# MAXIMUM LIKELIHOOD ESTIMATION AND UNIFORM INFERENCE WITH SPORADIC IDENTIFICATION FAILURE

**By**

**Donald W. K. Andrews and Xu Cheng**

**October 2011**
**Revised October 2012**

**COWLES FOUNDATION DISCUSSION PAPER NO. 1824R**


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

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

# Maximum Likelihood Estimation and Uniform Inference with Sporadic Identification Failure

Donald W. K. Andrews*
Cowles Foundation
Yale University

Xu Cheng
Department of Economics
University of Pennsylvania

First Draft: August, 2007
Revised: November 3, 2012

*The first author gratefully acknowledges the research support of the National Science Foundation via grant numbers SES-0751517 and SES-1058376. The authors thank Xiaohong Chen, Sukjin Han, Yuichi Kitamura, Peter Phillips, Eric Renault, Frank Schorfheide, and Ed Vytlacil for helpful comments.

**Abstract**

This paper analyzes the properties of a class of estimators, tests, and confidence sets (CS's) when the parameters are not identified in parts of the parameter space. Specifically, we consider estimator criterion functions that are sample averages and are smooth functions of a parameter $\theta$. This includes log likelihood, quasi-log likelihood, and least squares criterion functions.

We determine the asymptotic distributions of estimators under lack of identification and under weak, semi-strong, and strong identification. We determine the asymptotic size (in a uniform sense) of standard $t$ and quasi-likelihood ratio (QLR) tests and CS's. We provide methods of constructing QLR tests and CS's that are robust to the strength of identification.

The results are applied to two examples: a nonlinear binary choice model and the smooth transition threshold autoregressive (STAR) model.

*Keywords*: Asymptotic size, binary choice, confidence set, estimator, identification, likelihood, nonlinear models, test, smooth transition threshold autoregression, weak identification.

*JEL Classification Numbers*: C12, C15.

# 1. Introduction

This paper provides a set of maximum likelihood (ML) regularity conditions under which the asymptotic properties of ML estimators and corresponding $t$ and QLR tests and confidence sets (CS's) are obtained. The novel feature of the conditions is that they allow the information matrix to be singular in parts of the parameter space. In consequence, the parameter vector is unidentified and weakly identified in some parts of the parameter space, while semi-strongly and strongly identified in other parts. The conditions maintain the usual assumption that the log-likelihood satisfies a stochastic quadratic expansion. The results also apply to quasi-log likelihood and nonlinear least squares procedures.

Compared to standard asymptotic results in the literature for ML estimators, tests, and CS's, the results given here cover both fixed and drifting sequences of true parameters. The latter are necessary to treat cases of weak identification and semi-strong identification. In particular, they are necessary to determine the asymptotic sizes of tests and CS's (in a uniform sense).

This paper is a sequel to Andrews and Cheng (2012a) (AC1). The method of establishing the results outlined above and in the Abstract is to provide a set of sufficient conditions for the high-level conditions of AC1 for estimators, tests, and CS's that are based on smooth sample-average criterion functions. The high-level conditions in AC1 involve the behavior of the estimator criterion function under certain drifting sequences of distributions. In contrast, the assumptions given here are much more primitive. They only involve mixing, smoothness, and moment conditions, plus conditions on the parameter space.

The paper considers models in which the parameter $\theta$ of interest is of the form $\theta = (\beta, \zeta, \pi)$, where $\pi$ is identified if and only if $\beta \neq 0$, $\zeta$ is not related to the identification of $\pi$, and $\psi = (\beta, \zeta)$ is always identified. For examples, the nonlinear binary choice model is of the form: $Y_i = 1(Y_i^* > 0)$ and $Y_i^* = \beta \cdot h(X_i, \pi) + Z_i'\zeta - U_i$, where $(Y_i, X_i, Z_i)$ is observed and $h(\cdot, \cdot)$ is a known function. The STAR model is of the form: $Y_t = \zeta_1 + \zeta_2 Y_{t-1} + \beta \cdot m(Y_{t-1}, \pi) + U_t$, where $Y_t$ is observed and $m(\cdot, \cdot)$ is a known function.

In general, the parameters $\beta$, $\zeta$, and $\pi$ may be scalars or vectors. We determine the asymptotic properties of ML estimators, tests, and CS's under drifting sequences of parameters/distributions. Suppose the true value of the parameter is $\theta_n = (\beta_n, \zeta_n, \pi_n)$ for $n \geq 1$, where $n$ indexes the sample size. The behavior of ML estimators and test

statistics 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): $\beta_n = 0 \ \forall n \geq 1$, $\pi$ is unidentified; Category I(b): $\beta_n \neq 0$ and $n^{1/2}\beta_n \to b \in R^{d_\beta}$, $\pi$ is weakly identified; Category II: $\beta_n \to 0$ and $n^{1/2}||\beta_n|| \to \infty$, $\pi$ is semi-strongly identified; and Category III: $\beta_n \to \beta_0 \neq 0$, $\pi$ is strongly identified.

For Category I sequences, we obtain the following results: the estimator of $\pi$ is inconsistent, the estimator of $\psi = (\beta, \zeta)$ and the $t$ and QLR test statistics have non-standard asymptotic distributions, and the standard tests and CS's (that employ standard normal or $\chi^2$ critical values) have asymptotic null rejection probabilities and coverage probabilities that may or may not be correct depending on the model.[1] (In many cases, they are not correct). For Category II sequences, estimators and standard tests and CS's are found to have standard asymptotic properties, but the rate of convergence of the estimator of $\pi$ is less than $n^{1/2}$. Specifically, the estimators are asymptotically normal and the test statistics have asymptotic chi-squared distributions. For Category III sequences, the estimators and standard tests and CS's have standard asymptotic properties and the estimators converge at rate $n^{1/2}$.

We also consider $t$ and QLR tests and CS's that are robust to the strength of identification. These procedures use different critical values from the standard ones. First, we consider critical values based on asymptotically least-favorable sequences of distributions. Next, we consider data-dependent critical values that employ an identification-category selection procedure that determines whether $\beta$ is near the value 0 that yields lack of identification of $\pi$, and if it is, the critical value is adjusted (in a smooth way) to take account of the lack of identification or weak identification. We show that the robust procedures have correct asymptotic size (in a uniform sense). The data-dependent robust critical values yield more powerful tests than the least favorable critical values.

In the numerical results for the STAR and nonlinear binary choice models, $\pi$ is taken to be a scalar to ease computation. In the STAR model, the transition parameter is fixed, as in the empirical work in Lundbergh and Teräsvirta (2006), and the unknown parameter $\pi$ is the threshold parameter. The numerical results in both models are summarized as follows. The asymptotic distributions of the estimators of $\beta$ and $\pi$ are far from the normal distribution under weak identification and lack of identification. The asymptotic distributions range from being strongly bimodal, to being close to uniform, to

[1] Here, by "correct" we mean $\alpha$ or less for tests and $1-\alpha$ or greater for CS's, where $\alpha$ and $1-\alpha$ are the nominal sizes of the tests or CS's.

being extremely peaked. The asymptotics provide remarkably accurate approximations to the finite-sample distributions.

In the STAR model, the standard $t$ and QLR confidence intervals (CI's) for $\beta$ have substantial asymptotic size distortions with asymptotic sizes equaling .56 and .72, respectively, for nominal .95 CI's. This is also true for the $t$ and QLR CI's for $\pi$, where the asymptotic sizes are .40 and .84, respectively. Note that the size distortions are noticeably larger for the standard $t$ than QLR CI. In the binary choice model, the standard $t$ and QLR CI's for $\beta$ have incorrect asymptotic sizes: .68 versus .92, respectively, for nominal .95 CI's. However, the standard $t$ and QLR CI's for $\pi$ have small and no size distortion, respectively. In both models, the asymptotic sizes provide very good approximations to the finite-sample sizes for the cases considered.

In both models, the robust CI's have correct asymptotic sizes and finite-sample sizes that are quite close to the asymptotic size for the QLR CI's and fairly close for the $t$ CI's. (As mentioned above, for the STAR model, these results are for the case of a fixed transition parameter.)

In sum, the numerical results indicate that the asymptotic results of the paper are quite useful in determining the finite-sample behavior of estimators and standard tests and CI's under weak identification and lack of identification. They are also quite useful in designing robust tests and CI's whose finite-sample size is close to their nominal size.

The results of this paper apply when the criterion function satisfies a stochastic quadratic expansion in the parameter $\theta$. This rules out a number of interesting models that exhibit lack of identification in parts of the parameter space, including regime switching models, mixture models, abrupt transition structural change models, and abrupt transition threshold autoregressive models, such as in Hansen (2000).[2]

Now, we briefly discuss the literature related to this paper. See AC1 for a more detailed discussion. The following are companion papers to this one: AC1, Andrews and Cheng (2012b) (AC1-SM), and Andrews and Cheng (2011a) (AC3). 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. AC3 provides results for estimators and tests based on generalized method of moments (GMM) criterion functions. It provides applications to an endogenous nonlinear regression model and an endogenous binary choice model.

Cheng (2008) provides results for a nonlinear regression model with multiple sources

[2]See AC1 for other references concerning results for these models.

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 $\pi$ only appears under the alternative. Such tests have been considered in the literature starting from Davies (1977). The results of this paper cover tests of this sort, as well as tests for a whole range of linear and nonlinear hypotheses that involve $(\beta, \zeta, \pi)$ and corresponding CS's.

The weak instrument (IV) literature is closely related to this paper. However, papers in that literature focus on criterion functions that are indexed by parameters that do not determine the strength of identification. In contrast, in this paper, the parameter $\beta$, which determines the strength of identification of $\pi$, appears as one of parameters in the criterion function. Selected papers from the weak IV literature include Nelson and Startz (1990), Dufour (1997), Staiger and Stock (1997), Stock and Wright (2000), Kleibergen (2002, 2005), and Moreira (2003).

Andrews and Mikusheva (2011) and Qu (2011) consider Lagrange multiplier (LM) tests in a maximum likelihood context where identification may fail, with emphasis on dynamic stochastic general equilibrium models. The results of the present paper apply to $t$ and QLR statistics, but not to LM statistics. The consideration of LM statistics is in progress. Andrews and Mikusheva (2012) consider Anderson-Rubin-type tests based on minimum distance statistics for models with weak identification.

Antoine and Renault (2009, 2010) and Caner (2010) consider GMM estimation with IV's that lie in the semi-strong category, using our terminology. Nelson and Startz (2007) and Ma and Nelson (2008) analyze models like those considered in this paper. However, they do not provide asymptotic results or robust tests and CS's of the type given in this paper. Sargan (1983), Phillips (1989), and Choi and Phillips (1992) provide finite-sample and asymptotic results for linear simultaneous equations models when some parameters are not identified. Phillips and Shi (2012) provide results for a nonlinear regression model with non-stationary regressors in which identification may fail.

The remainder of the paper is organized as follows. Section 2 introduces the smooth sample average extremum estimators, criterion functions, tests, CS's, and drifting sequences of distributions considered in the paper. Section 3 states the assumptions employed. Section 4 provides the asymptotic results for the extremum estimators. Section 5 establishes the asymptotic distributions of QLR statistics, determines the asymptotic size of standard QLR CS's, and introduces robust QLR tests and CS's, whose asymp-

totic size is equal to their nominal size. Section 6 considers $t$-based CS's. The nonlinear binary choice model is used as a running example in the previous sections. Section 7 provides results for the smooth transition threshold autoregressive model (STAR) model. Section 8 provides numerical results for the STAR and binary choice models. An Appendix provides some additional material. Five Supplemental Appendices to this paper are given in Andrews and Cheng (2011b). Supplemental Appendix A provides proofs of the results given in the paper. Supplemental Appendix B provides some miscellaneous results. Supplemental Appendix C provides additional numerical results for the nonlinear binary choice and STAR models. Supplemental Appendices D and E verify the assumptions for the nonlinear binary choice model and the STAR model, respectively.

All limits below are taken "as $n \to \infty$." 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 $d_\beta$-vector of zeros, where $d_\beta$ is the dimension of a parameter $\beta$. Let $R_{[\pm\infty]} = R \cup \{\pm\infty\}$. Let $R^p_{[\pm\infty]} = R_{[\pm\infty]} \times ... \times R_{[\pm\infty]}$ with $p$ copies. Let $\Rightarrow$ denote weak convergence of a sequence of stochastic processes indexed by $\pi \in \Pi$ for some space $\Pi$.[3]

# 2. Estimator and Criterion Function

## 2.1. Smooth Sample Average Estimators

We consider an extremum estimator $\widehat{\theta}_n$ that is defined by minimizing a sample criterion function of the form

$$Q_n(\theta) = n^{-1} \sum_{i=1}^{n} \rho\left(W_i, \theta\right), \tag{2.1}$$

where $\{W_i : i \leq n\}$ are the observations and $\rho(w, \theta)$ is a known function that is twice continuously differentiable in $\theta$. This includes ML and LS estimators. The observations $\{W_i : i \leq n\}$ may be i.i.d. or strictly stationary. Formal assumptions are provided in Section 3 below.

[3] In the definition of weak convergence, we employ the uniform metric $d$ on the space $\mathcal{E}_v$ of $R^v$-valued functions on $\Pi$. See the Outline of the Supplemental Appendix of AC1 for more details.

The paper considers the case where $\theta$ is not identified (by the criterion function $Q_n(\theta)$) at some points in the parameter space. Lack of identification occurs when the $Q_n(\theta)$ is flat wrt some sub-vector of $\theta$. To model this identification problem, $\theta$ is partitioned into three sub-vectors:

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

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 effect the identification of $\pi$. These conditions allow for a wide range of cases, including cases in which reparameterization is used to convert a model into the framework considered here.

