1}(\widehat{\beta}_n) \widehat{\Sigma}_n B^{-1}(\widehat{\beta}_n) r_\theta^A(\widehat{\theta}_n)')^{-1} A(\widehat{\theta}_n)(r(\widehat{\theta}_n) - v), \tag{5.8}$$

where the first $d_r - d_\pi^*$ rows of $A(\widehat{\theta}_n) r(\widehat{\theta}_n)$ only depend on the randomness in $\widehat{\psi}_n$, not $\widehat{\pi}_n$, asymptotically by the choice of $A(\widehat{\theta}_n)$.

Define a $d_r \times d_\theta$ matrix

$$r_\theta^*(\theta) = \begin{bmatrix} r_\psi^*(\theta) & 0 \\ 0 & r_\pi^*(\theta) \end{bmatrix}. \tag{5.9}$$

The matrix $r_\theta^*(\theta)$, rather than $r_\theta^A(\theta)$, appears in the asymptotic distribution in Section 5.5. The reason is as follows. Because $\widehat{\psi}_n$ converges faster than $\widehat{\pi}_n$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, as shown in Theorems 4.1 and 4.2, the effect of randomness in $\widehat{\pi}_n$ is an order of magnitude larger than that in $\widehat{\psi}_n$. As a result, the limit of $r_\psi^0(\widehat{\theta}_n)$ in (5.7) does not show up in the asymptotic distributions of the Wald and t statistics. On the other hand, the limit of $r_\psi^*(\widehat{\theta}_n)$ does appear in the asymptotic distribution because it is the effect of randomness in $\widehat{\psi}_n$ separated from that in $\widehat{\pi}_n$.

When $r_\pi(\theta)$ has full row rank, i.e., $d_\pi^* = d_r$, for all $\theta \in \Theta_\delta$, we have $A(\theta) = I_{d_r}$, $r_\theta^A(\theta) = r_\theta(\theta)$, and $r_\theta^*(\theta) = [0 : r_\pi(\theta)]$. In this case, rotation is not needed to concentrate the randomness in $\widehat{\pi}_n$. Also, when $d_r = 1$, we have $A(\theta) = 1$, and so no rotation is employed.

Define

$$\eta_n(\theta) = \begin{cases} n^{1/2} A_1(\theta)(r(\psi_n, \pi) - r(\psi_n, \pi_n)) & \text{if } d_\pi^* < d_r \\ 0 & \text{if } d_\pi^* = d_r. \end{cases} \tag{5.10}$$

5.3. Function $\mathbf{r}(\theta)$ of Interest

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

Assumption R1.

(i) $r(\theta)$ is continuously differentiable on Θ.

(ii) $r_\theta(\theta)$ is full row rank d_r $\forall \theta \in \Theta$.

(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 $\delta > 0$.

Assumption R2. $\eta_n(\widehat{\theta}_n) \to_p 0$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ $\forall b \in (R \cup \{\pm\infty\})^{d_\beta}$.

Three different sufficient conditions for the high-level Assumption R2 are given by Assumptions R2*(i)–(iii), which follow. Any one of them is sufficient for Assumption R2 (under the conditions in Lemma 5.1 later in this section).

Assumption R2*.

(i) $d_\pi^* = d_r$.

(ii) $d_r = 1$.

(iii) The column space of $r_\pi(\theta)$ is the same $\forall \theta \in \Theta_\delta$ for some $\delta > 0$.

Assumption R2*(i) requires that the restrictions only involve π. Alternatively, Assumption R2*(ii) requires that only one restriction appears. Alternatively, Assumption R2*(iii) is satisfied when $r_\pi(\theta) = a(\theta) R_\pi$, where $a(\theta) : \Theta_\delta \to R$, $a(\theta) \neq 0$, and $R_\pi \in R^{d_r \times d_\pi}$. A special case is when $r_\pi(\theta)$ is constant because of the restrictions being linear.

Assumption $\mathrm{R_L}$. $r(\theta) = R\theta$, where $R \in R^{d_r \times d_\theta}$ has full row rank d_r.

Assumption $\mathrm{R_L}$ is a sufficient condition for Assumptions R1 and R2.

LEMMA 5.1. *Assumptions R2*(i) and R2*(ii) each (separately) implies Assumption R2. Assumption R2*(iii) combined with Assumption GMM1 (or Assumptions A and B3(i)–(ii) of AC1) implies Assumption R2.*

LEMMA 5.2. *Assumption R_L implies Assumptions R1 and R2.*

5.4. Variance Matrix Estimators

The estimators of the components of the asymptotic variance matrix are assumed to satisfy the following assumptions. Two forms are given for Assumption V1, which follows. The first applies when β is a scalar, and the second applies when β is a vector. The reason for the difference is that the normalizing matrix $B(\beta)$ is different in these two cases.

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 (3.20) (or as in Assumptions D2 and D3 of AC1). Let

$$\Sigma(\theta;\gamma_0) = J^{-1}(\theta;\gamma_0)V(\theta;\gamma_0)J^{-1}(\theta;\gamma_0) \quad \text{and} \quad \Sigma(\pi;\gamma_0) = \Sigma(\psi_0,\pi;\gamma_0). \tag{5.11}$$

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

Assumption V1 applies when β is a scalar.

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 $\beta_0 = 0$.

When β is a vector, i.e., $d_\beta > 1$, we reparameterize β as $(||\beta||,\omega)$, where $\omega = \beta/||\beta||$ if $\beta \neq 0$ and by definition $\omega = 1_{d_\beta}/||1_{d_\beta}||$ with $1_{d_\beta} = (1,...,1) \in R^{d_\beta}$ if $\beta = 0$. Correspondingly, θ is reparameterized as $\theta^+ = (||\beta||,\omega,\zeta,\pi)$. Let $\Theta^+ = \{\theta^+ : \theta^+ = (||\beta||,\beta/||\beta||,\zeta,\pi),\ \theta \in \Theta\}$. Let $\widehat{\theta}_n^+$ and θ_0^+ be the counterparts of $\widehat{\theta}_n$ and θ_0 after reparametrization.

When β is a vector, let $J(\theta^+;\gamma_0)$ and $V(\theta^+;\gamma_0)$ denote some nonstochastic $d_\theta \times d_\theta$ matrix-valued functions such that $J(\theta_0^+;\gamma_0) = J(\gamma_0)$ and $V(\theta_0^+;\gamma_0) = V(\gamma_0)$. Let

$$\Sigma(\theta^+;\gamma_0) = J^{-1}(\theta^+;\gamma_0)V(\theta^+;\gamma_0)J^{-1}(\theta^+;\gamma_0) \quad \text{and}$$
$$\Sigma(\pi,\omega;\gamma_0) = \Sigma(||\beta_0||,\omega,\zeta_0,\pi;\gamma_0). \tag{5.12}$$

Let $\Sigma_{\beta\beta}(\pi,\omega;\gamma_0)$ denote the upper left $d_\beta \times d_\beta$ submatrix of $\Sigma(\pi,\omega;\gamma_0)$.

Assumption V1, which follows, applies when β is a vector.

Assumption V1 (Vector 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 $||b|| < \infty$.[11]

(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) $\lambda_{\min}(\Sigma(\pi,\omega;\gamma_0)) > 0$ and $\lambda_{\max}(\Sigma(\pi,\omega;\gamma_0)) < \infty$ $\forall \pi \in \Pi$, $\forall \omega \in R^{d_\beta}$ with $||\omega|| = 1$, $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$.

(iv) $P(\tau_\beta(\pi^*(\gamma_0,b),\gamma_0,b) = 0) = 0$ $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$ and $\forall b$ with $||b|| < \infty$.

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 $\widehat{V}_n \to_p V(\gamma_0)$.

Example 1 (cont.)
In this example, β is a scalar. The estimators of $J(\gamma_0)$ and $V(\gamma_0)$ are

$$\widehat{J}_n = \widehat{J}_n(\widehat{\theta}_n) \quad \text{and} \quad \widehat{V}_n = \widehat{V}_n(\widehat{\theta}_n), \tag{5.13}$$

respectively, where

$$\widehat{J}_n(\theta) = \widehat{J}_{g,n}(\theta)'\mathcal{W}_n \widehat{J}_{g,n}(\theta),$$
$$\widehat{V}_n(\theta) = \widehat{J}_{g,n}(\theta)'\mathcal{W}_n \widehat{V}_{g,n}(\theta)\mathcal{W}_n \widehat{J}_{g,n}(\theta),$$
$$\widehat{J}_{g,n}(\theta)' = n^{-1}\sum_{i=1}^n Z_i d_i(\pi)', \quad \text{and} \quad \widehat{V}_{g,n}(\theta) = n^{-1}\sum_{i=1}^n U_i^2(\theta) Z_i Z_i'. \tag{5.14}$$

The key quantities in Assumption V1 (scalar β) are

$$J(\theta;\gamma_0) = J_g(\theta;\gamma_0)'\mathcal{W}(\gamma_0)J_g(\theta;\gamma_0) \quad \text{and}$$

$$V(\theta;\gamma_0) = J_g(\theta;\gamma_0)'\mathcal{W}(\gamma_0)V_g(\theta;\gamma_0)\mathcal{W}(\gamma_0)J_g(\theta;\gamma_0), \quad \text{where}$$

$$J_g(\theta;\gamma_0) = -\mathbb{E}_{\phi_0} Z_i d_i(\pi)', \quad \mathcal{W}(\gamma_0) = (\mathbb{E}_{\phi_0} Z_i Z_i')^{-1}, \quad \text{and} \tag{5.15}$$

$$V_g(\theta;\gamma_0) = \mathbb{E}_{\phi_0} U_i^2 Z_i Z_i' + 2\mathbb{E}_{\phi_0}[\beta_0 h(X_{1,i},\pi_0) - \beta h(X_{1,i},\pi) + X_{2,i}(\zeta_0 - \zeta)] Z_i Z_i'$$
$$+ \mathbb{E}_{\phi_0}[\beta_0 h(X_{1,i},\pi_0) - \beta h(X_{1,i},\pi) + X_{2,i}'(\zeta_0 - \zeta)]^2 Z_i Z_i'.$$

Assumption V1(i) holds by the uniform LLN given in Lemma 12.1 in Supplemental Appendix D using the moment conditions in (2.11), Assumption GMM1(ii), and the continuous mapping theorem. Assumption V1(ii) holds by the continuity of $h(x,\pi)$ and $h_\pi(x,\pi)$ in π and the conditions in (2.11).