**Example 1.** This example is the nonlinear binary choice model

$$Y_i = 1(Y_i^* > 0) \text{ and } Y_i^* = \beta \cdot h(X_i, \pi) + Z_i'\zeta - U_i, \tag{2.3}$$

where $h(X_i, \pi) \in R$ is known up to the finite-dimensional parameter $\pi \in R^{d_\pi}$. Suppose $h(x, \pi)$ is twice continuously differentiable wrt $\pi$ for any $x$ in the support of $X_i$ and the first- and second-order partial derivatives are denoted by $h_\pi(x, \pi)$ and $h_{\pi\pi}(x, \pi)$, respectively.

The observed variables are $\{W_i = (Y_i, X_i, Z_i) : i = 1, ..., n\}$. The random variables $\{(X_i, Z_i, U_i) : i = 1, ..., n\}$ are i.i.d. The distribution of $(X_i, Z_i)$ is $\phi$, which is an infinite-dimensional nuisance parameter. The parameter of interest is $\theta = (\beta, \zeta, \pi)$. Conditional on $(X_i, Z_i)$, the distribution function (df) of $U_i$ is $L(u)$. The df $L(u)$ is known and does not depend on $(X_i, Z_i)$. For example, $L(u)$ is the standard normal distribution df in a probit model and the logistic df in a logit model. We assume that $L(u)$ is twice continuously differentiable and its first- and second-order derivatives are denoted by $L'(u)$ and $L''(u)$, respectively. Suppose $L'(u) > 0$ and $0 < L(u) < 1$ $\forall u \in R$.

In this model,

$$\begin{aligned} P(Y_i = 1|X_i, Z_i) &= P(U_i < \beta h(X_i, \pi) + Z_i'\zeta | X_i, Z_i) = L(g_i(\theta)), \text{ where} \\ g_i(\theta) &= \beta h(X_i, \pi) + Z_i'\zeta. \end{aligned} \tag{2.4}$$

We estimate $\theta = (\beta, \zeta, \pi)$ by the ML estimator. The sample criterion function is

$$Q_n(\theta) = -n^{-1} \sum_{i=1}^{n} [Y_i \log L(g_i(\theta)) + (1 - Y_i) \log(1 - L(g_i(\theta)))] \tag{2.5}$$

and the ML estimator minimizes $Q_n(\theta)$ over $\theta \in \Theta$. (Here we use the negative of the standard log-likelihood function so that the estimator minimizes the sample criterion function as in the general set-up of the paper.)

When $\beta = 0$, $g_i(\theta)$ and $Q_n(\theta)$ do not depend on $\pi$, and $\pi$ is not identified. □

The true distribution of the observations $\{W_i : i \leq n\}$ is denoted $F_\gamma$ for some parameter $\gamma \in \Gamma$. We let $P_\gamma$ and $E_\gamma$ denote probability and expectation under $F_\gamma$. The parameter space $\Gamma$ for the true parameter, referred to as the "true parameter space," is compact and is of the form:

$$\Gamma = \{\gamma = (\theta, \phi) : \theta \in \Theta^*, \phi \in \Phi^*(\theta)\}, \tag{2.6}$$

where $\Theta^*$ is a compact subset of $R^{d_\theta}$ and $\Phi^*(\theta) \subset \Phi^*$ $\forall \theta \in \Theta^*$ for some compact metric space $\Phi^*$ with a metric that induces weak convergence of the bivariate distributions $(W_i, W_{i+m})$ for all $i, m \geq 1$.[4] In unconditional likelihood scenarios, no parameter $\phi$ appears. In conditional likelihood scenarios, with conditioning variables $\{X_i : i \geq 1\}$, $\phi$ indexes the distribution of $\{X_i : i \geq 1\}$. In nonlinear regression models estimated by least squares, $\theta$ indexes the regression functions and possibly a finite-dimensional feature of the distribution of the errors, such as its variance, and $\phi$ indexes the remaining characteristics of the distribution of the errors, which may be infinite dimensional.

By definition, the estimator $\widehat{\theta}_n$ (approximately) minimizes $Q_n(\theta)$ over an "optimization parameter space" $\Theta$:[5]

$$\widehat{\theta}_n \in \Theta \text{ and } Q_n(\widehat{\theta}_n) = \inf_{\theta \in \Theta} Q_n(\theta) + o(n^{-1}). \tag{2.7}$$

We assume that the interior of $\Theta$ includes the true parameter space $\Theta^*$ (see Assump-

[4]Thus, the metric satisfies: if $\gamma \to \gamma_0$, then $(W_i, W_{i+m})$ under $\gamma$ converges in distribution to $(W_i, W_{i+m})$ under $\gamma_0$. Note that $\Gamma$ 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 $d_{\Phi^*}$ is the metric on $\Phi^*$.

[5]The $o(n^{-1})$ term in (2.7), and in (4.1) and (4.2) below, is a fixed sequence of constants that does not depend on the true parameter $\gamma \in \Gamma$ and does not depend on $\pi$ in (4.1). The $o(n^{-1})$ term allows for some numerical inaccuracy in practice and circumvents the issue of the existence of parameter values that achieve the infima.

tion B1 below). This ensures that the asymptotic distribution of $\widehat{\theta}_n$ is not affected by boundary constraints for any sequence of true parameters in $\Theta^*$. The focus of this paper is not on boundary effects.

Without loss of generality (wlog), the optimization parameter space $\Theta$ can be written as

$$\begin{aligned} \Theta &= \{\theta = (\psi, \pi) : \psi \in \Psi(\pi), \pi \in \Pi\}, \text{ where} \\ \Pi &= \{\pi : (\psi, \pi) \in \Theta \text{ for some } \psi\} \text{ and} \\ \Psi(\pi) &= \{\psi : (\psi, \pi) \in \Theta\} \text{ for } \pi \in \Pi. \end{aligned} \tag{2.8}$$

We allow $\Psi(\pi)$ to depend on $\pi$ and, hence, $\Theta$ need not be a product space between $\psi$ and $\pi$. For example, this is needed in the STAR model and in the ARMA(1, 1) example in AC1.

**Example 1 (cont.).** The true parameter space for $\theta$ is

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

$b_1^* \geq 0$, $b_2^* \geq 0$, $b_1^*$ 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.

The ML estimator of $\theta$ minimizes $Q_n(\theta)$ over $\theta \in \Theta$. The optimization parameter space $\Theta$ is

$$\Theta = \mathcal{B} \times \mathcal{Z} \times \Pi, \text{ where } \mathcal{B} = [-b_1, b_2] \subset R, \tag{2.10}$$

$b_1 > b_1^*$, $b_2 > b_2^*$, $\mathcal{Z}$ ($\subset R^{d_\zeta}$) is compact, $\Pi$ ($\subset R^{d_\pi}$) is compact, $\mathcal{Z}^* \in int(\mathcal{Z})$, and $\mathcal{B}^* \in int(\mathcal{B})$.

## 2.2. Confidence Sets and Tests

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

CS's are obtained by inverting tests. A nominal $1 - \alpha$ CS for $r(\theta)$ is

$$CS_n = \{v : \mathcal{T}_n(v) \leq c_{n,1-\alpha}(v)\}, \tag{2.11}$$

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

$$P_\gamma(r(\theta) \in CS_n) = P_\gamma(\mathcal{T}_n(r(\theta)) \leq c_{n,1-\alpha}(r(\theta))), \tag{2.12}$$

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

We are interested in the finite-sample size of a 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 as follows:

$$AsySz = \liminf_{n\to\infty} \inf_{\gamma\in\Gamma} P_\gamma(r(\theta) \in CS_n) = \liminf_{n\to\infty} \inf_{\gamma\in\Gamma} P_\gamma(\mathcal{T}_n(r(\theta)) \leq c_{n,1-\alpha}(r(\theta))). \tag{2.13}$$

For a test, we are interested in the maximum null rejection probability, which is the finite-sample size of the test. A test's asymptotic size is an approximation to the latter. The asymptotic size of a test of $H_0 : r(\theta) = v$ is

$$AsySz = \limsup_{n\to\infty} \sup_{\gamma\in\Gamma: r(\theta)=v} P_\gamma(\mathcal{T}_n(v) > c_{n,1-\alpha}(v)). \tag{2.14}$$

## 2.3. Drifting Sequences of Distributions

The uniformity over $\gamma \in \Gamma$ for any given sample size $n$ in (2.13) and (2.14) is crucial for the asymptotic size to be a good approximation to the finite-sample size. The value of $\gamma$ at which the finite-sample size of a CS or test is attained may vary with the sample size. Thus, to determine the asymptotic size we need to derive the asymptotic distribution of the test statistic $\mathcal{T}_n(v_n)$ under sequences of true parameters $\gamma_n = (\theta_n, \phi_n)$ and $v_n = r(\theta_n)$ that may depend on $n$.

As shown in Andrews and Guggenberger (2010) and Andrews, Cheng, and Guggenberger (2009), the asymptotic size of CS's and tests are determined by certain drifting sequences of distributions. The following sequences $\{\gamma_n\}$ are key:

$$\begin{aligned}
\Gamma(\gamma_0) &= \{\{\gamma_n \in \Gamma : n \geq 1\} : \gamma_n \to \gamma_0 \in \Gamma\}, \\
\Gamma(\gamma_0, 0, b) &= \left\{\{\gamma_n\} \in \Gamma(\gamma_0) : \beta_0 = 0 \text{ and } n^{1/2}\beta_n \to b \in R^{d_\beta}_{[\pm\infty]}\right\}, \text{ and} \\
\Gamma(\gamma_0, \infty, \omega_0) &= \left\{\{\gamma_n\} \in \Gamma(\gamma_0) : n^{1/2}||\beta_n|| \to \infty \text{ and } \beta_n/||\beta_n|| \to \omega_0 \in R^{d_\beta}\right\},
\end{aligned} \tag{2.15}$$

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 $\beta = 0$: $n^{1/2}||\beta_n|| \to \infty$.

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

# 3. Assumptions

## 3.1. Smooth Sample Average Assumptions

This section provides primitive sufficient conditions for many of the high-level assumptions given in AC1 for the class of sample average criterion functions that are smooth in $\theta$. Note that the high-level assumptions in AC1 concern limit behavior under drifting sequences of true distributions. In contrast, the assumptions given here concern behavior under fixed true distributions and do not involve the sample size $n$.[6]

In Assumptions S1-S4 below, the true distribution of $\{W_i : i \geq 1\}$ is $F_{\gamma_0}$. The conditions in Assumptions S1-S4 are assumed to hold for all $\gamma_0 = (\beta_0, \zeta_0, \pi_0, \phi_0) \in \Gamma$. Let $C$ be a generic finite positive constant that does not necessarily take the same value when it appears in two different places. None of the constants that appear in Assumptions S1-S4 depend on $\gamma_0 \in \Gamma$.

### 3.1.1. Assumption S1

The first assumption is the following.

**Assumption S1.** Under any $\gamma_0 \in \Gamma$, $\{W_i : i \geq 1\}$ is a strictly stationary and strong mixing sequence with mixing coefficients $\alpha_m \leq Cm^{-A}$ for some $A > d_\theta q/(q - d_\theta)$ and some $q > d_\theta \geq 2$, or $\{W_i : i \geq 1\}$ is an i.i.d. sequence and the constant $q$ (that appears in Assumption S3 below) equals $2 + \delta$ for some $\delta > 0$.

[6] The sufficient conditions given here imply Assumptions A, B3, C1-C8, and D1-D3 of AC1.

In Assumption S1, the decay rate of the strong mixing coefficients is used to obtain the stochastic equicontinuity of certain empirical processes using results in Hansen (1996). The WLLN and CLT for strong mixing arrays also hold under this decay rate, see Andrews (1988) and de Jong (1997). In the i.i.d. case, the constant $q$ is smaller than in the strong mixing case, which yields weaker moment restrictions in Assumption S3 below.

**Example 1 (cont.).** In this example, Assumption S1 holds with $q = 2 + \delta$ for some $\delta > 0$ because $\{W_i : i \geq 1\}$ are i.i.d. for each $\gamma_0 \in \Gamma$. □

### 3.1.2. Assumption S2

The second assumption is as follows.

**Assumption S2.** (i) For some function $\rho(w, \theta) \in R$, $Q_n(\theta) = n^{-1} \sum_{i=1}^n \rho(W_i, \theta)$, where $\rho(w, \theta)$ is twice continuously differentiable in $\theta$ on an open set containing $\Theta^*$ $\forall w \in \mathcal{W}$.
(ii) $\rho(w, \theta)$ does not depend on $\pi$ when $\beta = 0$ $\forall w \in \mathcal{W}$.
(iii) $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$, $E_{\gamma_0} \rho(W_i, \psi, \pi)$ is uniquely minimized by $\psi_0$ $\forall \pi \in \Pi$.
(iv) $\forall \gamma_0 \in \Gamma$ with $\beta_0 \neq 0$, $E_{\gamma_0} \rho(W_i, \theta)$ is uniquely minimized by $\theta_0$.
(v) $\Psi(\pi)$ is compact $\forall \pi \in \Pi$, and $\Pi$ and $\Theta$ are compact.
(vi) $\forall \varepsilon > 0$, $\exists \delta > 0$ 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.

For i.i.d. observations with density $f(w, \theta)$, the ML estimator is obtained by taking $\rho(W_i, \theta) = \log f(W_i, \theta)$. For a stationary $p$-th order Markov process $\{W_i^* : -p + 1 \leq i \leq n\}$, we let $W_i = (W_i^*, ..., W_{i-p}^*)$. If the conditional density of $W_i^*$ given $(W_{i-1}^*, ..., W_{i-p}^*)$ is $f(w^* | W_{i-1}^*, ..., W_{i-p}^*; \theta)$, then the ML estimator is obtained by taking $\rho(W_i, \theta) = \log f(W_i^* | W_{i-1}^*, ..., W_{i-p}^*; \theta)$.

**Example 1 (cont.).** Assumption S2(i) holds in this example with

$$\rho(W_i, \theta) = Y_i \log L(g_i(\theta)) + (1 - Y_i) \log(1 - L(g_i(\theta))) \tag{3.1}$$

by (2.5) and the smoothness conditions on $h(X_i, \pi)$ and $L(u)$. Assumption S2(ii) holds because $g_i(\theta)$ does not depend on $h(X_i, \pi)$ when $\beta = 0$. For brevity, Assumptions S2(iii) and S2(iv) are verified in Supplemental Appendix D. The argument for Assumption S2(iv) is a standard argument for ML estimators in well-identified scenarios. Assumption S2(v) holds because $\Psi(\pi) = \mathcal{B} \times \mathcal{Z}$, which does not depend on $\pi$, $\Theta = \mathcal{B} \times \mathcal{Z} \times \Pi$, and

$\mathcal{B}$, $\mathcal{Z}$, and $\Pi$ are all compact. Assumption S2(vi) holds because $\Psi(\pi)$ does not depend on $\pi$. $\square$

A class of examples of $\rho(w,\theta)$ functions that satisfy Assumption S2(ii) are functions of the form

$$\rho(w,\theta) = \rho^*(w, a(x,\beta)h(x,\pi), \zeta), \text{ where } a(x,0) = 0, \ \forall w \in \mathcal{W}, \tag{3.2}$$

$x$ is a sub-vector of $w$, and $a(x,\beta)$ and $h(x,\pi)$ are known functions. In (3.2), $\rho(w,\theta)$ does not depend on $\pi$ when $\beta = 0$ because $a(x,\beta) = 0$. Examples of $a(x,\beta)$ include (i) $a(x,\beta) = \beta$, (ii) $a(x,\beta) = \exp(\beta) - 1$, and (iii) $a(x,\beta) = x'\beta$. Example (i) covers the nonlinear regression example, where $\beta$ is the coefficient of the nonlinear regressor. Example (ii) demonstrates that $a(x,\beta)$ can be nonlinear in $\beta$ provided $a(x,\beta) = 0$ at $\beta = 0$. Example (iii) covers the weak IV example and the case in which $\beta$ enters the model through a single index. The form in (3.2) does not require a regression model and it allows for complicated structural models by allowing different functional forms for $a(x,\beta)$, $h(x,\pi)$, and $\rho(w,\theta)$.

Returning now to the general $\rho(w,\theta)$ case, Assumption S2(vi) holds immediately in cases where $\Psi(\pi)$ does not depend on $\pi$. When $\Psi(\pi)$ depends on $\pi$, the boundary of $\Psi(\pi)$ is often a continuous linear function of $\pi$, as in the STAR model and the ARMA(1,1) model considered in AC1. In such cases, it is simple to verify Assumption S2(vi).

### 3.1.3. Assumption S3

Let $\rho_\theta(w,\theta)$ and $\rho_{\theta\theta}(w,\theta)$ denote the first-order and second-order partial derivatives of $\rho(w,\theta)$ wrt $\theta$, respectively. Let $\rho_\psi(w,\theta)$ and $\rho_{\psi\psi}(w,\theta)$ denote the first-order and second-order partial derivatives of $\rho(w,\theta)$ wrt $\psi$.

We define a matrix $B(\beta)$ that is used to normalize the second-derivative matrix $\rho_{\theta\theta}(w,\theta)$ so that its sample average has a nonsingular probability limit. Let

$$B(\beta) = \begin{bmatrix} I_{d_\psi} & 0_{d_\psi \times d_\pi} \\ 0_{d_\pi \times d_\psi} & \iota(\beta) I_{d_\pi} \end{bmatrix} \in R^{d_\theta \times d_\theta}, \text{ where } \iota(\beta) = \begin{cases} \beta & \text{if } \beta \text{ is a scalar} \\ ||\beta|| & \text{if } \beta \text{ is a vector} \end{cases}. \tag{3.3}$$

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

For $\beta \neq 0$, let

$$B^{-1}(\beta)\rho_\theta(w,\theta) = \rho^\dagger_\theta(w,\theta) \text{ and}$$
$$B^{-1}(\beta)\rho_{\theta\theta}(w,\theta)B^{-1}(\beta) = \rho^\dagger_{\theta\theta}(w,\theta) + \iota^{-1}(\beta)\varepsilon(w,\theta), \tag{3.4}$$

where $\rho^\dagger_{\theta\theta}(w,\theta)$ is symmetric and $\rho^\dagger_\theta(w,\theta)$, $\rho^\dagger_{\theta\theta}(w,\theta)$, and $\varepsilon(w,\theta)$ satisfy Assumption S3 below. The re-scaling matrix $B^{-1}(\beta)$ in (3.4) is used to deal with the singularity issue that arises when $\beta = 0$. In particular, the covariance matrix of $\rho_\theta(W_i,\theta)$ is singular when $\beta = 0$ and close to singular when $\beta$ is close to 0. In contrast, the rescaled quantity $\rho^\dagger_\theta(W_i,\theta)$ has a covariance matrix that is not close to being singular even when $\beta$ is close to 0. Similarly, $E_{\gamma_0}\rho_{\theta\theta}(W_i,\theta)$ is singular when $\beta = 0$ and close to singular when $\beta$ is close to 0. Re-scaling of $\rho_{\theta\theta}(W_i,\theta)$ yields a quantity $\rho^\dagger_{\theta\theta}(W_i,\theta)$ whose expectation is not close to singular even when $\beta$ is close to 0 plus another term $\varepsilon(W_i,\theta)$ that is asymptotically negligible.

Below we illustrate the form of $\rho^\dagger_\theta(w,\theta)$, $\rho^\dagger_{\theta\theta}(w,\theta)$, and $\varepsilon(w,\theta)$ in Example 1 and for $\rho(w,\theta)$ functions as in (3.2), see Section 9.1 of the Appendix.

Next, define

$$V^\dagger(\theta_1,\theta_2;\gamma_0) = \sum_{m=-\infty}^{\infty} Cov_{\gamma_0}(\rho^\dagger_\theta(W_i,\theta_1), \rho^\dagger_\theta(W_{i+m},\theta_2)), \tag{3.5}$$

which does not depend on $i$ because the observations are stationary under Assumption S1. Under Assumptions S1 and S3(iii) below, $V^\dagger(\theta_1,\theta_2;\gamma_0)$ exists by a standard strong mixing inequality.

The form of Assumption S3 differs depending on whether $\beta$ is a scalar or vector. We state Assumption S3 for the scalar $\beta$ case first because it is simpler.

**Assumption S3 (scalar $\beta$).** (i) $E_{\gamma_0}\varepsilon(W_i,\theta_0) = 0$ and $|\beta_0|^{-1}||E_{\gamma_0}\varepsilon(W_i,\psi_0,\pi)|| \leq C||\pi - \pi_0|| \ \forall\gamma_0 \in \Gamma$ with $0 < |\beta_0| < \delta$ for some $\delta > 0$.
(ii) For all $\delta > 0$ and some functions $M_1(w) : \mathcal{W} \to R_+$ and $M_2(w) : \mathcal{W} \to R_+$, $||\rho_{\psi\psi}(w,\theta_1) - \rho_{\psi\psi}(w,\theta_2)|| + ||\rho^\dagger_{\theta\theta}(w,\theta_1) - \rho^\dagger_{\theta\theta}(w,\theta_2)|| \leq M_1(w)\delta$ and $||\rho^\dagger_\theta(w,\theta_1) - \rho^\dagger_\theta(w,\theta_2)|| + ||\varepsilon(w,\theta_1) - \varepsilon(w,\theta_2)|| \leq M_2(w)\delta$, $\forall\theta_1,\theta_2 \in \Theta$ with $||\theta_1 - \theta_2|| \leq \delta$, $\forall w \in \mathcal{W}$.
(iii) $E_{\gamma_0} \sup_{\theta\in\Theta}\{|\rho(W_i,\theta)|^{1+\delta} + ||\rho_{\psi\psi}(W_i,\theta)||^{1+\delta} + ||\rho^\dagger_{\theta\theta}(W_i,\theta)||^{1+\delta} + M_1(W_i) + ||\rho^\dagger_\theta(W_i,\theta)||^q + ||\varepsilon(W_i,\theta)||^q + M_2(W_i)^q\} \leq C$ for some $\delta > 0$ $\forall\gamma_0 \in \Gamma$, where $q$ is as in Assumption S1.
(iv) $\lambda_{\min}(E_{\gamma_0}\rho_{\psi\psi}(W_i,\psi_0,\pi)) > 0$ $\forall\pi \in \Pi$ when $\beta_0 = 0$ and $E_{\gamma_0}\rho^\dagger_{\theta\theta}(W_i,\theta_0)$ is positive definite $\forall\gamma_0 \in \Gamma$.

(v) $V^\dagger(\theta_0, \theta_0; \gamma_0)$ is positive definite $\forall \gamma_0 \in \Gamma$.

In Assumption S3(iii), the last three terms have bounded $q$th moments in order to establish the stochastic equicontinuity of empirical processes based on $\rho^\dagger_\theta(W_i, \theta)$ and $\varepsilon(W_i, \theta)$ using Lemma 11.4 in Supplemental Appendix A.

In Assumptions S1-S3, Assumptions S2(ii), S2(iii), S3(i), S3(iii), S3(iv) and S3(v) are related to the weak identification problem. Assumption S2(ii) implies that the sample criterion function is flat in $\pi$ when $\beta = 0$, as in Assumption A of AC1. Assumption S2(iii) differs from a standard condition in the sense that the population criterion function is not uniquely minimized by the true value when $\beta_0 = 0$. The Lipschitz condition in Assumption S3(i) typically holds because the partial derivative of $E_{\gamma_0}\varepsilon(W_i, \psi_0, \pi)$ wrt $\pi$ is approximately proportional to $||\beta_0||$ when $||\beta_0||$ is close to 0. Because parts of $B^{-1}(\beta)$ diverge as $\beta$ converges to 0, the moment conditions for $\rho^\dagger_\theta(W_i, \theta)$ and $\rho^\dagger_{\theta\theta}(W_i, \theta)$ in Assumption S3(iii) are stronger than standard moment conditions on the first-order and second-order derivatives. These conditions hold in typical examples, see below, because the partial derivative of $\rho(w, \theta)$ wrt $\pi$ is small when $\beta$ is close to 0 under Assumption S2(ii). Hence, the rhs moments are uniformly bounded even after the scaling by $B^{-1}(\beta)$. In Assumptions S3(iv) and S3(v), $E_{\gamma_0}\rho^\dagger_{\theta\theta}(W_i, \theta_0)$ and $V^\dagger(\theta_0, \theta_0; \gamma_0)$ typically are positive definite due to the re-scaling in (3.4).

Under Assumptions S1-S3, the criterion function $Q_n(\theta)$ has probability limit $Q(\theta; \gamma) = E_\gamma \rho(W_i, \theta)$ under any sequence of parameters $\gamma_n \to \gamma$.

**Example 1 (cont.).** In this example, $\rho^\dagger_\theta(W_i, \theta)$, $\rho^\dagger_{\theta\theta}(W_i, \theta)$, and $\varepsilon(w, \theta)$ are defined as follows. For notational simplicity, let $L_i(\theta)$, $L'_i(\theta)$, and $L''(\theta)$ abbreviate $L(g_i(\theta))$, $L'(g_i(\theta))$, and $L''(g_i(\theta))$, respectively. Let

$$d_{\psi,i}(\pi) = (h(X_i, \pi), Z_i')', \; d_i(\pi) = (h(X_i, \pi), Z_i', h_\pi(X_i, \pi)')', \text{ and}$$
$$D_i(\theta) = \begin{bmatrix} 0 & 0_{1\times d_\zeta} & h_\pi(X_i, \pi)' \\ 0_{d_\zeta \times 1} & 0_{d_\zeta \times d_\zeta} & 0_{d_\zeta \times d_\pi} \\ h_\pi(X_i, \pi) & 0_{d_\pi \times d_\zeta} & h_{\pi\pi}(X_i, \pi)\beta \end{bmatrix}. \tag{3.6}$$

The first-and second-order partial derivatives of $\rho(W_i, \theta)$ wrt to $\psi$ and $\theta$ are

$$\begin{aligned}
\rho_\psi(W_i,\theta) &= w_{1,i}(\theta)(Y_i - L_i(\theta))d_{\psi,i}(\pi), \\
\rho_\theta(W_i,\theta) &= w_{1,i}(\theta)(Y_i - L_i(\theta))B(\beta)d_i(\pi), \\
\rho_{\psi\psi}(W_i,\theta) &= [w_{1,i}^2(\theta)(Y_i - L_i(\theta))^2 + w_{2,i}(\theta)(Y_i - L_i(\theta))]d_{\psi,i}(\pi)d_{\psi,i}(\pi)', \\
\rho_{\theta\theta}(W_i,\theta) &= [w_{1,i}^2(\theta)(Y_i - L_i(\theta))^2 + w_{2,i}(\theta)(Y_i - L_i(\theta))]B(\beta)d_i(\pi)d_i(\pi)'B(\beta) \\
&\quad + w_{1,i}(\theta)(Y_i - L_i(\theta))D_i(\theta), \text{ where} \\
w_{1,i}(\theta) &= \frac{-L_i'(\theta)}{L_i(\theta)(1-L_i(\theta))} \text{ and } w_{2,i}(\theta) = \frac{-L_i''(\theta)}{L_i(\theta)(1-L_i(\theta))}.
\end{aligned} \tag{3.7}$$

See Section 14.11 in Supplemental Appendix D for the calculation of these derivatives.