To verify Assumption V1(iii), note that

$$\Sigma(\pi;\gamma_0) = J^{-1}(\psi_0,\pi;\gamma_0)V(\psi_0,\pi;\gamma_0)J^{-1}(\psi_0,\pi;\gamma_0), \quad \text{where}$$
$$J_g(\psi_0,\pi;\gamma_0) = -\mathbb{E}_{\phi_0} Z_i d_i(\pi)' \quad \text{and} \quad V_g(\psi_0,\pi;\gamma_0) = \mathbb{E}_{\phi_0} U_i^2 Z_i Z_i' \tag{5.16}$$

when $\beta_0 = 0$. We have the following results: $\Sigma(\pi;\gamma_0)$ is positive definite (pd) and finite $\forall\pi \in \Pi$ because both $J(\psi_0,\pi;\gamma_0)$ and $V(\psi_0,\pi;\gamma_0)$ are pd and finite, which in turn holds because (a) $\mathcal{W}(\gamma_0)$ is pd and finite by Assumption GMM1(vii), (b) $J_g(\psi_0,\pi;\gamma_0) \in R^{k\times d_\theta}$ has full rank by (2.11), and (c) $V_g(\psi_0,\pi;\gamma_0)$ is pd and finite by (2.11). This completes the verification of Assumption V1.

Assumptions V1(i) and V1(ii) hold not only under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ but also under $\{\gamma_n\} \in \Gamma(\gamma_0,\infty,\omega_0)$ in this example. This and $\widehat{\theta}_n \rightarrow_p \theta_0$ under $\{\gamma_n\} \in \Gamma(\gamma_0,\infty,\omega_0)$, which holds by Theorem 4.2 (because Assumptions GMM1–GMM5, B1, and B2 have been verified previously), imply that Assumption V2 holds. This completes the verification of Assumption V2.

5.5. Asymptotic Null Distribution of the Wald Statistic

The asymptotic null distribution of the Wald statistic under H_0 depends on the following quantities. The limit distribution of $\widehat{\omega}_n(\pi) = \widehat{\beta}_n(\pi)/||\widehat{\beta}_n(\pi)||$ under $\Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$ is given by

$$\omega^*(\pi;\gamma_0,b) = \frac{\tau_\beta(\pi;\gamma_0,b)}{||\tau_\beta(\pi;\gamma_0,b)||} \quad \text{for } \pi \in \Pi, \tag{5.17}$$

where $\tau_\beta(\pi;\gamma_0,b)$ is defined in (4.5). Let $\overline{B}(\pi;\gamma_0,b)$ be a $d_r \times d_r$ matrix-valued function of $\tau_\beta(\pi;\gamma_0,b)$ defined as

$$\overline{B}(\pi;\gamma_0,b) = \begin{bmatrix} I_{(d_r - d_\pi^*)} & 0 \\ 0 & \imath(\tau_\beta(\pi;\gamma_0,b))I_{d_\pi^*} \end{bmatrix}, \tag{5.18}$$

where $\imath(\beta) = \beta$ when β is a scalar and $\imath(\beta) = ||\beta||$ when β is a vector.

Let

$$r_\theta^*(\pi) = r_\theta^*(\psi_0,\pi), \quad r_\psi^*(\pi) = r_\psi^*(\psi_0,\pi) \quad \text{and}$$

$$\overline{\Sigma}(\pi;\gamma_0,b) = \begin{cases} \Sigma(\pi;\gamma_0) & \text{if } \beta \text{ is a scalar} \\ \Sigma(\pi,\omega^*(\pi;\gamma_0,b);\gamma_0) & \text{if } \beta \text{ is a vector,} \end{cases} \tag{5.19}$$

where $\Sigma(\pi;\gamma_0)$ and $\Sigma(\pi,\omega;\gamma_0)$ are defined in (5.11) and (5.12), respectively.

Define a stochastic process $\{\lambda(\pi;\gamma_0,b) : \pi \in \Pi\}$ by

$$\begin{aligned} \lambda(\pi;\gamma_0,b) = {} & \tau^A(\pi;\gamma_0,b)'\overline{B}(\pi;\gamma_0,b)(r_\theta^*(\pi)\overline{\Sigma}(\pi;\gamma_0,b)r_\theta^*(\pi)')^{-1} \\ & \times \overline{B}(\pi;\gamma_0,b)\tau^A(\pi;\gamma_0,b), \quad \text{where} \end{aligned}$$

$$\tau^A(\pi;\gamma_0,b) = \begin{pmatrix} r_\psi^*(\pi)\tau(\pi;\gamma_0,b) \\ A_2(\psi_0,\pi)(r(\psi_0,\pi) - r(\psi_0,\pi_0)) \end{pmatrix} \in R^{d_r}. \tag{5.20}$$

With linear restrictions, the stochastic process $\lambda(\pi;\gamma_0,b)$ can be simplified. Under Assumption $\mathrm{R_L}$, $r_\theta(\theta) = R$ does not depend on θ, and hence $A(\theta)$ and

$r^*_\theta(\theta)$ do not depend on θ. Define $R^* = r^*_\theta(\theta)$ under Assumption $\mathrm{R_L}$. Specifically,

$$R^A = AR = \begin{bmatrix} R^*_\psi & 0 \\ R^0_\psi & R^*_\pi \end{bmatrix} \quad \text{and} \quad R^* = \begin{bmatrix} R^*_\psi & 0 \\ 0 & R^*_\pi \end{bmatrix}, \tag{5.21}$$

where $R^*_\psi \in R^{(d_r - d^*_\pi)\times d_\psi}$ and $R^*_\pi \in R^{d^*_\pi \times d_\pi}$.

Define a stochastic process $\{\lambda_L(\pi;\gamma_0,b) : \pi \in \Pi\}$ by

$$\begin{aligned} \lambda_L(\pi;\gamma_0,b) &= \overline{\tau}(\pi;\gamma_0,b)' R^{*\prime} \overline{B}(\pi;\gamma_0,b)(R^* \overline{\Sigma}(\pi;\gamma_0,b) R^{*\prime})^{-1} \\ &\quad \times \overline{B}(\pi;\gamma_0,b) R^* \overline{\tau}(\pi;\gamma_0,b), \qquad \text{where} \\ \overline{\tau}(\pi;\gamma_0,b) &= (\tau(\pi;\gamma_0,b)', (\pi - \pi_0)')' \in R^{d_\theta}. \end{aligned} \tag{5.22}$$

Under the linear restriction of Assumption $\mathrm{R_L}$, $\lambda_L(\pi;\gamma_0,b) = \lambda(\pi;\gamma_0,b)$ and the asymptotic distribution of the Wald statistic can be simplified by replacing the stochastic process $\{\lambda(\pi;\gamma_0,b) : \pi \in \Pi\}$ with $\{\lambda_L(\pi;\gamma_0,b) : \pi \in \Pi\}$ in the asymptotic results that follow.

The following theorem establishes the asymptotic null distribution of the Wald statistic for nonlinear restrictions that satisfy Assumption R2. (The null holds by the definition $W_n = W_n(v_n)$ in (5.4).)

THEOREM 5.1. *Suppose Assumptions B1–B2, R1–R2, and V1–V2 hold. In addition, suppose Assumptions GMM1–GMM5 hold (or Assumptions A, B3, C1–C8, and D1–D3 of AC1 hold).*

(a) *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $||b|| < \infty$, $W_n \rightarrow_d \lambda(\pi^*(\gamma_0,b);\gamma_0,b)$.
(b) *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $W_n \rightarrow_d \chi^2_{d_r}$.

A special case of Theorem 5.1 is the following result for linear restrictions.

COROLLARY 5.1. *Suppose Assumptions B1–B2,* R_L*, and V1–V2 hold. In addition, suppose Assumptions GMM1–GMM5 hold (or Assumptions A, B1–B3, C1–C8, and D1–D3 of AC1 hold).*

(a) *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $||b|| < \infty$, $W_n \rightarrow_d \lambda_L(\pi^*(\gamma_0,b);\gamma_0,b)$.
(b) *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $W_n \rightarrow_d \chi^2_{d_r}$.

Specific forms of the stochastic process $\lambda(\pi;\gamma_0,b)$ are provided in the following examples. In Examples r1–r4, $r(\theta)$ is linear in θ and Corollary 5.1 applies. In Example r5, $r(\theta)$ is nonlinear in θ and Assumption R2 is verified.

Example r1

When $r(\theta) = \psi$, $R = R^* = [I_{d_\psi} : 0]$, and $\lambda_L(\pi;\gamma_0,b) = \tau(\pi;\gamma_0,b)'\overline{\Sigma}^{-1}_{\psi\psi}(\pi;\gamma_0,b)\tau(\pi;\gamma_0,b)$, where $\overline{\Sigma}_{\psi\psi}(\pi;\gamma_0,b)$ is the upper left $d_\psi \times d_\psi$ block of $\overline{\Sigma}(\pi;\gamma_0,b)$.

Example r2

When $r(\theta)=\pi$, $R=R^*=[0:I_{d_\pi}]$, and $\lambda_L(\pi;\gamma_0,b)=||\tau_\beta(\pi;\gamma_0,b)||^2(\pi-\pi_0)'\overline{\Sigma}_{\pi\pi}^{-1}(\pi;\gamma_0,b)(\pi-\pi_0)$, where $\overline{\Sigma}_{\pi\pi}(\pi;\gamma_0,b)$ is the lower right $d_\pi\times d_\pi$ block of $\overline{\Sigma}(\pi;\gamma_0,b)$.

Example r3

When $d_\psi=d_\pi$ and $r(\theta)=\psi+\pi$, $R=[I_{d_\psi}:I_{d_\pi}]$, $R^*=[0_{d_\psi}:I_{d_\pi}]$, and $\lambda_L(\pi;\gamma_0,b)=||\tau_\beta(\pi;\gamma_0,b)||^2(\pi-\pi_0)'\overline{\Sigma}_{\pi\pi}^{-1}(\pi;\gamma_0,b)(\pi-\pi_0)$. Note that $\lambda_L(\pi;\gamma_0,b)$ is the same in this example as in Example r2. This occurs because $d_\pi^*=d_r$ so that the randomness in $\widehat{\psi}_n$ is completely dominated by that in $\widehat{\pi}_n$. Although R is different in Examples r2 and r3, R^* is the same in both examples.

Example r4

When $r(\theta)=\theta$, $R=R^*=I_{d_\theta}$, and $\lambda_L(\pi;\gamma_0,b)=\overline{\tau}(\pi;\gamma_0,b)'\overline{B}(\pi;\gamma_0,b)\overline{\Sigma}^{-1}(\pi;\gamma_0,b)\overline{B}(\pi;\gamma_0,b)\overline{\tau}(\pi;\gamma_0,b)$.

Example r5

When $\theta=(\beta,\pi)'$, $r(\theta)=(\beta,\pi^2)'$, and β and π are scalars, we have

$$r_\theta(\theta)=r_\theta^*(\theta)=\begin{bmatrix}1 & 0\\ 0 & 2\pi\end{bmatrix} \quad \text{and} \quad A(\theta)=I_2. \tag{5.23}$$

Assumption R2*(iii) holds because $A_2(\theta)$ does not depend on θ. This implies that Assumption R2 holds. The stochastic process $\{\tau^A(\pi;\gamma_0,b):\pi\in\Pi\}$ can be simplified to $\tau^A(\pi;\gamma_0,b)=(\tau(\pi;\gamma_0,b),\pi^2-\pi_0^2)$.

Next we show that Assumption R2 is not superfluous. In certain cases, the Wald statistic diverges to infinity in probability under H_0.

THEOREM 5.2. *Suppose Assumptions B1–B2, R1, and V1 hold. In addition, suppose Assumptions GMM1–GMM4 hold (or Assumptions A, B1–B3, and C1–C8 of AC1 hold). Under* $\{\gamma_n\}\in\Gamma(\gamma_0,0,b)$, $W_n\to_p\infty$ *if* $||\eta_n(\widehat{\theta}_n)||\to_p\infty$.

Remark. This theorem provides a high-level condition under which the Wald statistic diverges to infinity in probability under the null. This result holds for sequences $\{\gamma_n\}$ in both the weak and semistrong identification categories. The Wald statistic, which uses $r_\theta(\widehat{\theta}_n)$ in the covariance matrix estimation, is designed for the standard case in which $\widehat{\theta}_n$ converges to θ_n at rate $n^{-1/2}$. When $\widehat{\pi}_n$ is inconsistent or converges to π_n more slowly than $n^{-1/2}$, the estimator of the covariance matrix does not necessarily provide a proper normalization for the Wald statistic to have a nondegenerate limit.

Example r6

We now demonstrate that restrictions exist for which Assumption R2 fails to hold. Suppose $\theta=(\beta,\pi)'$, $r(\theta)=((\beta+1)\pi,\pi^2)'$, and β and π are both scalars. Then, we have

$$r_\theta(\theta) = \begin{bmatrix} \pi & \beta+1 \\ 0 & 2\pi \end{bmatrix}, \qquad A_1(\theta) = \frac{1}{||(-2\pi, \beta+1)||}(-2\pi, \beta+1), \quad \text{and}$$

$$\eta_n(\theta) = -\frac{n^{1/2}}{||(-2\pi, \beta+1)||}\left[-2\pi(\beta_n+1)(\pi-\pi_n) + (\beta+1)(\pi^2-\pi_n^2)\right]. \quad \textbf{(5.24)}$$

Consider a sequence $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$. Suppose Assumptions B1, B2, and GMM1–GMM5 hold. If $|b| < \infty$, assume $P(\pi^*(\gamma_0, b) = 0) = 0$ (which typically holds when Π contains a nondegenerate interval). Some calculations show that under $\{\gamma_n\}$, we have $\eta_n(\widehat{\theta}_n) = ||(-2\pi_0, 1)||^{-1}[n^{1/2}\beta_n(\widehat{\pi}_n - \pi_n)]^2(n^{1/4}\beta_n)^{-2}(1 + o(1)) + O_p(1)$.[12] In consequence, if $n^{1/4}\beta_n \to 0$, then $\eta_n(\widehat{\theta}_n) \to_p \infty$ and Theorem 5.2 applies.[13]

Sequences for which $n^{1/2}\beta_n \to \infty$ and $n^{1/4}\beta_n \to 0$ are in the semistrong identification category. Hence, this example shows that even for sequences in the semistrong identification category, in which case both $\widehat{\beta}_n$ and $\widehat{\pi}_n$ are consistent and asymptotically normal, the Wald test can diverge to infinity for nonlinear restrictions because of the different rates of convergence of $\widehat{\beta}_n$ and $\widehat{\pi}_n$.

Stock and Yogo (2005) specify several tests for weak instruments in a linear instrumental variables regression model. Wright (2003) specifies a test for lack of identification in a GMM context. All of these tests reject the null hypothesis of weak identification or no identification with probability that goes to one as $n \to \infty$ in Example r6 when $n^{1/2}\beta_n \to \infty$ and $n^{1/4}\beta_n \to 0$. (For the Stock and Yogo (2005) test, this is true for any fixed finite choice of the critical value for the test.) Hence, these tests are not able to detect situations where problems arise with some Wald tests as in Example r6. (Note that the version of the Stock and Yogo (2005) test that is designed to control the size of a Wald test applies to a Wald test of the null hypothesis that completely specifies the value of the endogenous variable vector. It is not designed for the null hypothesis specified in Example r6.)

Armstrong et al. (2012) provide results that are related to those in Theorem 5.2. Their results apply to simple (and hence linear) null hypotheses in nonlinear models, whereas Theorem 5.2 applies to nonlinear hypotheses in linear or nonlinear models. In both cases, it is shown that Wald tests can have incorrect asymptotic size in semistrong identification scenarios.

5.6. Asymptotic Distribution of the Wald Statistic under the Alternative

Next, we provide the asymptotic distributions of the Wald test under alternative hypotheses, which yield power results for the Wald test and false coverage probabilities for Wald CSs. Suppose the conditions of Theorem 5.1 hold. The following results are obtained by altering of the proof of Theorem 5.1. Suppose the sequence of null hypothesis values of $r(\theta)$ are $\{v_{n,0}^{null} : n \geq 1\}$.[14] We consider the case where the true parameters $\{\gamma_n\}$ satisfy $r(\theta_n) \neq v_{n,0}^{null}$.

First, consider the alternative hypothesis distributions $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $b \in R^{d_\beta}$. Suppose the sequence of true values $\{\theta_n\}$ satisfies $n^{1/2}(r(\theta_n) - v_{n,0}^{null}) \to d$

for some $d \in R^{d_r}$. Then, the asymptotic distribution of $W_n(v_{n,0}^{null})$ is given by the expression in Theorem 5.1(a), but with $\tau^A(\pi;\gamma_0,b)$ in the definition of $\lambda(\pi;\gamma_0,b)$ replaced by $\tau^{A*}(\pi;\gamma_0,b) = \tau^A(\pi;\gamma_0,b) + (A_1(\psi_0,\pi)d, 0_{d_\pi^*})$. Alternatively, suppose the sequence of true values satisfies $r(\theta_n) - v_{n,0}^{null} \to d_0 \in R^{d_r}$ and $d_0 \neq 0$. When $A_1(\theta) \neq 0$ $\forall \theta \in \Theta$, $W_n(v_{n,0}^{null}) \to_p \infty$. When $A_1(\theta) = 0$ $\forall \theta \in \Theta$, the asymptotic distribution of $W_n(v_{n,0}^{null})$ is given by the expression in Theorem 5.1(a), but with $\tau^A(\pi;\gamma_0,b)$ in the definition of $\lambda(\pi;\gamma_0,b)$ replaced by $\tau^{A**}(\pi;\gamma_0,b) = \tau^A(\pi;\gamma_0,b) + (0_{d_r-d_\pi^*}, A_2(\psi_0,\pi)d_0)$.

Next, consider the alternative hypothesis distributions $\{\gamma_n\} \in \Gamma(\gamma_0,\infty,\omega_0)$ with $\beta_0 \neq 0$. When $n^{1/2}(r(\theta_n) - v_{n,0}^{null}) \to d$ for some $d \in R^{d_r}$, $W_n(v_{n,0}^{null})$ converges in distribution to a non-central $\chi^2_{d_r}$ distribution with noncentrality parameter $\delta^2 = d'(r_\theta(\theta_0)B^{-1}(\beta_0)\Sigma(\gamma_0)B^{-1}(\beta_0)r_\theta(\theta_0)')^{-1}d$. Alternatively, when $r(\theta_n) - v_{n,0}^{null} \to d_0$ for some $d_0 \in R^{d_r}$ with $d_0 \neq 0$, $W_n \to_p \infty$.

Lastly, consider the alternative hypothesis distributions $\{\gamma_n\} \in \Gamma(\gamma_0,\infty,\omega_0)$ with $\beta_0 = 0$. Suppose the restrictions satisfy $r(\theta) = (r_1(\psi), r_2(\theta))$ for $r_2(\theta) \in R^{d_\pi^*}$ with $d_\pi^* \geq 0$ and the $d_\pi^* \times d_\pi$ matrix $(\partial/\partial\pi')r_2(\theta)$ has full rank d_π^*.[15] Let $v_{n,0}^{null} = (v_{n,0,1}^{null}, v_{n,0,2}^{null})$ for $v_{n,0,2}^{null} \in R^{d_\pi^*}$. When

$$n^{1/2}(r_1(\theta_n) - v_{n,0,1}^{null}) \to d_1 \in R^{d_r - d_\pi^*} \quad \text{and} \quad n^{1/2}\iota(\beta_n)(r_2(\theta_n) - v_{n,0,2}^{null}) \to d_2 \in R^{d_\pi^*}, \tag{5.25}$$

the asymptotic distribution of $W_n(v_{n,0}^{null})$ is a noncentral $\chi^2_{d_r}$ distribution with noncentrality parameter $\delta^2 = d'(r_\theta^*(\theta_0)\Sigma(\gamma_0)r_\theta^*(\theta_0)')^{-1}d$, where $d = (d_1, d_2) \in R^{d_r}$. Note that the local alternatives in (5.25) are $n^{-1/2}$-alternatives for the $r_1(\psi)$ restrictions but are more distant $n^{-1/2}\iota(\beta_n)^{-1}$-alternatives for the $r_2(\theta)$ restrictions because of the slower $n^{1/2}\iota(\beta_n)$-rate of convergence of $\widehat{\pi}_n$ in the present context. Alternatively, when $r(\theta_n) - v_{n,0}^{null} \to d_0$ for some $d_0 \in R^{d_r}$ with $d_0 \neq 0$, $W_n \to_p \infty$.

5.7. Asymptotic Size of Standard Wald Confidence Sets

Here, we determine the asymptotic size of a standard CS for $r(\theta) \in R^{d_r}$ obtained by inverting a Wald statistic, i.e.,

$$CS_{W,n} = \{v : W_n(v) \leq \chi^2_{d_r,1-\alpha}\}, \tag{5.26}$$

where the Wald statistic $W_n(v)$ is as in (5.2), $\chi^2_{d_r,1-\alpha}$ is the $1-\alpha$ quantile of a chi-square distribution with d_r degree of freedom, and $1-\alpha$ is the nominal size of the CS.