The rescaled partial derivatives in (3.4) take the form

$$\begin{aligned}
\rho_\theta^\dagger(W_i,\theta) &= w_{1,i}(\theta)(Y_i - L_i(\theta))d_i(\pi), \\
\rho_{\theta\theta}^\dagger(W_i,\theta) &= [w_{1,i}^2(\theta)(Y_i - L_i(\theta))^2 + w_{2,i}(\theta)(Y_i - L_i(\theta))]d_i(\pi)d_i(\pi)', \text{ and} \\
\varepsilon(w,\theta) &= w_{1,i}(\theta)(Y_i - L_i(\theta)) \begin{bmatrix} 0 & 0_{1\times d_\zeta} & h_\pi(X_i,\pi)' \\ 0_{d_\zeta\times 1} & 0_{d_\zeta\times d_\zeta} & 0_{d_\zeta\times d_\pi} \\ h_\pi(X_i,\pi) & 0_{d_\pi\times d_\zeta} & h_{\pi\pi}(X_i,\pi) \end{bmatrix}.
\end{aligned} \tag{3.8}$$

Assumption S3 is verified in Supplemental Appendix D for this example. □

When $\beta$ is a vector, i.e., $d_\beta > 1$, we reparameterize $\beta$ 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, $\theta$ is reparameterized as $\theta^+ = (||\beta||, \omega, \zeta, \pi)$. Let $\Theta^+ = \{\theta^+ : \theta^+ = (||\beta||, \beta/||\beta||, \zeta, \pi),\ \theta \in \Theta\}$.

This new parameterization is needed when $\beta$ is a vector because $\rho_\theta^\dagger(w,\theta)$, $\rho_{\theta\theta}^\dagger(w,\theta)$, and $\varepsilon(w,\theta)$ typically involve $\beta/||\beta||$ due to the re-scaling in (3.4) and $\beta/||\beta||$ is not continuous in $\beta$ for $\theta \in \Theta$. In consequence, the Lipschitz conditions in Assumptions S3(ii) and S3(iii) (scalar $\beta$) can not be verified when $\beta$ is a vector. The new parameterization treats $||\beta||$ and $\omega = \beta/||\beta||$ as separate variables. In Assumption S3 (vector $\beta$) below, some Lipschitz conditions are specified in terms of $\theta^+ = (||\beta||, \omega, \zeta, \pi)$.

In Assumption S3 (vector $\beta$), both the original parameterization with $\theta$ and the alternative parameterization with $\theta^+$ are employed for convenience. Note that only conditions related to $\rho_\theta^\dagger(w,\theta)$, $\rho_{\theta\theta}^\dagger(w,\theta)$, and $\varepsilon(w,\theta)$ require the alternative parameterization with $\theta^+$.

**Assumption S3 (vector $\boldsymbol{\beta}$).** (i) $E_{\gamma_0}\varepsilon(W_i,\theta_0) = 0$ and $||\beta_0||^{-1}||E_{\gamma_0}\varepsilon(W_i,\theta^+)|| \leq C(||\pi-\pi_0|| + ||\omega-\omega_0||)$ $\forall\theta^+ = (||\beta_0||,\omega,\zeta_0,\pi)$ and $\forall\gamma_0 \in \Gamma$ with $0 < ||\beta_0|| < \delta$ for some $\delta > 0$.
(ii) For all $\delta > 0$ and some functions $M_1(w) : \mathcal{W} \to R_+$ and $M_2(w) : \mathcal{W} \to R_+$, $||\rho_{\psi\psi}(w,\theta_1) - \rho_{\psi\psi}(w,\theta_2)|| + ||\rho^\dagger_{\theta\theta}(w,\theta_1^+) - \rho^\dagger_{\theta\theta}(w,\theta_2^+)|| \leq M_1(w)\delta$ and $||\rho_\psi(w,\theta_1) - \rho_\psi(w,\theta_2)|| + ||\rho^\dagger_\theta(w,\theta_1^+) - \rho^\dagger_\theta(w,\theta_2^+)|| + ||\varepsilon(w,\theta_1^+) - \varepsilon(w,\theta_2^+)|| \leq M_2(w)\delta$, $\forall\theta_1,\theta_2 \in \Theta$ with $||\theta_1-\theta_2|| \leq \delta$, $\forall\theta_1^+,\theta_2^+ \in \Theta^+$ with $||\theta_1^+ - \theta_2^+|| \leq \delta$, $\forall w \in \mathcal{W}$.
(iii) Assumptions S3(iii)-(iv) (scalar $\beta$) hold with the definitions of $M_1(w)$ and $M_2(w)$ replaced by those given above.

Assumption S3(i) (vector $\beta$) typically holds because the partial derivatives of $E_{\gamma_0}\varepsilon(W_i, \theta^+)$ wrt $\pi$ and $\omega$ are approximately proportional to $||\beta_0||$.

### 3.1.4. Assumption S4

Next, we state an assumption that controls how the mean $E_{\gamma_0}\rho_{\psi,i}(\theta)$ changes as the true $\beta_0$ changes, where $\gamma_0 = (\beta_0,\zeta_0,\pi_0,\phi_0)$. Define the $d_\psi \times d_\beta$-matrix of partial derivatives of the average population moment function wrt the true $\beta$ value, $\beta_0$, to be

$$K(\theta;\gamma_0) = \frac{\partial}{\partial\beta_0'}E_{\gamma_0}\rho_\psi(W_i,\theta). \tag{3.9}$$

The domain of the function $K(\theta;\gamma_0)$ is $\Theta_\delta \times \Gamma_0$, where $\Theta_\delta = \{\theta \in \Theta : ||\beta|| < \delta\}$ and $\Gamma_0 = \{\gamma_a = (a\beta,\zeta,\pi,\phi) \in \Gamma : \gamma = (\beta,\zeta,\pi,\phi) \in \Gamma \text{ with } ||\beta|| < \delta \text{ and } a \in [0,1]\}$ for some $\delta > 0$.[7]

**Assumption S4.** (i) $K(\theta;\gamma_0)$ exists $\forall(\theta,\gamma_0) \in \Theta_\delta \times \Gamma_0$.
(ii) $K(\theta;\gamma^*)$ is continuous in $(\theta,\gamma^*)$ at $(\theta,\gamma^*) = ((\psi_0,\pi),\gamma_0)$ uniformly over $\pi \in \Pi$ $\forall\gamma_0 \in \Gamma$ with $\beta_0 = 0$, where $\psi_0$ is a sub-vector of $\gamma_0$.

Assumption S4 is not restrictive in most applications.[8]

For simplicity, $K(\psi_0,\pi;\gamma_0)$ is abbreviated as $K(\pi;\gamma_0)$.

**Example 1 (cont.).** It is shown in Supplemental Appendix D that Assumption S4

[7] The constant $\delta > 0$ is as in Assumption B2(iii) stated below. The set $\Gamma_0$ is not empty by Assumption B2(ii).

[8] Assumptions S1 and S4 imply Assumption C5 of AC1. A set of primitive sufficient conditions for Assumption C5 of AC1 is given in Appendix A of AC1-SM. These conditions also are sufficient for Assumption S4.

holds with

$$K(\pi;\gamma_0) = K(\psi_0,\pi;\gamma_0) = E_{\gamma_0}\frac{L_i'^2(\theta_0)}{L_i(\theta_0)(1-L_i(\theta_0))}h(X_i,\pi_0)d_{\psi,i}(\pi) \tag{3.10}$$

for $\gamma_0 = (\theta_0, \pi_0)$. □

## 3.2. Parameter Space Assumptions

Next, we specify conditions on the parameter spaces $\Theta$ and $\Gamma$.

Define $\Theta^*_\delta = \{\theta \in \Theta^* : ||\beta|| < \delta\}$, where $\Theta^*$ is the true parameter space for $\theta$, see (2.6). The optimization parameter space $\Theta$ satisfies:

**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 non-empty open set $\mathcal{Z}^0 \subset R^{d_\zeta}$ and $\Pi$ as in (2.8).
(iii) $\Pi$ is compact.

Because the optimization parameter space is user selected, Assumptions B1(ii)-(iii) can be made to hold by the choice of $\Theta$.

The true parameter space $\Gamma$ satisfies:

**Assumption B2.** (i) $\Gamma$ is compact and (2.6) 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 $\Gamma$ is not empty and that there are elements $\gamma$ of $\Gamma$ whose $\beta$ values are non-zero 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 $\Gamma$ is compatible with the existence of partial derivatives of certain expectations wrt the true parameter $\beta$ around $\beta = 0$, which arise in (3.9) and Assumption S4.

**Example 1 (cont.).** Let $\gamma = (\theta,\phi)$, where $\phi$ is the distribution of $(X_i, Z_i)$, and $\phi \in \Phi^*$, where $\Phi^*$ is a compact metric space with some metric that induces weak convergence. The parameter space for the true value of $\gamma$ is

$$\Gamma = \{\gamma = (\theta,\phi) : \theta \in \Theta^*, \phi \in \Phi^*(\theta)\}, \tag{3.11}$$

where $\Phi^*(\theta) \subset \Phi^*$ $\forall\theta \in \Theta^*$.

The parameter space $\Phi^*(\theta)$, which must be specified precisely to obtain the uniform asymptotic results, is defined as follows. For notational simplicity, let $\overline{h}_i = \sup_{\pi\in\Pi}|h(X_i,\pi)|$, $\overline{h}_{\pi,i} = \sup_{\pi\in\Pi}||h_\pi(X_i,\pi)||$, $\overline{h}_{\pi\pi,i} = \sup_{\pi\in\Pi}||h_{\pi\pi}(X_i,\pi)||$, $\overline{w}_{1,i} = \sup_{\theta\in\Theta}|w_{1,i}(\theta)|$, and $\overline{w}_{2,i} = \sup_{\theta\in\Theta}|w_{2,i}(\theta)|$. Let $q = 2+\delta$ for some $\delta > 0$.

For any $\theta_0 \in \Theta^*$, the true parameter space for $\phi$ is

$$
\begin{aligned}
&\Phi^*(\theta_0) = \{\phi_0 \in \Phi^* :\ E_{\gamma_0}(\overline{h}_i^{4q} + \overline{h}_{\pi,i}^{4q} + \overline{h}_{\pi\pi,i}^{4q} + ||Z_i||^{4q} + \overline{w}_{1,i}^{4q} + \overline{w}_{2,i}^{2+\delta}) \le C \\
&||w_{1,i}(\theta_1) - w_{1,i}(\theta_2)|| \le M_1(W_i)||\pi_1 - \pi_2||,\ ||w_{2,i}(\theta_1) - w_{2,i}(\theta_2)|| \\
&\le M_2(W_i)||\pi_1 - \pi_2||,\ ||h_{\pi\pi}(X_i,\pi_1) - h_{\pi\pi}(X_i,\pi_2)|| \le M_h(W_i)||\pi_1 - \pi_2||, \\
&\forall \pi_1,\pi_2 \in \Pi \text{ for some functions } M_1(W_i),\ M_2(W_i),\ M_h(W_i), \\
&E_{\gamma_0}(M_1(W_i)^{4q/3} + M_2(W_i)^{4/3} + M_h(W_i)^{4q/3}) \le C, \\
&E_{\gamma_0}\sup_{\theta\in\Theta}\left(|\log L_i(\theta)|^{1+\delta} + |\log(1-L_i(\theta))|^{1+\delta}\right) \le C, \\
&P_{\gamma_0}(a'(h(X_i,\pi_1),h(X_i,\pi_2),Z_i) = 0) < 1,\ \forall \pi_1,\pi_2 \in \Pi \text{ with } \pi_1 \ne \pi_2,\ \forall a \in R^{d_\zeta+2} \\
&\text{with } a \ne 0,\ E_{\gamma_0} d_i(\pi)d_i(\pi)' \text{ is positive definite } \forall \pi \in \Pi\}
\end{aligned} \tag{3.12}
$$

for some $C < \infty$, where $d_i(\pi) = (h(X_i,\pi), Z_i', h_\pi(X_i,\pi)')'$.[9] □

## 3.3. Key Quantities

Now, we define some of the key quantities that arise in the asymptotic distribution of the estimator $\widehat{\theta}_n$ and the test statistics considered. Let $S_\psi = [I_{d_\psi} : 0_{d_\psi \times d_\pi}]$ denote the $d_\psi \times d_\theta$ selector matrix that selects $\psi$ out of $\theta$. Define

$$
\begin{aligned}
\Omega(\pi_1,\pi_2;\gamma_0) &= S_\psi V^\dagger((\psi_0,\pi_1),(\psi_0,\pi_2);\gamma_0)S_\psi', \\
H(\pi;\gamma_0) &= E_{\gamma_0}\rho_{\psi\psi}(W_i,\psi_0,\pi), \\
J(\gamma_0) &= E_{\gamma_0}\rho_{\theta\theta}^\dagger(W_i,\theta_0),\ \text{and} \\
V(\gamma_0) &= V^\dagger(\theta_0,\theta_0;\gamma_0).
\end{aligned} \tag{3.13}
$$

**Example 1 (cont.).** The key quantities that determine the asymptotic behavior of the ML estimator in the binary choice model are as follows. The probability limit of the

[9] In (3.12), the expectation $E_{\gamma_0}(\cdot)$ only depends on $\phi_0$. Because $\theta_0$ shows up in some other expectations, we use $E_{\gamma_0}(\cdot)$ throughout the example for notational consistency.

criterion function $Q_n(\theta)$ when the true value is $\gamma_0 \in \Gamma$ is

$$\begin{aligned} Q(\theta;\gamma_0) &= E_{\gamma_0}\rho(W_i,\theta) = E_{\gamma_0}E_{\gamma_0}(\rho(W_i,\theta)|X_i,Z_i) \\ &= -E_{\gamma_0}[L_i(\theta_0)\log L_i(\theta) + (1-L_i(\theta_0))\log(1-L_i(\theta))]. \end{aligned} \tag{3.14}$$

By calculations given in Section 14.1 of Supplemental Appendix D, we have

$$\begin{aligned} \Omega(\pi_1,\pi_2;\gamma_0) &= E_{\gamma_0}\frac{L_i'^2(\theta_0)}{L_i(\theta_0)(1-L_i(\theta_0))}d_{\psi,i}(\pi_1)d_{\psi,i}(\pi_2)', \\ H(\pi;\gamma_0) &= E_{\gamma_0}\frac{L_i'^2(\theta_0)}{L_i(\theta_0)(1-L_i(\theta_0))}d_{\psi,i}(\pi)d_{\psi,i}(\pi)', \text{ and} \\ J(\gamma_0) &= V(\gamma_0) = E_{\gamma_0}\frac{L_i'^2(\theta_0)}{L_i(\theta_0)(1-L_i(\theta_0))}d_i(\pi_0)d_i(\pi_0)'. \end{aligned} \tag{3.15}$$

□

## 3.4. Quadratic Approximations

Here we specify certain quadratic approximations to $Q_n(\theta)$ and related results that hold under Assumptions S1-S4, B1, and B2. These results help to explain the form of the asymptotic distributions that arise in the results stated below.