The asymptotic size of the preceding CS above is determined using the asymptotic distribution of $W_n = W_n(r(\theta_n))$ under drifting sequences of true parameters, as given in Theorems 5.1 and 5.2. For $||b|| < \infty$, define

$$h = (b,\gamma_0), \quad H = \{h = (b,\gamma_0) : ||b|| < \infty, \gamma_0 \in \Gamma \text{ with } \beta_0 = 0\}, \quad \text{and}$$
$$W(h) = \lambda(\pi^*(\gamma_0,b);\gamma_0,b). \tag{5.27}$$

As defined, $W(h)$ is the asymptotic distribution of W_n under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ for $||b|| < \infty$ determined in Theorem 5.1(a).

Let $c_{W,1-\alpha}(h)$ denote the $1-\alpha$ quantile of $W(h)$ for $h \in H$.

As in (2.24), $AsySz$ denotes the asymptotic size of a CS of nominal level $1-\alpha$. The asymptotic size results use the following distribution function (df) continuity assumption, which typically is not restrictive.

Assumption V4. The df of $W(h)$ is continuous at $\chi^2_{d_r,1-\alpha}$ and $\sup_{h\in H} c_{W,1-\alpha}(h)\ \forall h \in H$.

THEOREM 5.3. *Suppose Assumptions B1–B2, R1–R2, V1–V2, and V4 hold. In addition, suppose Assumptions GMM1–GMM5 hold (or Assumptions A, B1–B3, C1–C8, and D1–D3 of AC1 hold). Then, the standard nominal* $1-\alpha$ *Wald CS satisfies*

$$AsySz = \min\{\inf_{h\in H} P(W(h) \leq \chi^2_{d_r,1-\alpha}),\ 1-\alpha\}.$$

Remark. Under Assumption $\mathrm{R_L}$ (i.e., linearity of $r(\theta)$), Theorem 5.3 holds with $W(h)$ replaced by the equivalent, but simpler, quantity $W_L(h) = \lambda_L(\pi^*(\gamma_0, b); \gamma_0, b)$ for $h = (b, \gamma_0)$. This holds by Corollary 5.1(a).

Theorem 5.2 shows that the Wald statistic W_n diverges to infinity in some circumstances, e.g., see Example r6 in Section 5.5. In such cases, the standard Wald CS has asymptotic size equal to 0.

COROLLARY 5.2. *Suppose Assumptions B1–B2, R1, and V1 hold. In addition, suppose Assumptions GMM1–GMM5 hold (or Assumptions A, B1–B3, C1–C8, and D1-D3 of AC1 hold). If* $||\eta_n(\widehat{\theta}_n)|| \to_p \infty$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *for some* $\gamma_0 \in \Gamma$ *and* $||b|| < \infty$*, the standard nominal* $1-\alpha$ *Wald CS has* $AsySz = 0$.

5.8. Robust Wald Confidence Sets

Next, we construct Wald CSs that have correct asymptotic size. These CSs are robust to the strength of identification. The CSs for $r(\theta)$ are constructed by inverting a robust Wald test that combines the Wald test statistic with a robust critical value that differs from the usual $\chi^2_{d_r}$-quantile, which is designed for the strong-identification case. The first robust CS uses the least favorable (LF) critical value. The second robust CS, called a type 2 robust CS, is introduced in AC1. It uses a data-dependent critical value. It is smaller than the LF robust CS under strong identification and hence is preferable.

5.8.1. Least Favorable Critical Value. The LF critical value is

$$c^{LF}_{W,1-\alpha} = \max\left\{\sup_{h\in H} c_{W,1-\alpha}(h),\ \chi^2_{d_r,1-\alpha}\right\}. \tag{5.28}$$

The LF critical value can be improved (i.e., made smaller) by exploiting the knowledge of the null hypothesis value of $r(\theta)$. For instance, if the null hypothesis specifies the value of π to be 3, then the supremum in (5.28) 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.[16]

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 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 formulas in (5.28), and the supremum over H 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.[17]

5.8.2. Type 2 Robust Critical Value. Next, we define the type 2 robust critical value. It improves on the LF critical value. It employs an identification category selection (ICS) procedure that uses the data to determine whether b is finite.[18] The ICS procedure chooses between the identification categories $\mathcal{IC}_0 : ||b|| < \infty$ and $\mathcal{IC}_1 : ||b|| = \infty$. The ICS statistic is

$$A_n = \left(n\widehat{\beta}_n' \widehat{\Sigma}_{\beta\beta,n}^{-1} \widehat{\beta}_n / d_\beta \right)^{1/2}, \tag{5.29}$$

where $\widehat{\Sigma}_{\beta\beta,n}$ is the upper left $d_\beta \times d_\beta$ block of $\widehat{\Sigma}_n$, which is defined in (5.1).

The type 2 robust critical value provides a continuous transition from a weak-identification critical value to a strong-identification critical value using a transition function $s(x)$. 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 (i) $s(x) = \exp(-c \cdot x)$ for some $c > 0$ and (ii) $s(x) = (1 + c \cdot x)^{-1}$ for some $c > 0$.[19] For example, in the nonlinear regression model with endogeneity, we use the function $s(x) = \exp(-2x)$.

The type 2 robust critical value is

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

$$c_B = c_{W,1-\alpha}^{LF} + \Delta_1, \qquad c_S = \chi^2_{d_r,1-\alpha} + \Delta_2, \tag{5.30}$$

and $\Delta_1 \ge 0$ and $\Delta_2 \ge 0$ are asymptotic size-correction factors that are defined subsequently. Here, B denotes Big, and S denotes Small. When $A_n \le \kappa$, $\widehat{c}_{W,1-\alpha,n}$ equals the LF critical value $c_{W,1-\alpha}^{LF}$ plus a size-correction factor Δ_1. When $A_n > \kappa$, $\widehat{c}_{W,1-\alpha,n}$ is a linear combination of $c_{W,1-\alpha}^{LF} + \Delta_1$ and $\chi^2_{d_r,1-\alpha} + \Delta_2$, where Δ_2 is another size-correction factor. The weight given to the standard critical value $\chi^2_{d_r,1-\alpha}$ increases with the strength of identification, as measured by $A_n - \kappa$.

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

$$A(h) = \left(\tau_\beta(\pi^*; \gamma_0, b)' \Sigma_{\beta\beta}^{-1}(\pi^*; \gamma_0) \tau_\beta(\pi^*; \gamma_0, b)/d_\beta\right)^{1/2}, \tag{5.31}$$

where π^* abbreviates $\pi^*(\gamma_0, b)$, $\tau_\beta(\pi; \gamma_0, b)$ is defined in (4.5), and $\Sigma_{\beta\beta}(\pi; \gamma_0)$ is the upper left (1,1) element of $\Sigma(\psi_0, \pi; \gamma_0)$ for $\Sigma(\theta; \gamma_0) = J^{-1}(\theta; \gamma_0) V(\theta; \gamma_0) J^{-1}(\theta; \gamma_0)$.[20]

Under $\gamma_n \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$, the asymptotic null rejection probability of a test based on the statistic W_n and the robust critical value $\widehat{c}_{W,1-\alpha,n}$ is equal to

$$\begin{aligned} NRP(\Delta_1, \Delta_2; h) &= P(W(h) > c_B \;\&\; A(h) \le \kappa) + P(W(h) > c_A(h) \;\&\; A(h) > \kappa) \\ &= P(W(h) > c_B) + P(W(h) \in (c_A(h), c_B] \;\&\; A(h) > \kappa), \quad \text{where} \\ c_A(h) &= c_S + (c_B - c_S) \cdot s(A(h) - \kappa). \end{aligned} \tag{5.32}$$

The constants Δ_1 and Δ_2 are chosen such that $NRP(\Delta_1, \Delta_2; h) \le \alpha$ $\forall h \in H$. In particular, we define $\Delta_1 = \sup_{h \in H_1} \Delta_1(h)$, where $\Delta_1(h) \ge 0$ solves $NRP(\Delta_1(h), 0; h) = \alpha$ (or $\Delta_1(h) = 0$ if $NRP(0, 0; h) < \alpha$), $H_1 = \{(b, \gamma_0) : (b, \gamma_0) \in H \;\&\; ||b|| \le ||b_{\max}|| + D\}$, $b_{\max}$ is defined such that $c_{W,1-\alpha}(h)$ 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. We define $\Delta_2 = \sup_{h \in H} \Delta_2(h)$, where $\Delta_2(h)$ solves $NRP(\Delta_1, \Delta_2(h); h) = \alpha$ (or $\Delta_2(h) = 0$ if $NRP(\Delta_1, 0; h) < \alpha$).[21] As defined, Δ_1 and Δ_2 can be computed sequentially, which eases computation.

Given the definitions of Δ_1 and Δ_2, the asymptotic rejection probability is always less than or equal to the nominal level α, and it is close to α when h is close to $h_{\max}$ (because of the adjustment by Δ_1) and when $||b||$ is large (because of the adjustment by Δ_2).

The type 2 robust critical value can be improved by employing NI and/or plug-in versions of it, denoted by $\widehat{c}_{W,1-\alpha,n}(v)$. These are defined by replacing $c^{LF}_{W,1-\alpha}$ in (5.30) by the NI-LF or plug-in NI-LF critical value and making c_B, Δ_1, and Δ_2 depend on the null value v, denoted $c_B(v)$, $\Delta_1(v)$, and $\Delta_2(v)$. We recommend using these versions whenever possible because they lead to smaller CSs.

For any given value of κ, the type 2 robust CS has correct asymptotic size as a result of the choice of Δ_1 and Δ_2. In consequence, a good choice of κ depends on the false coverage probabilities (FCPs) of the robust CS. (A FCP of a CS for $r(\theta)$ is the probability that the CS includes a value different from the true value $r(\theta)$.) The numerical work in this paper and in AC1 and AC2 shows that if a reasonable value of κ is chosen, such as $\kappa = 1.5$ or 2.0, the FCPs of type 2 robust CSs are not sensitive to deviations from this value of κ. This is because the size-correction constants Δ_1 and Δ_2 have to adjust as κ is changed to maintain correct asymptotic size. The adjustments of Δ_1 and Δ_2 offset the effect of changing κ.

One can select κ in a simple way, i.e., by taking $\kappa = 1.5$ or 2.0, or one can select κ in a more sophisticated way that explicitly depends on FCPs. Both methods yield similar results for the cases that we have considered.

The more sophisticated method of choosing κ is to minimize the average FCP of the robust CS over a chosen set of κ values denoted by $\mathcal{K}$. First, for given $h \in H$, one chooses 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.[22] Second, one computes the FCP of the value $v_{H_0}(h)$ for each robust CS with $\kappa \in \mathcal{K}$. Third, one repeats steps one and two for each $h \in \mathcal{H}$, where $\mathcal{H}$ is a representative subset of H.[23] The optimal choice of κ is the value that minimizes over $\mathcal{K}$ the average over $h \in \mathcal{H}$ of the FCP's at $v_{H_0}(h)$.

In summary, the steps used to construct a type 2 robust Wald (or t) test are as follows: (1) Estimate the model using the standard GMM estimator, yielding $\widehat{\beta}_n$ and the covariance matrix $\widehat{\Sigma}_{\beta\beta,n}$. (2) Compute the Wald statistic using the formula in (5.2). (3) Construct the ICS statistic A_n defined in (5.29). (4) Simulate the LF critical value $c^{LF}_{W,1-\alpha}$ and the size correction factors Δ_1 and Δ_2 based on the asymptotic formulas in (5.27), (5.31), and (5.32) and the description following (5.32), for a given value of κ. (5) Compute the type 2 robust critical value $\widehat{c}_{W,1-\alpha,n}$ defined in (5.30), employing the NI and/or plug-in versions when applicable. (6) Choose κ by minimizing the FCP of the type 2 robust CI. The last step can be avoided when the type 2 robust CI constructed is not very sensitive to the choice of κ, which is typically the case found in our simulation studies. For a type 2 robust CI for a particular parameter, one takes the CI to consist of all null values of the parameter for which the type 2 robust test fails to reject the null hypothesis. This can be computed by grid search or some more sophisticated method, such as a multistep grid search where the fineness of the grid varies across the steps.

5.8.3. Asymptotic Size of Robust Wald CSs. In this section, we show that the LF and data-dependent robust CSs defined earlier have correct asymptotic size. The asymptotic size results rely on the following df continuity conditions, which are not restrictive in most examples.

Assumption LF.

(i) The df of $W(h)$ is continuous at $c_{W,1-\alpha}(h)$ $\forall h \in H$.
(ii) If $c^{LF}_{W,1-\alpha} > \chi^2_{d_r,1-\alpha}$, $c^{LF}_{W,1-\alpha}$ is attained at some $h_{\max} \in H$.

Assumption NI-LF.

(i) The df of $W(h)$ is continuous at $c_{W,1-\alpha}(h)$ $\forall h \in H(v)$, $\forall v \in V_r$.
(ii) For some $v \in V_r$, $c^{LF}_{W,1-\alpha}(v) = \chi^2_{d_r,1-\alpha}$ or $c^{LF}_{W,1-\alpha}(v)$ is attained at some $h_{\max} \in H$.

For $h \in H$, define

$$\widehat{c}_{W,1-\alpha}(h) = \begin{cases} c_B & \text{if } A(h) \leq \kappa \\ c_S + [c_B - c_S] \cdot s(A(h) - \kappa) & \text{if } A(h) > \kappa. \end{cases} \tag{5.33}$$

As defined, $\widehat{c}_{W,1-\alpha}(h)$ equals $\widehat{c}_{W,1-\alpha,n}$ with $A(h)$ in place of A_n. The asymptotic distribution of $\widehat{c}_{W,1-\alpha,n}$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ for $||b|| < \infty$ is the distribution of $\widehat{c}_{W,1-\alpha}(h)$.

Define $\widehat{c}_{W,1-\alpha}(h, v)$ analogously to $\widehat{c}_{W,1-\alpha}(h)$, but with $c^{LF}_{W,1-\alpha}$, Δ_1, and Δ_2 replaced by $c^{LF}_{W,1-\alpha}(v)$, $\Delta_1(v)$, and $\Delta_2(v)$, respectively, for $v \in V_r$. The asymptotic distribution of $\widehat{c}_{W,1-\alpha,n}(v)$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ for $||b|| < \infty$ is the distribution of $\widehat{c}_{W,1-\alpha}(h, v)$.

Assumption Rob2.

(i) $P(W(h) = \widehat{c}_{W,1-\alpha}(h)) = 0 \; \forall h \in H$.

(ii) If $\Delta_2 > 0$, $NRP(\Delta_1, \Delta_2; h^*) = \alpha$ for some point $h^* \in H$, where Δ_1 and Δ_2 are defined following (5.32).

Assumption NI-Rob2.

(i) $P(W(h) = \widehat{c}_{W,1-\alpha}(h, v)) = 0 \; \forall h \in H(v), \forall v \in V_r$.

(ii) For some $v \in V_r$, $\Delta_2(v) = 0$ or $NRP(\Delta_1(v), \Delta_2(v); h^*) = \alpha$ for some point $h^* \in H(v)$, where $\Delta_1(v)$ and $\Delta_2(v)$ are defined following (5.32).

THEOREM 5.4. *Suppose Assumptions B1–B2, R1–R2, and V1–V2 hold. In addition, suppose Assumptions GMM1–GMM5 hold (or Assumptions A, B1–B3, C1–C8, and D1–D3 of AC1 hold). Then, the nominal* $1 - \alpha$ *robust Wald CS has* $AsySz = 1 - \alpha$ *when based on the following critical values: (i) LF, (ii) NI–LF, (iii) type 2 robust, and (iv) type 2 NI robust, provided the following additional assumptions hold, respectively: (i) LF, (ii) NI–LF, (iii) Rob2, and (iv) NI–Rob2.*

Remarks.

1. Plug-in versions of the robust Wald CSs considered in Theorem 5.4 also have asymptotically correct size under continuity assumptions on $c_{W,1-\alpha}(h)$ that typically are not restrictive. For brevity, we do not provide formal results here.
2. If part (ii) of Assumption LF, NI-LF, Rob2, or NI-Rob2 does not hold, then the corresponding part of Theorem 5.4 still holds but with $AsySz \geq 1 - \alpha$.
3. A third type of robust critical value, referred to as type 1, is considered in AC1. Critical values of this type can be employed with Wald statistics. The resulting type 1 robust CSs outperform LF robust CSs in terms of FCPs but are inferior to type 2 robust CSs. However, they are easier to compute than type 2 robust CSs.

6. QUASI–LIKELIHOOD RATIO CONFIDENCE SETS AND TESTS

In this section, we introduce CSs based on the QLR statistic. For brevity, theoretical results for the QLR procedures are given in AC1. However, we define QLR procedures here because numerical results are reported for them in the numerical results section.

We consider CSs for a function $r(\theta)$ ($\in R^{d_r}$) of θ obtained by inverting QLR tests. The function $r(\theta)$ is assumed to be smooth and to be of the form

$$r(\theta) = \begin{bmatrix} r_1(\psi) \\ r_2(\pi) \end{bmatrix}, \tag{6.1}$$

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 $d_r = d_{r_1} + d_{r_2}$.

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,

$$\widetilde{\theta}_n(v) \in \Theta, \qquad r(\widetilde{\theta}_n(v)) = v, \quad \text{and} \quad Q_n(\widetilde{\theta}_n(v)) = \inf_{\theta \in \Theta : r(\theta) = v} Q_n(\theta) + o(n^{-1}). \tag{6.2}$$

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

$$QLR_n(v) = 2n(Q_n(\widetilde{\theta}_n(v)) - Q_n(\widehat{\theta}_n))/\widehat{s}_n, \tag{6.3}$$

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 AC1 for details.

Let $c_{n,1-\alpha}(v)$ denote a nominal level $1-\alpha$ critical value to be used with the QLR test statistic. It may be stochastic or nonstochastic. The usual choice, based on the asymptotic distribution of the QLR statistic under standard regularity conditions, is the $1-\alpha$ quantile of the $\chi^2_{d_r}$ distribution: $c_{n,1-\alpha}(v) = \chi^2_{d_r,1-\alpha}$.

A critical value that delivers a robust QLR CS for $r(\theta)$ that has correct asymptotic size can be constructed using the same approach as in Section 5.8.3. Details are in AC1.

Given a critical value $c_{n,1-\alpha}(v)$, the nominal level $1-\alpha$ QLR CS for $r(\theta)$ is

$$CS^{QLR}_{r,n} = \{v \in r(\Theta) : QLR_n(v) \leq c_{n,1-\alpha}(v)\}. \tag{6.4}$$

7. NUMERICAL RESULTS: NONLINEAR REGRESSION MODEL WITH ENDOGENEITY

In this section, we provide asymptotic and finite-sample simulation results for the nonlinear regression model with endogeneity.

The model we consider consists of a structural equation with two right-hand-side endogenous variables X_1 and X_2, where X_1 is a nonlinear regressor and X_2 is a linear regressor, and two reduced-form equations for X_1 and X_2, respectively:

$$\begin{aligned} Y_i &= \zeta_1 + \beta \cdot h(X_{1,i}, \pi) + \zeta_2 X_{2,i} + U_i, \\ X_{1,i} &= \lambda_1 + \lambda_2 Z_{1,i} + V_{1,i}, \\ X_{2,i} &= \lambda_3 + \lambda_4 Z_{2,i} + \lambda_5 Z_{3,i} + V_{2,i}, \end{aligned} \tag{7.1}$$

where $Y_i, X_{1,i}, X_{2,i} \in R$ are endogenous variables, $Z_{1,i}, Z_{2,i}, Z_{3,i} \in R$ are excluded exogenous variables, $h(x,\pi) = (|x|^{\pi} - 1)/\pi$, and $\theta = (\beta, \zeta_1, \zeta_2, \pi)' \in R^4$ is the unknown parameter.[24] The data generating process (DGP) satisfies $(\zeta_1, \zeta_2) = (-2, 2), (\lambda_1, \lambda_2) = (3, 1)$, $(\lambda_3, \lambda_4, \lambda_5) = (0, 1, 1)$, $\{(Z_{1,i}, Z_{2,i}, Z_{3,i}, U_i, V_{1,i}, V_{2,i}) : i = 1, ..., n\}$ are i.i.d., $(Z_{1,i}, Z_{2,i}, Z_{3,i})$ and $(U_i, V_{1,i}, V_{2,i})$ are independent, $(Z_{1,i}, Z_{2,i}, Z_{3,i}) \sim N(0, I_3)$, $U_i \sim N(0, 0.25)$, $V_{k,i} \sim N(0,1)$ and $\text{Corr}(U_i, V_{k,i}) = 0.5$ for $k = 1$ and 2, and $\text{Corr}(V_{1,i}, V_{2,i}) = 0.5$.

The IVs for the GMM estimator of θ are $Z_i = (1, Z_{1,i}, Z_{1,i}^2, Z_{2,i}, Z_{3,i})' \in R^5$. Thus, five moment conditions are used to estimate four parameters.

The true parameter space for π is [1.5, 3.5] and the optimization space for π is $[1, 4]$. The finite-sample results are for $n = 500$. The number of simulation repetitions is 20,000.[25]

Figures 1 and 2 provide the asymptotic and finite-sample densities of the GMM estimators of β and π when the true π value is $\pi_0 = 1.5$. Each figure gives the densities for $b = 0$, 4, 10, and 30, where b indexes the magnitude of β. Specifically, for the finite-sample results, $b = n^{1/2}\beta$. Figures S-1 and S-2 in Supplemental Appendix E provide analogous results for $\pi_0 = 3.0$.