(i) Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ (defined in (2.15) above), the sample criterion function $Q_n(\theta)$ $(= Q_n(\psi,\pi))$ has a quadratic expansion in $\psi$ around the point $\psi_{0,n} = (0,\zeta_n)$ for given $\pi$ for the form:

$$\begin{aligned} Q_n(\psi,\pi) = {} & Q_n(\psi_{0,n},\pi) + D_\psi Q_n(\psi_{0,n},\pi)'(\psi-\psi_{0,n}) + \\ & \frac{1}{2}(\psi-\psi_{0,n})'D_{\psi\psi}Q_n(\psi_{0,n},\pi)(\psi-\psi_{0,n}) + R_n(\psi,\pi), \end{aligned} \tag{3.16}$$

where $D_\psi Q_n(\psi_{0,n},\pi)$ and $D_{\psi\psi}Q_n(\psi_{0,n},\pi)$ denote the vector and matrix of first and second partial derivatives of $Q_n(\psi,\pi)$ with respect to $\psi$, respectively, evaluated at $\psi = \psi_{0,n}$, and $R_n(\psi,\pi)$ is a remainder term that is small uniformly in $\pi \in \Pi$ for $\psi$ close to $\psi_{0,n}$.[10]

(ii) Under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, the sample criterion function $Q_n(\theta)$ has a quadratic

[10] The precise conditions that the remainder $R_n(\psi,\pi)$ satisfies are specified in Assumption C1 of AC1. The quadratic approximation result (i) and results (ii)-(iv) that follow are established in the proof of Theorem 4.1 given in Supplemental Appendix A.

expansion in $\theta$ around the true value $\theta_n$ of the form:

$$Q_n(\theta) = Q_n(\theta_n) + DQ_n(\theta_n)'(\theta - \theta_n) + \frac{1}{2}(\theta - \theta_n)D^2Q_n(\theta_n)(\theta - \theta_n) + R_n^*(\theta), \quad (3.17)$$

where $DQ_n(\theta_n)$ and $D^2Q_n(\theta_n)$ denote the vector and matrix of first and second partial derivatives of $Q_n(\theta)$ with respect to $\theta$, respectively, evaluated at $\theta = \theta_n$, and $R_n^*(\theta)$ is a remainder term that is small for $\theta$ close to $\theta_n$.[11]

(iii) Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, the recentered and rescaled first derivative of $Q_n(\theta)$ wrt $\psi$ satisfies an empirical process CLT:

$$\begin{aligned} G_n(\cdot) &\Rightarrow G(\cdot; \gamma_0), \text{ where} \\ G_n(\pi) &= n^{-1/2} \sum_{i=1}^{n} \left(\rho_{\psi,i}(\psi_{0,n}, \pi) - E_{\gamma_n}\rho_{\psi,i}(\psi_{0,n}, \pi)\right) \end{aligned} \quad (3.18)$$

and $G(\cdot; \gamma_0)$ is 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 $\pi_1, \pi_2 \in \Pi$.

(iv) Under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, the rescaled first and second derivatives of $Q_n(\theta)$ satisfy

$$n^{1/2}B^{-1}(\beta_n)DQ_n(\theta_n) \to_d G^*(\gamma_0) \sim N(0_{d_\theta}, V(\gamma_0)) \quad (3.19)$$

and

$$J_n = B^{-1}(\beta_n)D^2Q_n(\theta_n)B^{-1}(\beta_n) \to_p J(\gamma_0) \in R^{d_\theta \times d_\theta} \ \forall \gamma_0 \in \Gamma. \quad (3.20)$$

## 3.5. Assumptions C6 and C7

In this section, we state assumptions that concern the minimum of the limit of the normalized criterion function after $\psi$ has been concentrated out.[12]

Define a "weighted non-central chi-square" process $\{\xi(\pi; \gamma_0, b) : \pi \in \Pi\}$ and a non-stochastic function $\{\eta(\pi; \gamma_0, \omega_0) : \pi \in \Pi\}$ by

$$\begin{aligned} \xi(\pi; \gamma_0, b) &= -\frac{1}{2}\left(G(\pi; \gamma_0) + K(\pi; \gamma_0)\, b\right)' H^{-1}(\pi; \gamma_0)\left(G(\pi; \gamma_0) + K(\pi; \gamma_0)b\right) \text{ and} \\ \eta(\pi; \gamma_0, \omega_0) &= -\frac{1}{2}\omega_0' K(\pi; \gamma_0)' H^{-1}(\pi; \gamma_0) K(\pi; \gamma_0)\omega_0. \end{aligned} \quad (3.21)$$

[11] The precise conditions that the remainder $R_n^*(\theta)$ satisfies are specified in Assumption D1 of AC1.

[12] Assumptions C6 and C7 are the same as in AC1, which is why the numbering starts at C6, rather than C1.

The process $\xi(\pi;\gamma_0,b)$ is the limit under $\{\gamma_n\} \in \Gamma(\gamma_0,0,b)$ for $||b|| < \infty$, defined in (2.15), and the function $\eta(\pi;\gamma_0,\omega_0)$ is the limit under $\{\gamma_n\} \in \Gamma(\gamma_0,0,b)$ for $||b|| = \infty$. Under Assumptions S1-S4, $\{\xi(\pi;\gamma_0,b) : \pi \in \Pi\}$ has bounded continuous sample paths a.s.

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

**Assumption C6.** Each sample path of the stochastic process $\{\xi(\pi;\gamma_0,b) : \pi \in \Pi\}$ in some set $A(\gamma_0,b)$ with $P_{\gamma_0}(A(\gamma_0,b)) = 1$ is minimized over $\Pi$ 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 C6, $\pi^*(\gamma_0,b)$ is random.

Next, we give a primitive sufficient condition for Assumption C6 for the case where $\beta$ is a scalar parameter. Let $\rho_\psi(w,\theta) = (\rho_\beta(w,\theta)', \rho_\zeta(w,\theta)')'$. When $\beta = 0$, $\rho_\zeta(w,\theta)$ does not depend on $\pi$ by Assumption S2(ii) and is denoted by $\rho_\zeta(w,\psi)$. When $d_\beta = 1$ and $\beta_0 = 0$, define

$$\begin{aligned}\rho^*_\psi(W_i,\psi_0,\pi_1,\pi_2) &= (\rho_\beta(W_i,\psi_0,\pi_1), \rho_\beta(W_i,\psi_0,\pi_2), \rho_\zeta(W_i,\psi_0)')' \text{ and} \\ \Omega_G(\pi_1,\pi_2;\gamma_0) &= \sum_{m=-\infty}^{\infty} Cov_{\gamma_0}(\rho^*_\psi(W_i,\psi_0,\pi_1,\pi_2), \rho^*_\psi(W_{i+m},\psi_0,\pi_1,\pi_2)). \quad (3.22)\end{aligned}$$

**Assumption C6$^\dagger$.** (i) $d_\beta = 1$ (i.e., $\beta$ is a scalar).
(ii) $\Omega_G(\pi_1,\pi_2;\gamma_0)$ is positive definite $\forall\pi_1,\pi_2 \in \Pi$ with $\pi_1 \neq \pi_2$, $\forall\gamma_0 \in \Gamma$ with $\beta_0 = 0$.

**Lemma 3.1.** *Assumptions* S1-S3 *and* C6$^\dagger$ *imply Assumption* C6.

**Example 1 (cont.).** For this example, Assumption C6$^\dagger$ is verified in Supplemental Appendix D with the covariance matrix in Assumption C6$^\dagger$(ii) equal to

$$\begin{aligned}\Omega_G(\pi_1,\pi_2;\gamma_0) &= E_{\gamma_0}\frac{L'^2(Z_i'\zeta_0)}{L(Z_i'\zeta_0)(1-L(Z_i'\zeta_0))}h_{Z,i}(\pi_1,\pi_2)h_{Z,i}(\pi_1,\pi_2)', \text{ where} \\ h_{Z,i}(\pi_1,\pi_2) &= (h(X_i,\pi_1), h(X_i,\pi_2), Z_i')'. \quad (3.23)\end{aligned}$$

□

The following assumption is used in the proof of consistency of $\widehat{\pi}_n$ for the case where the true parameter $\beta_n$ satisfies $\beta_n \to 0$ and $n^{1/2}||\beta_n|| \to \infty$.

**Assumption C7.** The non-stochastic function $\eta(\pi;\gamma_0,\omega_0)$ is uniquely minimized over $\pi \in \Pi$ at $\pi_0$ $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$.

In Assumption C7, $\pi_0$ is non-random. Assumption C7 can be verified using the Cauchy-Schwarz inequality or a matrix version of it, see Tripathi (1999), when $K(\pi;\gamma_0)$ and $H(\pi;\gamma_0)$ take proper forms, as in our examples.

**Example 1 (cont.).** Assumption C7 is verified in this example as follows. By (2.4) and (3.15), when $\beta_0 = 0$,

$$H(\pi;\gamma_0) = E_{\gamma_0}\frac{L'^2(Z_i'\zeta_0)}{L(Z_i'\zeta_0)(1-L(Z_i'\zeta_0))}d_{\psi,i}(\pi)d_{\psi,i}(\pi)'. \tag{3.24}$$

By (2.4) and (3.10), when $\beta_0 = 0$,

$$K(\pi;\gamma_0) = E_{\gamma_0}\frac{L'^2(Z_i'\zeta_0)}{L(Z_i'\zeta_0)(1-L(Z_i'\zeta_0))}h(X_i,\pi_0)d_{\psi,i}(\pi). \tag{3.25}$$

Hence, when $\beta_0 = 0$,

$$K(\pi;\gamma_0)'H^{-1}(\pi;\gamma_0)K(\pi;\gamma_0) \leq E_{\gamma_0}\frac{L'^2(Z_i'\zeta_0)}{L(Z_i'\zeta_0)(1-L(Z_i'\zeta_0))}h^2(X_i,\pi_0) \tag{3.26}$$

by the matrix Cauchy-Schwarz inequality in Tripathi (1999). The "$\leq$" holds as an equality if and only if $h(X_i,\pi_0)a + d_{\psi,i}(\pi)'b = 0$ with probability 1 for some $a \in R$ and $b \in R^{d_\zeta+1}$ with $(a,b') \neq 0$. The "$\leq$" holds as an equality uniquely at $\pi = \pi_0$ because for any $\pi \neq \pi_0$, $P_{\gamma_0}(c'(h(X_i,\pi_0),h(X_i,\pi),Z_i')' = 0) < 1$ for any $c \neq 0$ by (3.12). □

# 4. Estimation Results

This section provides the asymptotic results of the paper for the extremum estimator $\widehat{\theta}_n$. The results are given under the drifting sequences of distributions defined in Section 2.3. Define a concentrated extremum estimator $\widehat{\psi}_n(\pi)$ $(\in \Psi(\pi))$ of $\psi$ 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 sample 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 extremum estimator $\widehat{\theta}_n$ in (2.7) 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 (2.7) 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 $\pi$ by Assumption S2(ii).

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 (3.21) 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* S1-S4, B1, B2, *and* C6 *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) $n\left(Q_n(\widehat{\theta}_n) - Q_{0,n}\right) \to_d \inf_{\pi \in \Pi} \xi(\pi; \gamma_0, b)$.

**Comments. 1.** The results of Theorem 4.1 and Theorem 4.2 below are like those of Theorems 5.1 and 5.2 of AC1. However, Theorems 4.1 and Theorem 4.2 are obtained under assumptions that are much more primitive and easier to verify, though less general, than the results in AC1. In particular, Assumptions S1-S4 impose conditions for fixed parameters, not conditions on the behavior of random variables under sequences of parameters. In addition, explicit formulae for the components of the asymptotic results are provided here based on the sample average form of $Q_n(\theta)$ that is considered.

**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 $\beta$ out of $\psi$. The

asymptotic distribution of $n^{1/2}\widehat{\beta}_n$ (without centering at $\beta_n$) under $\Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$ is given by $\tau_\beta(\pi^*(\gamma_0, b); \gamma_0, b)$.

**3.** Assumption C6 is not needed for Theorem 4.1(b).

Let

$$G^*(\gamma_0) \sim N(0_{d_\theta}, V(\gamma_0)). \tag{4.6}$$

**Theorem 4.2.** *Suppose Assumptions* S1-S3, B1, B2, *and* C7 *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) $n(Q_n(\widehat{\theta}_n) - Q_n(\theta_n)) \to_d -\frac{1}{2}G^*(\gamma_0)'J^{-1}(\gamma_0)G^*(\gamma_0)$.

# 5. QLR Confidence Sets

In this section, we consider CS's based on the quasi-likelihood ratio (QLR) statistic. We establish (i) the asymptotic distribution of the QLR statistic under the drifting sequences of distributions defined in Section 2.3, (ii) the asymptotic size of standard QLR CS's, which often are size-distorted, and (iii) the correct asymptotic size of robust QLR CS's, which are designed to be robust to the strength of identification. The proofs of the results given here rely on results given in Supplemental Appendix A and AC1.

## 5.1. Definition of the QLR Test Statistic

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

$$r(\theta) = \begin{bmatrix} r_1(\psi) \\ r_2(\pi) \end{bmatrix}, \tag{5.1}$$

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

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

$$\widetilde{\theta}_n(v) \in \Theta,\ r(\widetilde{\theta}_n(v)) = v,\ \text{and}\ Q_n(\widetilde{\theta}_n(v)) = \inf_{\theta \in \Theta: r(\theta) = v} Q_n(\theta) + o(n^{-1}). \tag{5.2}$$

The QLR test statistic for testing $H_0 : r(\theta) = v$ is

$$QLR_n(v) = 2n(Q_n(\widetilde{\theta}_n(v)) - Q_n(\widehat{\theta}_n))/\widehat{s}_n, \tag{5.3}$$

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

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 non-stochastic. 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, which we denote by $\chi^2_{d_r,1-\alpha}$.

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

$$CS^{QLR}_{r,n} = \{v \in r(\Theta) : QLR_n(v) \leq c_{n,1-\alpha}(v)\}. \tag{5.4}$$

## 5.2. QLR Assumptions

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

$$\Pi_r(v_2) = \{\pi \in \Pi : r_2(\pi) = v_2 \text{ for some } \theta = (\psi, \pi) \in \Theta\}. \tag{5.5}$$

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

We assume $r(\theta)$ satisfies:

**Assumption RQ1.** (i) $r(\theta)$ is continuously differentiable on $\Theta$.
(ii) $r_\theta(\theta)$ $(= (\partial/\partial\theta')r(\theta))$ is full row rank $d_r$ $\forall\theta \in \Theta$.
(iii) $r(\theta)$ satisfies (5.1).
(iv) $d_H(\Pi_r(v_2), \Pi_r(v_{0,2})) \to 0$ as $v_2 \to v_{0,2}$ $\forall v_{0,2} \in r_2(\Theta^*)$.
(v) $Q(\psi, \pi; \gamma_0)$ is continuous in $\psi$ at $\psi_0$ uniformly over $\pi \in \Pi$ (i.e., $\sup_{\pi\in\Pi} |Q(\psi, \pi; \gamma_0) - Q(\psi_0, \pi; \gamma_0)| \to 0$ as $\psi \to \psi_0$) $\forall\gamma_0 \in \Gamma$ with $\beta_0 = 0$.
(vi) $Q(\theta; \gamma_0)$ is continuous in $\theta$ at $\theta_0$ $\forall\gamma_0 \in \Gamma$ with $\beta_0 \neq 0$.

In Assumption RQ1(iv), $d_H$ denotes the Hausdorff distance. In Assumptions RQ1(iv) and (v), $Q(\theta; \gamma_0) = E_{\gamma_0}\rho(W_i, \theta)$.

Assumptions RQ1(i) and RQ1(ii) are standard and are not restrictive. Assumption RQ1(iii) rules out the case where any single restriction depends on both $\psi$ and $\pi$. This is restrictive. But, in some cases, a reparametrization can be used to obtain results for such restrictions, see AC1 for details. Assumption RQ1(iv) is not very restrictive and is easy to verify in most cases. Assumptions RQ1(v) and RQ1(vi) are not restrictive.

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

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

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

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

Results for the QLR test without imposing Assumption RQ2 could be obtained fairly straightforwardly from the results given below. Without Assumption RQ2, the asymptotic distribution of the QLR statistic would not be $\chi^2_{d_r}$ under strong or semi-strong identification. Rather, it would have a mixture of $\chi^2$ distributions with weights that depend on unknown parameters. One could simulate its distribution using estimates of the weights, rather than using the $\chi^2_{d_r}$, in those scenarios when the $\chi^2_{d_r}$ is employed below.

## 5.3. QLR Asymptotic Distributions

To obtain the asymptotic size of QLR CS's, we need to determine the limits of the coverage probabilities of the QLR CS's under all sequences $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ and $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ when the null hypotheses are true. That is, we need to know

these limits when $v = v_n = r(\theta_n)$ for $\gamma_n = (\theta_n, \phi_n)$ $\forall n \geq 1$. To obtain these coverage probabilities, we first determine the asymptotic null distributions of the QLR statistic under these sequences.

In the results below, we use the following notational simplifications:

$$QLR_n = QLR_n(v_n) \text{ and } \widetilde{\theta}_n = \widetilde{\theta}_n(v_n), \text{ where } v_n = r(\theta_n) \text{ and } \gamma_n = (\theta_n, \phi_n).^{13} \tag{5.6}$$

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

Next, we introduce the limit under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$ of the restricted concentrated criterion function after suitable normalization. Define the process $\{\xi_r(\pi; \gamma_0, b) : \pi \in \Pi\}$ by

$$\begin{aligned} \xi_r(\pi; \gamma_0, b) &= \xi(\pi; \gamma_0, b) + \frac{1}{2}\tau(\pi; \gamma_0, b)' P_\psi(\pi; \gamma_0)' H(\pi; \gamma_0) P_\psi(\pi; \gamma_0) \tau(\pi; \gamma_0, b), \text{ where} \\ P_\psi(\pi; \gamma_0) &= H^{-1}(\pi; \gamma_0) r_{1,\psi}(\psi_0)' \left( r_{1,\psi}(\psi_0) H^{-1}(\pi; \gamma_0) r_{1,\psi}(\psi_0)' \right)^{-1} r_{1,\psi}(\psi_0), \end{aligned} \tag{5.7}$$

$r_{1,\psi}(\psi) = (\partial/\partial\psi') r_1(\psi) \in R^{d_{r_1} \times d_\psi}$, and $\tau(\pi; \gamma_0, b)$ is defined in (4.3). The $d_\psi \times d_\psi$-matrix $P_\psi(\pi; \gamma_0)$ is an oblique projection matrix that projects onto the space spanned by the rows of $r_{1,\psi}(\psi_0)$.

The following Theorem shows that the QLR statistic converges in distribution to $\lambda_{QLR}(\gamma_0)/s(\gamma_0)$ under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, where $\lambda_{QLR}(\gamma_0)$ is defined by

$$\begin{aligned} \lambda_{QLR}(\gamma_0) &= G^*(\gamma_0)' J^{-1}(\gamma_0) P_\theta(\gamma_0)' J(\gamma_0) P_\theta(\gamma_0) J^{-1}(\gamma_0) G^*(\gamma_0), \\ P_\theta(\gamma_0) &= J^{-1}(\gamma_0) r_\theta(\theta_0)' \left( r_\theta(\theta_0) J^{-1}(\gamma_0) r_\theta(\theta_0)' \right)^{-1} r_\theta(\theta_0), \end{aligned} \tag{5.8}$$

$r_\theta(\theta_0) = (\partial/\partial\theta') r(\theta_0)$, and $J(\gamma_0)$ and $G^*(\gamma_0)$ are defined in (3.13) and (3.19), respectively. The $d_\theta \times d_\theta$-matrix $P_\theta(\gamma_0)$ is an oblique projection matrix that projects onto the space spanned by the rows of $r_\theta(\theta_0)$.

**Theorem 5.1.** *Suppose Assumptions* S1-S4, B1, B2, RQ1, *and* RQ3 *hold.*
(a) *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $||b|| < \infty$,
$QLR_n \to_d 2(\inf_{\pi \in \Pi_{r,0}} \xi_r(\pi; \gamma_0, b) - \inf_{\pi \in \Pi} \xi(\pi; \gamma_0, b))/s(\gamma_0)$.

[13] As a consequence of these definitions, the asymptotic results given below for the statistics $QLR_n$ and $\widetilde{\theta}_n$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ and under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ are results that hold when the restrictions are true.

(b) *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $QLR_n \to_d \lambda_{QLR}(\gamma_0)/s(\gamma_0)$ *provided Assumption* C7 *also holds.*

**Comment.** By Theorem 5.1(b) and some calculations, when Assumptions RQ2 also holds,

$$QLR_n \to_d \lambda_{QLR}(\gamma_0)/s(\gamma_0) \sim \chi^2_{d_r}. \tag{5.9}$$

## 5.4. Asymptotic Size of Standard QLR Confidence Sets

Here we establish the asymptotic size of a standard nominal $1-\alpha$ CS for $r(\theta) \in R^{d_r}$ obtained by inverting the QLR statistic, defined in (5.4), using the $\chi^2_{d_r}$ critical value. The asymptotic size is determined using Theorem 5.1 combined with Lemma 2.1 in AC1.

Let

$$\begin{aligned} &h = (b, \gamma_0),\ H = \{h = (b, \gamma_0) : ||b|| < \infty, \gamma_0 \in \Gamma \text{ with } \beta_0 = 0\}, \text{ and} \\ &QLR(h) = 2(\inf_{\pi \in \Pi_{r,0}} \xi_r(\pi; \gamma_0, b) - \inf_{\pi \in \Pi} \xi(\pi; \gamma_0, b))/s(\gamma_0) \end{aligned} \tag{5.10}$$

for $||b|| < \infty$. Note that $QLR(h)$ is the asymptotic distribution of $QLR_n$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ for $||b|| < \infty$ by Theorem 5.1(a). Let $c_{QLR,1-\alpha}(h)$ denote the $1-\alpha$ quantile of $QLR(h)$ for $h \in H$.

The asymptotic size results given below use the following df continuity assumption, which typically is not restrictive.

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

**Theorem 5.2.** *Suppose Assumptions* S1-S4, B1, B2, C7, RQ1-RQ3, *and* RQ4(i) *hold. Then, the asymptotic size of the standard nominal* $1-\alpha$ *QLR CS is*

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

**Comment.** Depending on the distribution of $\{QLR(h) : h \in H\}$, the standard QLR CS has asymptotic size equal to $1-\alpha$ or less than $1-\alpha$. Often, it is less than $1-\alpha$ and the standard QLR CS is size distorted.

## 5.5. Robust QLR Confidence Sets

In this section, we construct two QLR CS's that have correct asymptotic size. These CS's are *robust* to the strength of identification. We construct CS's for $r(\theta)$ by inverting a robust QLR test that combines the QLR test statistic with a robust critical value that differs from the standard strong-identification critical value, which is a $\chi^2_{d_r}$ quantile. The first robust CS uses the least favorable (LF) critical value. The second robust CS is introduced in AC1. It is more sophisticated and uses a data-dependent critical value. It is called a type 2 robust CS. It is smaller than the LF robust CS under strong identification.

### 5.5.1. Least Favorable Critical Value

The LF critical value is

$$c^{LF}_{QLR,1-\alpha} = \max\{\sup_{h\in H} c_{QLR,1-\alpha}(h), \chi^2_{d_r,1-\alpha}\}. \tag{5.11}$$

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 $\pi$ to be 3, then the supremum in (5.11) 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.[14,15]

When part of $\gamma$ 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 $\gamma$ with consistent estimators in the formulae in (5.11) 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 $\gamma$.[16]

[14]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 is $H$ that is consistent with the null hypothesis $H_0 : r(\theta_0) = v$, where $\theta_0$ denotes the true value. The NI-LF critical value, denoted $c^{LF}_{QLR,1-\alpha}(v)$, is defined by replacing $H$ by $H(v)$ in (5.11) when the null hypothesis value is $r(\theta_0) = v$. Note that $v$ takes values in the set $V_r = \{v_0 : r(\theta_0) = v_0 \text{ for some } h = (b, \gamma_0) \in H\}$.

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

[16]For example, if $\zeta$ is consistently estimated by $\widehat{\zeta}_n$, then $H$ is replaced by $\widehat{H}_n = \{h = (b, \gamma) \in H :$

### 5.5.2. Type 2 Robust Critical Value

Next, we improve on the LF critical value by employing an identification category selection (ICS) procedure that uses the data to determine whether $b$ is finite.[17]

By Theorem 4.2, the asymptotic covariance matrix of $\widehat{\theta}_n$ under strong identification is $\Sigma(\gamma_0) = J^{-1}(\gamma_0)' V(\gamma_0) J^{-1}(\gamma_0)$. Let $\widehat{\Sigma}_n = \widehat{J}_n^{-1}(\widehat{\theta}_n)\widehat{V}_n(\widehat{\theta}_n)\widehat{J}_n^{-1}(\widehat{\theta}_n)$ denote an estimator of $\Sigma(\gamma_0)$, where $\widehat{J}_n(\theta)$ and $\widehat{V}_n(\theta)$ are estimators with probability limits $J(\theta;\gamma_0)$ and $V(\theta;\gamma_0)$, respectively, under $\gamma_n \to \gamma_0$ and $J(\gamma_0) = J(\theta_0;\gamma_0)$ and $V(\gamma_0) = V(\theta_0;\gamma_0)$. For brevity, we state the formal consistency Assumptions V1 and V2 concerning $\widehat{J}_n(\theta)$ and $\widehat{V}_n(\theta)$ in Supplemental Appendix B.

**Example 1 (cont.)** In this example, we estimate $J(\gamma_0) = V(\gamma_0)$ by $\widehat{J}_n(\widehat{\theta}_n) = \widehat{V}_n(\widehat{\theta}_n)$, where

$$\widehat{J}_n(\theta) = \widehat{V}_n(\theta) = n^{-1}\sum_{i=1}^{n} \frac{L_i'^2(\theta)}{L_i(\theta)(1-L_i(\theta))} d_i(\pi) d_i(\pi)'. \tag{5.12}$$

□

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

$$A_n = \left( n\widehat{\beta}_n' \widehat{\Sigma}_{\beta\beta,n}^{-1} \widehat{\beta}_n / d_\beta \right)^{1/2}, \tag{5.13}$$

where $\widehat{\Sigma}_{\beta\beta,n}$ is the upper left $d_\beta \times d_\beta$ block of $\widehat{\Sigma}_n$. We use $A_n$ to assess the strength of identification.