Figure 1 shows that the ML estimator of β has a distribution that is very far from a normal distribution in the unidentified and weakly identified cases. The figure shows a build-up of mass at 0 in the unidentified case and a bimodal distribution in the weakly identified case. Figure 2 shows that there is a build-up of mass at the boundaries of the optimization space for the estimator of π in the unidentified and weakly identified cases. Figures 1 and 2 indicate that the asymptotic approximations developed here work very well.

Figures S-3 to S-6 in Supplemental Appendix E provide the asymptotic and finite-sample ($n = 500$) densities of the t and QLR statistics for β and π when

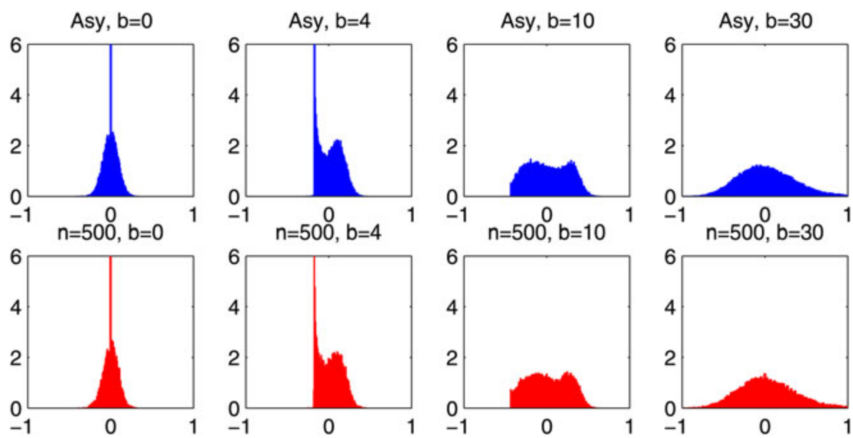

FIGURE 1. Asymptotic and finite-sample ($n = 500$) densities of the estimator of β in the nonlinear regression model with endogeneity when $\pi_0 = 1.5$.

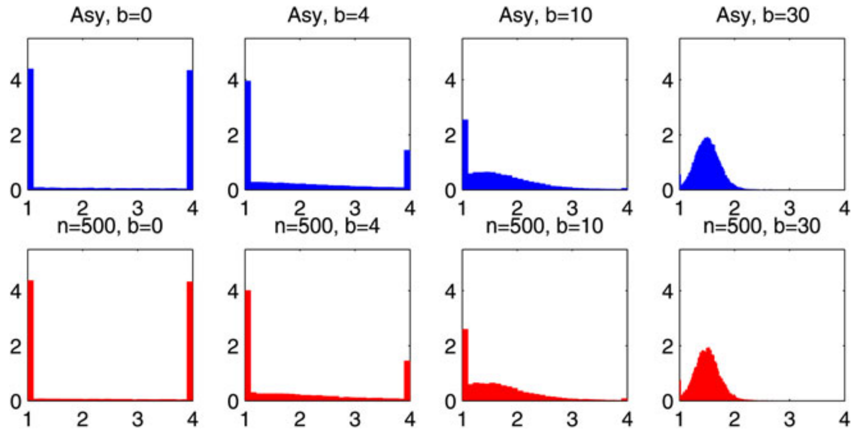

FIGURE 2. Asymptotic and finite-sample ($n = 500$) densities of the estimator of π in the nonlinear regression model with endogeneity when $\pi_0 = 1.5$.

$\pi_0 = 1.5$. These figures show that in the case of weak identification the t and QLR statistics are not well approximated by standard normal and χ_1^2 distributions. However, the asymptotic approximations developed here work very well.

Figure 3 provides graphs of the 0.95 asymptotic quantiles of the $|t|$ and QLR statistics concerning β and π as a function of b for $\pi_0 = 1.5$, 2.0, 3.0, and 3.5. For the $|t|$ statistic concerning β, 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 tests and CIs that employ the $|t|$ statistic for β and the standard critical value (based on the normal distribution) have incorrect size. For the QLR statistic for β, the graphs slightly exceed the 0.95 quantile under strong identification when b is 0 or almost 0 and fall below the 0.95 quantile under strong identification for other small to medium b values. The graphs in Figure 3(b) imply that tests and CIs that employ the QLR statistic for β and the standard critical value (based on the χ_1^2 distribution) have small size distortions as a result of the undercoverage for b values close to 0. Given the heights of the graphs in Figures 3(c) and 3(d), tests and CIs that employ the $|t|$ statistic for π have correct asymptotic size when $\pi_0 = 1.5$ and 2.0 and have slight size distortion when $\pi_0 = 3.0$ and 3.5, whereas those that employ the QLR statistic for π always have correct asymptotic size.

Figure 4 reports the asymptotic and finite-sample coverage probabilities (CPs) of nominal 0.95 standard $|t|$ and QLR CIs for β and π when $\pi_0 = 1.5$. For example, the smallest asymptotic and finite-sample CPs (over b) are around 0.68 and 0.93 for the $|t|$ and QLR CIs for β, respectively. There is no size distortion for the $|t|$ and QLR CIs for π. Note that the asymptotic CPs provide a good approximation to the finite-sample CPs. Figure S-7 in Supplemental Appendix E provides analogous results for $\pi_0 = 3.0$.

Next, we consider CIs that are robust to weak identification. For the robust CI for β, we impose the null value of $b = n^{1/2}\beta_0$, where β_0 is the true value of

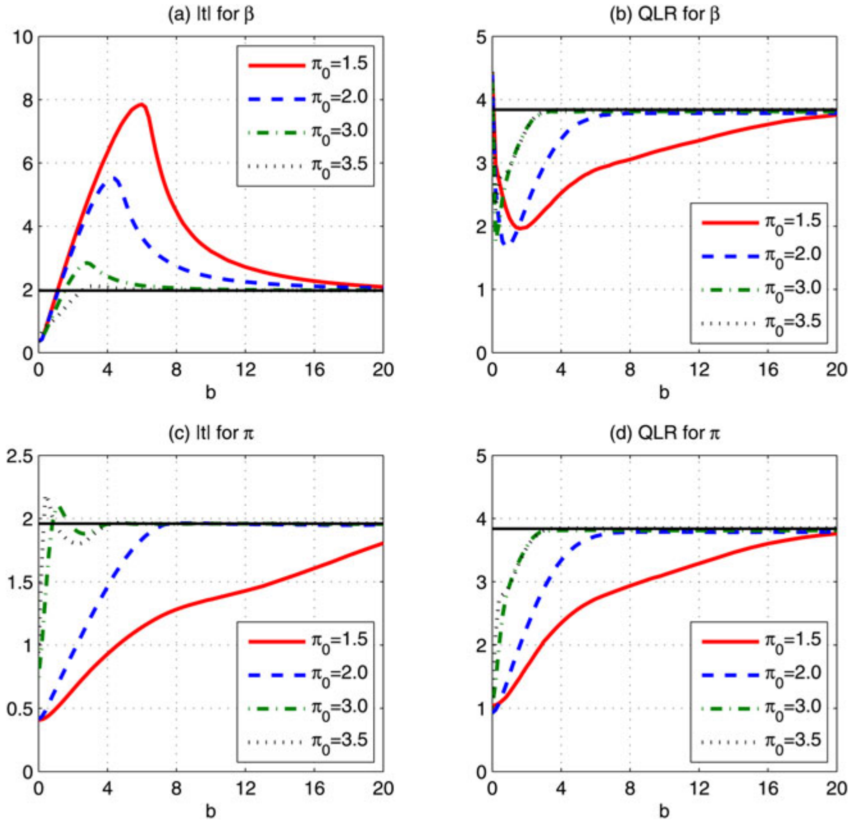

FIGURE 3. Asymptotic 0.95 quantiles of the $|t|$ and QLR statistics for tests concerning β and π in the nonlinear regression model with endogeneity.

β under the null. With the knowledge of b under the null, no ICS procedure is needed. Imposing the null value of b also results in a smaller LF critical value. As indicated in Figure 3(a), the NI-LF critical values for the $|t|$ CI for β are attained at $\pi_0 = 1.5$ for all b values. In consequence, the robust $|t|$ CI for β is asymptotically similar when $\pi_0 = 1.5$, as shown in Figure 5(a). Figure 5(a) also reports the finite-sample ($n = 500$) CPs of the robust $|t|$ CI for β. The smallest and largest finite-sample CPs are around 0.91 and 0.97, as opposed to 0.68 and 1.00 for the standard $|t|$ CI. Figure 5(b) shows that the robust QLR CI for β tends to overcover for a range of small to medium b values, but the asymptotic size is correct. Figures S-8(a) and S-8(b) in Supplemental Appendix E provide analogous results for $\pi_0 = 3.0$. The robust CIs for β are not asymptotically similar when $\pi_0 = 3.0$, but they have correct asymptotic size, and the asymptotic and finite-sample CPs are close for all b values.

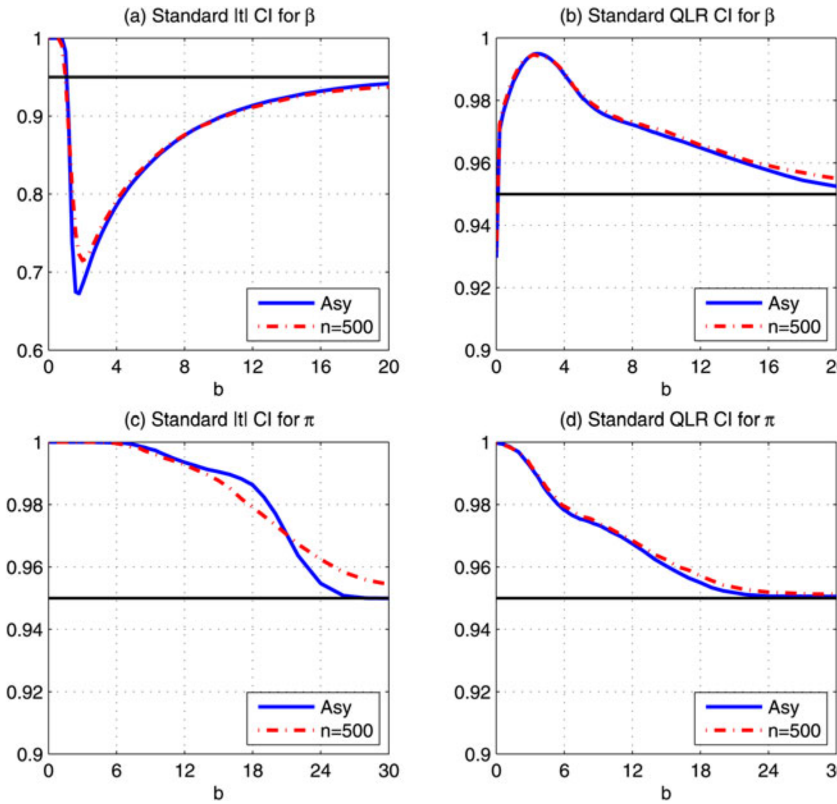

FIGURE 4. Coverage probabilities of standard $|t|$ and QLR CIs for β and π in the nonlinear regression model with endogeneity when $\pi_0 = 1.5$.

The robust CIs for π are constructed with the null value π_0 imposed. When $\pi_0 = 1.5$, the robust $|t|$ and QLR CIs are the same as the standard $|t|$ and QLR CIs, respectively, because the NI-LF critical values equal the standard critical values in both cases. In consequence, Figures 5(c) and 5(d) are the same as Figures 4(c) and 4(d), respectively. The robust $|t|$ and QLR CIs for π when $\pi_0 = 3.0$ are reported in Figures S-8(c) and S-8(d) in Supplemental Appendix E. In this case, the NI-LF critical value for the robust $|t|$ CI for π is slightly larger than the standard critical value, as shown in Figure 3(c). We apply the smooth transition in (5.33) to obtain critical values for the robust $|t|$ CI for π, where the transition function is $s(x) = \exp(-2x)$ and the constants are $\kappa = 1.5$ and $D = 1$. The choices of $s(x)$ and D were determined via some experimentation to be good choices in terms of yielding CPs that are relatively close to the nominal size 0.95 across different values of b. A wide range of κ values yield similar results (because the constants Δ_1 and Δ_2 adjust to maintain correct asymptotic size as κ is changed).

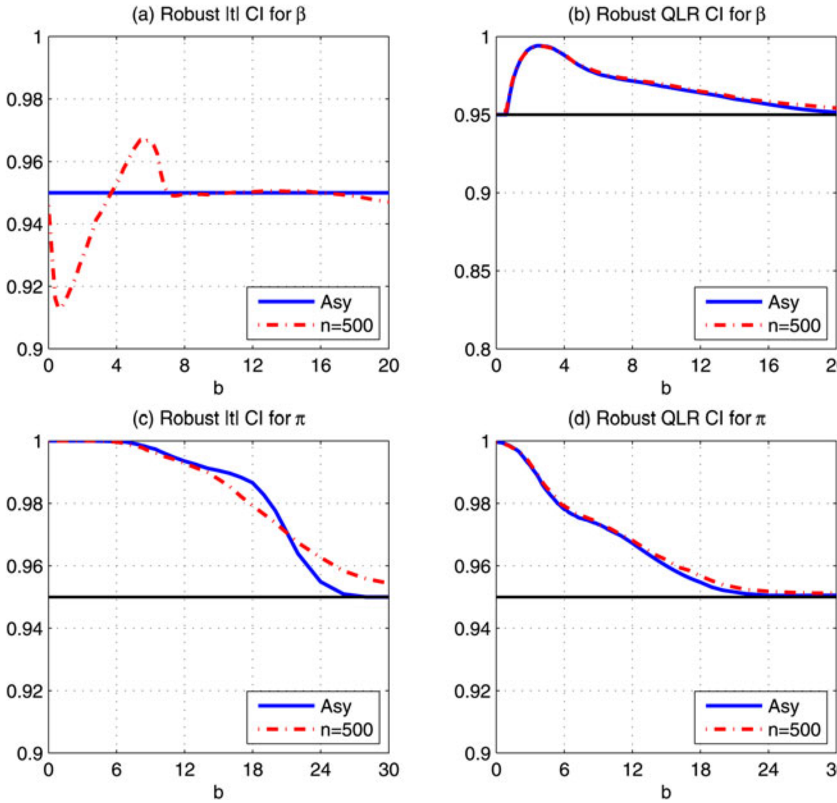

FIGURE 5. Coverage probabilities of robust $|t|$ and QLR CIs for β and π in the nonlinear regression model with endogeneity when $\pi_0 = 1.5$. No smooth transition is employed.

Figures S-7(c) and S-8(c) show that, when $\pi_0 = 3.0$, the standard $|t|$ CI for π suffers from size distortion but the robust $|t|$ CI for π has correct asymptotic size. When $\pi_0 = 3.0$, the robust QLR CI for π is the same the standard QLR CI for π, as shown in Figures S-7(d) and S-8(d).

Besides b and π_0, the construction of a robust CI also requires the ζ value to obtain the LF (or NI-LF) critical value through simulation. In this model, $\zeta = (\zeta_1, \zeta_2)'$. Because ζ can be consistently estimated, we recommend plugging in the estimator $\widehat{\zeta}_n$ in place of ζ_0 in practice. To ease the computational burden required to simulate the CPs, the finite-sample CPs of the robust CIs reported in Figures 5 and S-8 are constructed using the true value ζ_0, rather than the estimated value $\widehat{\zeta}_n$.[26] However, the difference between the robust CIs constructed with $\widehat{\zeta}_n$ and ζ_0 typically is relatively minor. A comparison is reported in Table S-1 of AC2 in the context of a STAR model.

NOTES

1. Throughout the paper, we use the term *identification/lack of identification* in the sense of identification by a GMM or MD criterion function $Q_n(\theta)$. Lack of identification by $Q_n(\theta)$ means that $Q_n(\theta)$ is flat in some directions in part of the parameter space. See Assumption GMM1(i) in Section 3.1 for a precise definition. 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 often a close relationship.

2. For references concerning results for these models, see AC1.

3. That is, the metric satisfies the following 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 for all $i, m \geq 1$. For example, in an i.i.d. situation, the metric on Φ^* can be the uniform metric on the distribution of W_i. In a stationary time series context, it can be the supremum over $m \geq 1$ of the uniform metric on the space of distributions of the vectors (W_i, W_{i+m}). 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 $j = 1, 2$ and d_{Φ^*} is the metric on Φ^*.

4. The $o(n^{-1})$ term in (2.4), and in (4.1) and (4.2), is a fixed sequence of constants that does not depend on the true parameter $\gamma \in \Gamma$ and does not depend on π in (4.1).

5. Note that $\mathcal{Z}$ and $\mathcal{Z}^*$ are not related to the support of Z_i. Rather, they are the optimization and true parameter spaces for ζ, which has dimension $2d_X$.

6. This follows from the combination of Lemma 10.1 in Supplemental Appendix A and Lemma 3.1 of AC1.

7. The matrix $B(\beta)$ is defined differently in the scalar and vector β cases because in the scalar case the use of β, rather than $||\beta||$, produces noticeably simpler (but equivalent) formulas, but in the vector case $||\beta||$ is required.

8. The constant $\delta > 0$ is as in Assumption B2(iii) stated in Section 3.7. The set Γ_0 is not empty by Assumption B2(ii).

9. The sufficient conditions are for Assumption C5 of AC1, which is the same as Assumption GMM3(iv) but with $m(W_i, \theta)$ of AC1 in place of $g(W_i, \theta)$.

10. In the vector β case, $J_g(\gamma_0)$ may depend on ω_0 in addition to γ_0.

11. The functions $J(\theta^+; \gamma_0)$ and $V(\theta^+; \gamma_0)$ do not depend on ω_0, only γ_0.

12. This holds because $\eta_n(\theta) = -\left(\frac{n^{1/2}}{||(-2\pi, \beta+1)||}\right)[-2\pi(\beta_n + 1)(\pi - \pi_n) + (\beta_n + 1)(\pi^2 - \pi_n^2) + (\beta - \beta_n)(\pi^2 - \pi_n^2)] = \left(\frac{n^{1/2}}{||(-2\pi, \beta+1)||}\right)[(\beta_n + 1)(\pi - \pi_n)^2 - (\beta - \beta_n)(\pi^2 - \pi_n^2)]$. Hence, $\eta_n(\widehat{\theta}_n)||(-2\widehat{\pi}_n, \widehat{\beta}_n + 1)|| = n^{1/2}(\widehat{\pi}_n - \pi_n)^2(1 + o(1)) - n^{1/2}(\widehat{\beta}_n - \beta_n)(\widehat{\pi}_n^2 - \pi_n^2) = [n^{1/2}\beta_n(\widehat{\pi}_n - \pi_n)]^2(n^{1/4}\beta_n)^{-2}(1 + o(1)) + O_p(1)$ using Theorem 4.1(a) or 4.2(a). (The $O_p(1)$ term is $o_p(1)$ if $|b| = \infty$.) Because $||(-2\widehat{\pi}_n, \widehat{\beta}_n + 1)|| \to_p ||(-2\pi_0, 1)|| < \infty$, the claim follows.

13. When $|b| = \infty$, this holds because $n^{1/2}\beta_n(\widehat{\pi}_n - \pi_n)$ has an asymptotic normal distribution by Theorem 4.2(a). When $|b| < \infty$, this holds because $[n^{1/2}\beta_n(\widehat{\pi}_n - \pi_n)]^2(n^{1/4}\beta_n)^{-2} = n^{1/2}(\widehat{\pi}_n - \pi_n)^2$, $\widehat{\pi}_n \to_d \pi^*(\gamma_0, b)$ by Theorem 4.1(a), and $P(\pi^*(\gamma_0, b) = 0) = 0$.

14. By allowing $v_{n,0}^{null}$ to depend on n, we obtain results for drifting null values. For example, if $r(\theta) = \beta$, this provides results when the null and local alternative values of β are $n^{-1/2}$-local to zero. This is useful for obtaining asymptotic false coverage probabilities of CSs for β when the true value of β is close to zero. In this case, the relevant null values also are close to zero, in an $n^{-1/2}$-local to zero sense.

15. Under these conditions on $r(\theta)$, one can take $A(\theta) = I_{d_r}$.

16. To be precise, let $H(v) = \{h = (b, \gamma_0) \in H : ||b|| < \infty, r(\theta_0) = v\}$, 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 $c_{W,1-\alpha}^{LF}(v)$, is defined by replacing H by $H(v)$ in (5.28) when the null hypothesis value is $r(\theta_0) = v$. Note that v takes values in the set $V_r = \{v_0 : r(\theta_0) = v_0 \text{ for some } h = (b, \gamma_0) \in H\}$. 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 in the text for NI-LF CSs and NI robust CSs hold in this case.

17. For example, if ζ is consistently estimated by $\widehat{\zeta}_n$, then H is replaced by $\widehat{H}_n = \{h = (b, \gamma) \in H : \gamma = (\beta, \widehat{\zeta}_n, \pi, \phi)\}$. If a plug-in NI-LF critical value is employed, $H(\upsilon)$ is replaced by $H(\upsilon) \cap \widehat{H}_n$, where $H(\upsilon)$ is defined in note 16. The parameter b is not consistently estimable, and so it cannot be replaced by a consistent estimator.