Now, we define the type 2 robust critical value, which 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 non-increasing 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 $c > 0$.[18] For example, in the binary choice example, we use

$\gamma = (\beta, \widehat{\zeta}_n, \pi, \phi)\}$. If a plug-in NI-LF critical value is employed, $H(v)$ is replaced by $H(v) \cap \widehat{H}_n$, where $H(v)$ is defined in footnote 14. Note that the parameter $b$ is not consistently estimable, so it cannot be replaced by a consistent estimator.

[17] When the null hypothesis specifies the value of $\beta$, it is not necessary to use an ICS procedure. Instead, we recommend using a (possibly plug-in) NI-LF critical value, see footnotes 14 and 16.

[18] If $c_{QLR,1-\alpha}^{LF} = \infty$, one should take $s(x)$ to equal 0 for $x$ sufficiently large and define $\infty \times 0$ in (5.14) to equal 0. Then, the critical value $\widehat{c}_{QLR,1-\alpha,n}$ is infinite if $A_n$ is small and is finite if $A_n$ is sufficiently large.

the function $s(x) = \exp(-x/2)$.

The type 2 robust critical value is

$$\widehat{c}_{QLR,1-\alpha,n} = \begin{cases} c_B & \text{if } A_n \leq \kappa \\ c_S + [c_B - c_S] \cdot s(A_n - \kappa) & \text{if } A_n > \kappa, \text{ where} \end{cases}$$
$$c_B = c^{LF}_{QLR,1-\alpha} + \Delta_1, \; c_S = \chi^2_{d_r,1-\alpha} + \Delta_2, \tag{5.14}$$

and $\Delta_1 \geq 0$ and $\Delta_2 \geq 0$ are asymptotic size-correction factors that are defined below. Here, "$B$" denotes Big, and "$S$" denotes Small. When $A_n \leq \kappa$, $\widehat{c}_{QLR,1-\alpha,n}$ equals the LF critical value $c^{LF}_{QLR,1-\alpha}$ plus a size-correction factor $\Delta_1$. When $A_n > \kappa$, $\widehat{c}_{QLR,1-\alpha,n}$ is a convex combination of $c^{LF}_{QLR,1-\alpha} + \Delta_1$ and $\chi^2_{d_r,1-\alpha} + \Delta_2$, where $\Delta_2$ is another size-correction factor and 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

$$A(h) = \left(\tau_\beta(\pi^*; \gamma_0, b)' \Sigma^{-1}_{\beta\beta}(\pi^*; \gamma_0) \tau_\beta(\pi^*; \gamma_0, b)/d_\beta\right)^{1/2}, \tag{5.15}$$

where $\pi^*$ 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 $\Sigma(\theta; \gamma_0) = J^{-1}(\theta; \gamma_0) V(\theta; \gamma_0) J^{-1}(\theta; \gamma_0)$.[19,20,21]

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

$$\begin{aligned} NRP(\Delta_1, \Delta_2; h) &= P(QLR(h) > c_B \;\&\; A(h) \leq \kappa) + P(QLR(h) > c_A(h) \;\&\; A(h) > \kappa) \\ &= P(QLR(h) > c_B) + P(QLR(h) \in (c_A(h), c_B] \;\&\; A(h) > \kappa), \text{ where} \\ c_A(h) &= c_S + (c_B - c_S) \cdot s(A(h) - \kappa). \end{aligned} \tag{5.16}$$

The constants $\Delta_1$ and $\Delta_2$ are chosen such that $NRP(\Delta_1, \Delta_2; h) \leq \alpha \; \forall h \in H$. In par-

[19] The convergence in distribution follows from Theorem 4.1(a) and Assumption V1.

[20] In the vector $\beta$ case, $\Sigma^{-1}_{\beta\beta}(\pi^*; \gamma_0)$ is replaced in (5.15) by a slightly different expression, see footnote 51 of AC1. When the type 2 robust critical value is considered in the vector $\beta$ 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 $\omega_0$ affects the asymptotic distribution of $A_n$.

[21] Alternatively to the ICS statistic $A_n$, one can use a NI-ICS statistic $A_n(v)$, which employs the restricted estimator $\widetilde{\beta}_n(v)$ of $\beta$ in place of $\widehat{\beta}_n$ and a different weight matrix. See AC1 for details.

ticular, we define $\Delta_1 = \sup_{h \in H_1} \Delta_1(h)$, where $\Delta_1(h) \geq 0$ solves $NRP(\Delta_1(h), 0; h) = \alpha$ (or $\Delta_1(h) = 0$ if $NRP(0, 0; h) < \alpha$), $H_1 = \{(b, \gamma_0) : (b, \gamma_0) \in H \ \& \ ||b|| \leq ||b_{\max}|| + D\}$, $b_{\max}$ is defined such that $c_{QLR,1-\alpha}(h)$ 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 non-negative 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 $NRP(\Delta_1, 0; h) < \alpha$).[22,23] As defined, $\Delta_1$ and $\Delta_2$ can be computed sequentially, which is computationally convenient.

Given the definitions of $\Delta_1$ and $\Delta_2$, the asymptotic rejection probability is always less than or equal to the nominal level $\alpha$ and it is close to $\alpha$ when $h$ is close to $h_{\max}$ (due to the adjustment by $\Delta_1$) and when $||b||$ is large (due to the adjustment by $\Delta_2$).

The type 2 robust critical value can be improved by employing NI and/or plug-in versions of it, denoted by $\widehat{c}_{QLR,1-\alpha,n}(v)$. These are defined by replacing $c^{LF}_{QLR,1-\alpha}$ in (5.14) by the NI-LF or plug-in NI-LF critical value and making $\Delta_1$ and $\Delta_2$ depend on the null value $v$. We recommend employing these versions whenever possible because they lead to smaller CS's.

The asymptotic sizes of QLR CS's based on LF and type 2 robust critical values (possibly with NI and/or plug-in features) are always $1 - \alpha$ or greater and are exactly $1 - \alpha$ under some mild df continuity conditions. For brevity, these results are stated formally in Theorem 12.1 in Supplemental Appendix B.

For any given value of $\kappa$, the type 2 robust CS has correct asymptotic size due to the choice of $\Delta_1$ and $\Delta_2$. In consequence, a good choice of $\kappa$ depends on the false coverage probabilities (FCP's) of the robust CS. (An 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 shows that if a reasonable value of $\kappa$ is chosen, such as $\kappa = 1.5$ or 2.0, the FCP's of type 2 robust CS's are not sensitive to deviations from this value of $\kappa$. The reason is that the size-correction constants $\Delta_1$ and $\Delta_2$ have to adjust as $\kappa$ is changed in order to maintain correct asymptotic size. The adjustments of $\Delta_1$ and $\Delta_2$ offset the

[22] 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 $\Delta_1$ and is typically strictly decreasing and continuous in $\Delta_1$. 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 $NRP(\Delta_1(h), 0; h) \leq \alpha$ and $\Delta_1(h) \geq 0$ is as small as possible. Analogous comments apply to the equation $NRP(\Delta_1, \Delta_2(h); h) = \alpha$ and the definition of $\Delta_2(h)$.

[23] When the LF critical value is achieved at $||b|| = \infty$, i.e., $\chi^2_{d_r,1-\alpha} \geq \sup_{h \in H} c_{QLR,1-\alpha}(h)$, the standard asymptotic critical value $\chi^2_{d_r,1-\alpha}$ yields a test or CI with correct asymptotic size and constants $\Delta_1$ and $\Delta_2$ 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 $h_{\max} = (b_{\max}, \gamma_{\max}) \in H$.

effect of changing $\kappa$.

One can select $\kappa$ in a simple way, i.e., by taking $\kappa = 1.5$ or 2.0, or one can select $\kappa$ in a more sophisticated way that explicitly depends on FCP's. (See Supplemental Appendix B for a description of the more sophisticated method.) Both methods yield quite similar results for the cases that we have considered.

# 6. t Confidence Intervals

In this section, we introduce confidence intervals (CI's) based on $t$ statistics. Theoretical results for the $t$ CI's are obtained using the asymptotic distributions of the unrestricted estimator $\widehat{\theta}_n$ in Theorems 4.1 and 4.2. Details are given in AC1.[24] In this section, the number of restrictions, $d_r$, equals one.

The $t$ statistic takes the form

$$T_n(v) = \frac{n^{1/2}(r(\widehat{\theta}_n) - v)}{(r_\theta(\widehat{\theta}_n)B^{-1}(\widehat{\beta}_n)\widehat{\Sigma}_n B^{-1}(\widehat{\beta}_n)r_\theta(\widehat{\theta}_n)')^{1/2}}, \tag{6.1}$$

where $r_\theta(\theta) = (\partial/\partial\theta')r(\theta) \in R^{d_r \times d_\theta}$ and $\widehat{\Sigma}_n$ is defined as in Section 5.5. Although this definition of the $t$ statistic involves $B^{-1}(\widehat{\beta}_n)$, it is the same as the standard definition used in practice, see AC1.

For testing $H_0 : r(\theta) = v$ against two-sided, upper-one-sided, and lower-one-sided alternatives, the $t$ statistic is $|T_n(v)|$, $T_n(v)$, and $-T_n(v)$, respectively.

Let $c_{n,1-\alpha}(v)$ denote a nominal level $1-\alpha$ critical value to be used with the $t$ test statistic. It may be stochastic or non-stochastic. The usual choice, based on the asymptotic distribution of the $t$ statistic under standard regularity conditions, is the $1-\alpha/2$ or $1-\alpha$ quantile of the $N(0,1)$ distribution: $z_{1-\alpha/2}$ or $z_{1-\alpha}$ depending on whether a two-sided or one-sided CI is desired.

Critical values that deliver robust $t$ CS's for $r(\theta)$ that have correct asymptotic size can be constructed using the same approaches as in Section 5.5.

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

$$CS^t_{r,n} = \{v \in r(\Theta) : |T_n(v)| \leq c_{n,1-\alpha}(v)\}. \tag{6.2}$$

[24] See Theorems 4.1, 4.4(a), and 5.1(a) in Sections 4.1, 4.7, and 5, respectively, in AC1. Lemma 11.1 in Supplemental Appendix A shows that Assumptions B1, B2, and S1-S3 imply the high-level conditions B3, C1-C4, C8, and D1-D3 employed in the results just stated in AC1.

For one-sided $t$ CI's, $|T_n(v)|$ is replaced by $T_n(v)$ or $-T_n(v)$ depending on whether one desires an upper or lower CI, respectively.

# 7. Smooth Transition Autoregressive (STAR) Model

## 7.1. STAR Model and Criterion Function

In this section, we apply the results above to the STAR model. This model and its applications are considered in Luukkonen, Saikkonen, and Teräsvirta (1988), Teräsvirta and Anderson (1992), and Teräsvirta (1994) among others. To fit the STAR model into our identification set-up, we write the model as

$$\begin{aligned} Y_t &= X_t'\zeta + X_t'\beta \cdot m(Z_t, \pi) + U_t, \text{ where} \\ X_t &= (1, Y_{t-1}, ..., Y_{t-p})', \ Z_t = Y_{t-d}, \end{aligned} \tag{7.1}$$

$\{Y_t : t = 1, ..., n\}$ are observed random variables, $\{U_t : t = 1, ..., n\}$ are unobserved innovations, $m(\cdot, \cdot)$ is a known transition function, and $W_t = (Y_t, X_t', Z_t')'$. We assume $p$ and $d$ are known and $1 \leq d \leq p$.

As in the literature, two different forms of the transition function $m(Z_t, \pi)$ are considered. The first one is the logistic function

$$m(Z_t, \pi) = (1 + \exp[-\pi_1(Z_t - \pi_2)])^{-1} \tag{7.2}$$

and the second one is the exponential function

$$m(Z_t, \pi) = 1 - \exp[-\pi_1(Z_t - \pi_2)^2], \tag{7.3}$$

where $\pi = (\pi_1, \pi_2)' \in R^2$, $\pi_1 > 0$ measures the slope of the transition, and $\pi_2$ measures the location of the transition. For both the logistic function and the exponential function, $m(Z_t, \pi) \in [0, 1]$.

We consider the LS estimator of $\theta = (\beta, \zeta, \pi)$. The LS sample criterion function is

$$Q_n(\theta) = n^{-1} \sum_{t=1}^{n} U_t^2(\theta)/2, \text{ where } U_t(\theta) = Y_t - X_t'\zeta - X_t'\beta \cdot m(Z_t, \pi). \tag{7.4}$$

The LS estimator of $\theta$ minimizes $Q_n(\theta)$ over $\theta \in \Theta$. The optimization parameter space

$\Theta$ takes the form

$$\Theta = \{(\beta, \zeta, \pi) : \beta \in \mathcal{B},\ \zeta \in \mathcal{Z}(\beta),\ \pi \in \Pi\}. \tag{7.5}$$

We show in Supplemental Appendix E that under the assumptions given below Assumptions S1-S4, B1, B2, C6 , and C7 hold and Assumptions V1 and V2 in Supplemental Appendix B also hold. Hence, all of the asymptotic results given above apply to the STAR model considered here.

The distribution of $\{U_t : t = ..., -1, 0, 1, ...\}$ is $\phi \in \Phi$, where $\Phi$ is a compact metric space with some metric $d_\Phi$ that induces weak convergence of the bivariate distributions $(Y_t, Y_{t+m})$ for all $t, m \geq 1$.[25] In this model, $\phi$ is an infinite-dimensional nuisance parameter. The true value of $\gamma = (\theta, \phi)$, denoted by $\gamma_0$, belongs to a compact set $\Gamma$. Let $\mathcal{F}_t$ denote some increasing set of sigma-fields to which $U_t$ and $Y_t$ are adapted. The data generating process (DGP) is assumed to satisfy Assumption STAR1 below.