18. When β is specified by the null hypothesis, it is not necesary to use an ICS procedure. Instead, we recommend using a (possibly plug-in) NI-LF critical value, see note 17.

19. If $c^{LF}_{W,1-\alpha} = \infty$, $s(x)$ should be taken to equal 0 for x sufficiently large, where $\infty \times 0$ equals 0 in (5.30). Then, the critical value $\widehat{c}_{W,1-\alpha,n}$ is infinite if A_n is small and is finite if A_n is sufficiently large.

20. The convergence in distribution follows from Theorem 4.1(a) and Assumption V1. In the vector β case, $\Sigma^{-1}_{\beta\beta}(\pi^*; \gamma_0)$ is replaced in (5.31) by a slightly different expresssion; see note 51 of AC1. When the type 2 robust critical value is considered in the vector β case, h is defined to include $\omega_0 = \lim_{n\to\infty} \beta_n/||\beta_n|| \in R^{d_\beta}$ as an element, i.e., $h = (b, \gamma_0, \omega_0)$ and $H = \{h = (b, \gamma_0, \omega_0) : ||b|| < \infty, \gamma_0 \in \Gamma$ with $\beta_0 = 0, ||\omega_0|| = 1\}$ because the true value ω_0 affects the asymptotic distribution of A_n. Alternatively to the ICS statistic A_n, one can use an NI-ICS statistic $A_n(\upsilon)$, which employs the restricted estimator $\widetilde{\beta}_n(\upsilon)$ of β in place of $\widehat{\beta}_n$ and a different weight matrix. See AC1 for details.

21. When $NRP(0, 0; h) > \alpha$, a unique solution $\Delta_1(h)$ typically exists because $NRP(\Delta_1, 0; h)$ is always non-increasing in Δ_1 and is typically strictly decreasing and continuous in Δ_1. If no exact solution to $NRP(\Delta_1(h), 0; h) = \alpha$ exists, then $\Delta_1(h)$ is taken to be any value for which $NRP(\Delta_1(h), 0; h) \le \alpha$ and $\Delta_1(h) \ge 0$ is as small as possible. Analogous comments apply to the equation $NRP(\Delta_1, \Delta_2(h); h) = \alpha$ and the definition of $\Delta_2(h)$. When the LF critical value is achieved at $||b|| = \infty$, i.e., $\chi^2_{d_r,1-\alpha} \ge \sup_{h\in H} c_{QLR,1-\alpha}(h)$, the standard asymptotic critical value $\chi^2_{d_r,1-\alpha}$ 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 $\sup_{h\in H} c_{QLR,1-\alpha}(h)$ is not attained at any point $h_{\max}$, then $b_{\max}$ can be taken to be any point such that $c_{QLR,1-\alpha}(h_{\max})$ is arbitrarily close to $\sup_{h\in H} c_{QLR,1-\alpha}(h)$ for some $h_{\max} = (b_{\max}, \gamma_{\max}) \in H$.

22. When b is close to 0, the FCP may be larger than 0.50 for all admissible υ because of weak identification. In such cases, $\upsilon_{H_0}(h)$ is taken to be the admissible value that minimizes the FCP for the selected value of κ that is being used to obtain $\upsilon_{H_0}(h)$.

23. When $r(\theta) = \pi$, 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.

24. The absolute value of x is employed in $h(x, \pi)$ to guarantee $h(x, \pi) \in R$ when π is not an integer. With the DGP specified in the text, $X_{1,i}$ is positive with probability close to 1. Hence, $h(X_{1,i}, \pi)$ is approximately the Box–Cox transformation of $X_{1,i}$.

25. The discrete values of b for which computations are made run from 0 to 30, with a grid of 0.2 for b between 0 and 10, a grid of 1 for b between 10 and 20, and a grid of 2 for b between 20 and 30.

26. With a single sample, the computational burden is the same whether the true value ζ_0 or the estimated value $\widehat{\zeta}_n$ is employed. However, in a simulation study, it is much faster to simulate the critical values for a range of true values of b and π_0 and the single true value of ζ_0 one time and then use them in each of the simulation repetitions, rather than to simulate a new critical value for each simulation repetition, which is required if $\widehat{\zeta}_n$ is employed.

REFERENCES

Amemiya, T. (1974) Multivariate regression and simultaneous-equation models when the dependent variables are truncated normal. *Econometrica* 42, 999–1012.

Andrews, D.W.K. (2002) Generalized method of moments estimation when a parameter is on a boundary. *Journal of Business & Economic Statistics* 20, 530–544.

Andrews, D.W.K. & X. Cheng (2012a) Estimation and inference with weak, semi-strong, and strong identification. *Econometrica* 80, 2153–2211.

Andrews, D.W.K. & X. Cheng (2012b) Supplemental material for "Estimation and inference with weak, semi-strong, and strong identification." Econometric Society website, http://www.econometricsociety.org/ecta/Supmat/9456_miscellaneous.pdf.

Andrews, D.W.K. & X. Cheng (2013a) Maximum likelihood estimation and uniform inference with sporadic identification failure. *Journal of Econometrics* 173, 36–56.

Andrews, D.W.K. & X. Cheng (2013b) Supplementary material for "Generalized method of moments estimation and uniform subvector inference with possible identification failure."

Andrews, D.W.K., X. Cheng, & P. Guggenberger (2009) Generic Results for Establishing the Asymptotic Size of Confidence Sets and Tests. Cowles Foundation Discussion Paper Number 1813, Yale University.

Andrews, D.W.K. & P. Guggenberger (2009) Validity of subsampling and "plug-in asymptotic" inference for parameters defined by moment inequalities. *Econometric Theory* 25, 669–709.

Andrews, D.W.K. & P. Guggenberger (2010) Asymptotic size and a problem with subsampling and with the m out of n bootstrap. *Econometric Theory* 26, 426–468.

Andrews, I. & A. Mikusheva (2011) Maximum Likelihood Inference in Weakly Identified DSGE Models. Manuscript, Department of Economics, MIT.

Andrews, I. & A. Mikusheva (2012) A Geometric Approach to Weakly Identified Econometric Models. Manuscript, Department of Economics, MIT.

Antoine, B. & E. Renault (2009) Efficient GMM with nearly weak instruments. *Econometrics Journal* 12, S135–S171.

Antoine, B. & E. Renault (2010) Efficient inference with poor instruments, a general framework. In D. Giles & A. Ullah (eds.), *Handbook of Empirical Economics and Finance*, Chap. 2, pp. 29–70. Taylor and Francis.

Areosa, W.D., M. McAleer, & M.C. Medeiros (2011) Moment-based estimation of smooth transition regression models with endogenous variables. *Journal of Econometrics* 165, 100–111.

Armstrong, T.B., H. Hong, & D. Nekipelov (2012) How Strong Must Identification Be for Conventional Asymptotics in Nonlinear Models? Manuscript, Cowles Foundation, Yale University.

Caner, M. (2010) Testing, estimation in GMM and CUE with nearly weak identification. *Econometric Reviews* 29, 330–363.

Cheng, X. (2008) Robust Confidence Intervals in Nonlinear Regression under Weak Identification. Working Paper, Department of Economics, Yale University.

Choi, I. & 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.

Davies, R.B. (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64, 247–254.

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

Guggenberger, P., F. Kleibergen, S. Mavroeidis, & L. Chen (2012) On the asymptotic sizes of subset Anderson-Rubin and Lagrange multiplier tests in linear instrumental variables regression. *Econometrica* 80, 2649–2666.

Hansen, L.P. (1982) Large sample properties of generalized method of moments estimation. *Econometrica* 50, 1029–1054.

Heckman, J.J. (1978) Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46, 931–959.

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

Kleibergen, F. (2005) Testing parameters in GMM without assuming that they are identified. *Econometrica* 73, 1103–1123.

Lee, L.F. (1981) Simultaneous equations models with discrete endogenous variables. In C.F. Manski & D. McFadden (eds.), *Structural Analysis of Discrete Data and Econometric Applications*, Chap. 9, pp. 346–364. MIT Press.

Lee, L.-F. & A. Chesher (1986) Specification testing when score test statistics are identically zero. *Journal of Econometrics* 31, 121–149.

Ma, J. & C.R. Nelson (2008) Valid Inference for a Class of Models where Standard Inference Performs Poorly; Including Nonlinear Regression, ARMA, GARCH, and Unobserved Components. Manuscript, Department of Economics, University of Washington.

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

Nelson, C.R. & R. Startz (1990) Some further results on the exact small sample properties of the instrumental variables estimator. *Econometrica* 58, 967–976.

Nelson, C.R. & R. Startz (2007) The zero-information-limit condition and spurious inference in weakly identified models. *Journal of Econometrics* 138, 47–62.

Nelson, F. & L. Olson (1978) Specification and estimation of a simultaneous-equation model with limited dependent variables. *International Economic Review* 19, 695–709.

Pakes, A. & D. Pollard (1989) Simulation and the asymptotics of optimization estimators. *Econometrica* 57, 1027–1057.

Park, J.Y. & P.C.B. Phillips (1988) Statistical inference in regressions with integrated processes: Part 1. *Econometric Theory* 4, 468–497.

Phillips, P.C.B. (1989) Partially identified econometric models. *Econometric Theory* 5, 181–240.

Qu, Z. (2011) Inference and Specification Testing in DSGE Models with Possible Weak Identification. Working paper, Department of Economics, Boston University.

Rivers, D. & Q.H. Vuong (1988) Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics* 39, 347–366.

Rotnitzky, A., D.R. Cox, M. Bottai, & J. Robins (2000) Likelihood-based inference with singular information matrix. *Bernoulli* 6, 243–284.

Sargan, J.D. (1983) Identification and lack of identification. *Econometrica* 51, 1605–1633.

Schorfheide, F. (2011) Estimation and Evaluation of DSGE Models: Progress and Challenges. NBER Working paper 16781.

Shi, X. & P.C.B. Phillips (2011) Nonlinear cointegrating regression under weak identification. *Econometric Theory* 28, 1–39.

Smith, R.J. & R.W. Blundell (1986) An exogeneity test for a simultaneous equation tobit model with an application to labor supply. *Econometrica* 54, 679–685.

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

Stock, J.H. & J.H. Wright (2000) GMM with weak instruments. *Econometrica* 68, 1055–1096.

Stock, J.H. & M. Yogo (2005) Testing for weak instruments in linear IV regression. In D.W.K. Andrews & J.H. Stock (eds.), *Identification and Inference for Econometric Models: A Festschrift in Honor of Thomas J. Rothenberg*, Chap. 5, pp. 80–108. Cambridge University Press.

Wright, J.H. (2003) Detecting lack of identification in GMM. *Econometric Theory* 19, 322–330.