**Assumption STAR1.** (i) $E_{\gamma_0}(U_t|\mathcal{F}_{t-1}) = 0$ a.s., $E_{\gamma_0}(U_t^2|\mathcal{F}_{t-1}) = \sigma^2$ a.s. with $\sigma^2 > 0$, and $\sup_{t\geq 1} E_{\gamma_0}|U_t|^{4+\varepsilon} \leq C < \infty\ \forall \gamma_0 \in \Gamma$.
(ii) Under $\gamma_0$, $\{Y_t : t = 1, ..., n\}$ is a strictly stationary and strong mixing sequence with mixing coefficients $\alpha_m \leq Cm^{-A}$ for some $A > d_\theta q/(q - d_\theta)$ and $q > d_\theta = 2p + 4$, $\forall \gamma_0 \in \Gamma$.

By Bhattacharya and Lee (1995), a set of sufficient conditions for Assumption STAR1(ii) is (i) $\{U_t : t = ..., -1, 0, 1, ...\}$ is a sequence of i.i.d. real-valued random variables, (ii) the distribution of $U_t$ is absolutely continuous wrt the Lebesgue measure and has a density function that is positive almost everywhere, (iii) $E_{\gamma_0}|U_t| < \infty$, and (iv) $\sum_{i=1}^{p}(|\zeta_i| + |\beta_i|) < 1$, where $\zeta = (\zeta_{int}, \zeta_1, ..., \zeta_p)$, $\beta = (\beta_{int}, \beta_1, ..., \beta_p)$, and $\zeta_{int}$ and $\beta_{int}$ are the intercepts when $m(\cdot, \cdot) = 0$ and 1, respectively.

Let $m_\pi(Z_t, \pi) = (m_{\pi,1}(Z_t, \pi), m_{\pi,2}(Z_t, \pi))' \in R^2$ and $m_{\pi\pi}(Z_t, \pi) \in R^{2\times 2}$ denote the first- and second-order partial derivatives of $m(Z_t, \pi)$ wrt $\pi$. Suppose $||m_{\pi\pi}(Z_t, \pi_1) - m_{\pi\pi}(Z_t, \pi_2)|| \leq M_{\pi\pi}(Z_t)\delta$ for any $\pi_1, \pi_2 \in \Pi$ and $||\pi_1 - \pi_2|| \leq \delta$ and $M_{\pi\pi}(Z_t)$ satisfies Assumption STAR2(iii) below. In Assumption STAR2, the constants $\varepsilon > 0$ and $0 < C < \infty$ do not depend on $\gamma_0$.

[25] For example, the metric $d_\Phi$ can be defined as follows. Let $\{u_t\}_1$ and $\{u_t\}_2$ denote two infinite $\{u_t : t = ..., -1, 0, 1, ...\}$ sequences. The distribution of $\{u_t\}_i$ is denoted by $\mathcal{L}(\{u_t\}_i)$ for $i = 1, 2$. Let $Y_t(\{u_t\}_i, \theta)$ denote $Y_t$ generated with the innovation sequence $\{u_t\}_i$ and $\theta$, for $i = 1, 2$. Let $\mathcal{L}(Y_t(\{u_t\}_i, \theta), Y_{t+m}(\{u_t\}_i, \theta))$ denote the bivariate distribution of $(Y_t(\{u_t\}_i, \theta), Y_{t+m}(\{u_t\}_i, \theta))$ for $i = 1, 2$. The metric $d_\Phi$ can be defined as $d_\Phi(\mathcal{L}(\{u_t\}_1), \mathcal{L}(\{u_t\}_2)) = \sup_{m\geq 1} \sup_{\theta\in\Theta^*} d_2(\mathcal{L}(Y_t(\{u_t\}_1, \theta), Y_{t+m}(\{u_t\}_1, \theta)), \mathcal{L}(Y_t(\{u_t\}_2, \theta), Y_{t+m}(\{u_t\}_2, \theta)))$, where $\Theta^*$ is the true parameter space for $\theta$ and $d_2$ is some metric on the space of bivariate distributions that induces weak convergence.

**Assumption STAR2.** (i) $P_{\gamma_0}([X_t', X_t'm(Z_t,\pi), X_t'm(Z_t,\overline{\pi})]a = 0) < 1 \ \forall a \neq 0 \in R^{3d_\beta}$, $\forall \pi, \overline{\pi} \in \Pi$ with $\pi \neq \overline{\pi}$.
(ii) $P_{\gamma_0}([X_t', X_t'm(Z_t,\pi), X_t'm_{\pi,1}(Z_t,\pi), X_t'm_{\pi,2}(Z_t,\pi)]a = 0) < 1 \ \forall a \neq 0 \in R^{4d_\beta}$ and $\forall \pi \in \Pi$.
(iii) $E_{\gamma_0} \sup_{\pi\in\Pi}(|Y_t|^{4q} + ||m_\pi(Z_t,\pi)||^{4q} + ||m_{\pi\pi}(Z_t,\pi)||^{2q} + ||M_{\pi\pi}(Z_t)||^{2q}) \leq C$, where $q$ is as in Assumption STAR1.

Let $G(\cdot;\gamma_0)$ be 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 $\pi_1, \pi_2 \in \Pi$, where

$$\begin{aligned} \Omega(\pi_1,\pi_2;\gamma_0) &= E_{\gamma_0} U_t^2 d_{\psi,t}(\pi_1) d_{\psi,t}(\pi_2)' \text{ and} \\ d_{\psi,t}(\pi) &= (X_t'm(Z_t,\pi), X_t')'. \end{aligned} \tag{7.6}$$

Define a "weighted non-central chi-square" process $\{\xi(\cdot;\gamma_0,b) : \pi \in \Pi\}$ and a Gaussian process $\{\tau_\beta(\cdot;\gamma_0,b) : \pi \in \Pi\}$ by

$$\begin{aligned} \xi(\pi;\gamma_0,b) &= -\frac{1}{2}\left(G(\pi;\gamma_0) + K(\pi;\gamma_0)\, b\right)' H^{-1}(\pi;\gamma_0)\left(G(\pi;\gamma_0) + K(\pi;\gamma_0)b\right) \text{ and} \\ \tau_\beta(\pi;\gamma_0,b) &= -S_\beta H^{-1}(\pi;\gamma_0)\left(G(\pi;\gamma_0) + K(\pi;\gamma_0)b\right), \text{ where} \\ K(\pi;\gamma_0) &= -E_{\gamma_0} d_{\psi,t}(\pi) d_{\psi,t}(\pi_0)' \cdot S_\beta', \ S_\beta = [I_{d_\beta} : 0] \in R^{d_\beta \times d_\psi}, \text{ and} \\ H(\pi;\gamma_0) &= E_{\gamma_0} d_{\psi,t}(\pi) d_{\psi,t}(\pi)'. \end{aligned} \tag{7.7}$$

The quantities in (7.7) appear in Theorem 4.1.

**Assumption STAR3.** (i) 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 $\Pi$ 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 $||b|| < \infty$.
(ii) $P_{\gamma_0}(\tau_\beta(\pi^*(\gamma_0,b);\gamma_0,b) = 0) = 0 \ \forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$ and $\forall b$ with $||b|| < \infty$.

**Lemma 7.1.** *When $X_t = Y_{t-k}$ for some $k \geq 1$ or $X_t = 1$, Assumption* STAR2(i) *implies Assumption* STAR3(i).

## 7.2. Parameter Space

The true parameter space for $\theta = (\beta, \zeta, \pi)$ is

$$\Theta^* = \{(\beta, \zeta, \pi) : \beta \in \mathcal{B}^*,\ \zeta \in \mathcal{Z}^*(\beta),\ \pi \in \Pi^*\}. \tag{7.8}$$

In (7.8), $\Theta^*$ is not a product space. For any $\beta \in \mathcal{B}^*$, $\zeta^*$ belongs to $\mathcal{Z}^*(\beta)$ which is defined such that $\{Y_t : 1 \leq t \leq n\}$ is a strictly stationary and strong mixing sequence as in Assumption STAR1.

For any $\theta_0 \in \Theta^*$, let $\Phi(\theta_0) \subset \Phi$ denote the true parameter space for the nuisance parameter $\phi$. The true parameter spaces $\Theta^*$ and $\Phi(\theta_0)$ are assumed to satisfy Assumption STAR4.

**Assumption STAR4.** (i) $\Theta^*$ is compact.
(ii) $0_{d_\theta} \in int(\mathcal{B}^*)$.
(iii) $\Pi^* = \Pi_1^* \times \Pi_2^*$, where $\pi_1 \geq \varepsilon$ for some $\varepsilon > 0$ $\forall \pi_1 \in \Pi_1^*$.
(iv) For some set $\mathcal{Z}_0^*$ and some $\delta > 0$, $\mathcal{Z}^*(\beta) = \mathcal{Z}_0^*$ $\forall ||\beta|| < \delta$.

The parameter space $\Gamma$ is defined to be such that for any $\theta_0 \in \Theta^*$ and $\phi_0 \in \Phi(\theta_0)$, $\gamma_0 = (\theta_0, \phi_0) \in \Gamma$ satisfies Assumptions STAR1-STAR4. We also assume $\Gamma$ is compact.

Assumption STAR4(ii) guarantees that the region of non-identification ($\beta = 0$) and near lack of identification ($||\beta||$ close to 0) is in the true parameter space. Assumption STAR4(iii) bounds the true parameter space of $\pi_1$ away from 0 because our focus is on the weak identification of $\pi$ that occurs when $\beta$ is close to 0, rather than a different sort of weak identification that occurs when $\pi_1$ is close to 0. Assumption STAR4(iv) is employed in the verification of Assumption B2(iii).

The optimization parameter space $\Theta$ defined in (7.5) is assumed to satisfy Assumption STAR5 below. Let $\Psi = \{(\beta, \zeta) : \beta \in \mathcal{B} \text{ and } \zeta \in \mathcal{Z}(\beta)\}$.

**Assumption STAR5.** (i) $int(\Theta) \supset \Theta^*$.
(ii) $\Theta, \mathcal{B}, \Pi$, and $\Psi$ are compact, $\mathcal{Z}(\beta)$ is compact $\forall \beta \in \mathcal{B}$.
(iii) For some set $\mathcal{Z}_0$ and some $\delta > 0$, $\mathcal{Z}(\beta) = \mathcal{Z}_0$ $\forall ||\beta|| < \delta$ and $int(\mathcal{Z}_0) \supset \mathcal{Z}_0^*$, where $\mathcal{Z}_0^*$ is as in Assumption STAR4(iv).

For the STAR model, the quantities $\rho_\psi(W_i, \theta), ..., \rho_\theta^\dagger(W_i, \theta), ...$, and the variance matrix estimators are specified in the Appendix.

# 8. Numerical Results

In this section, we provide asymptotic and finite-sample simulation results for the STAR model and the binary choice model.

## 8.1. Numerical Results for the STAR Model

The STAR model considered is

$$Y_t = \zeta_1 + \zeta_2 Y_{t-1} + \beta \cdot m(Y_{t-1}, \pi) + U_t, \tag{8.1}$$

with $m(x, \pi) = x(1 + \exp(-10(x - \pi)))^{-1}$, $\{U_t : t = 1, ..., n\}$ are i.i.d., and $U_t \sim N(0, 1)$. For illustrative purposes and to ease the computational burden, we use the constant 10 in the transition function, rather than introduce another parameter. Note that a fixed a value of the transition parameter is used in the empirical work in Lundbergh and Teräsvirta (2006). Identification issues are evident in the results given below even without a second parameter in the transition function. The addition of another parameter will exacerbate the identification issues. The true values of $\zeta_1$ and $\zeta_2$ are $-1$ and 0.5, respectively. The true parameter space for $\pi$ is $[-3.5, -1.5]$ and the optimization space for $\pi$ is $[-4, -1]$. The number of simulation repetitions is 20,000.[26]

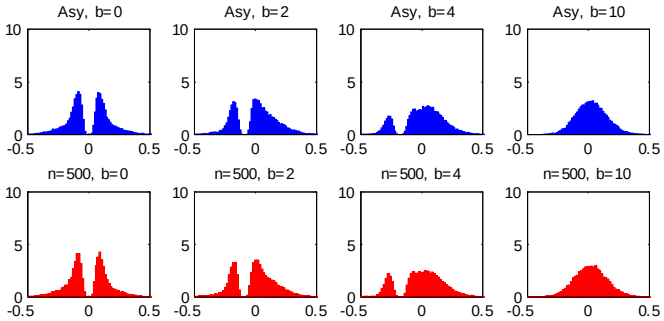


Figure 1. Asymptotic and Finite-Sample ($n = 500$) Densities of the Estimator of $\beta$ in the STAR Model when $\pi_0 = -1.5$.

Figures 1 and 2 provide the asymptotic and finite-sample densities of the ML estimators of $\beta$ and $\pi$ in the STAR model when the true $\pi$ value is $\pi_0 = -1.5$. Each

[26] For the STAR model, the discrete values of $b$ for which computations are made run from 0 to 12, with a grid of 0.2 for $b$ between 0 and 5, a grid of 0.5 for $b$ between 5 and 8, and a grid of 1 for $b$ between 8 and 12.

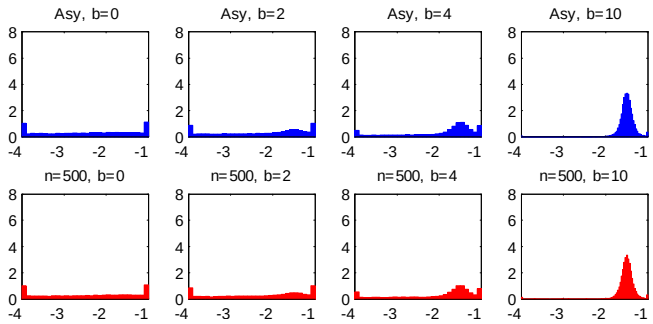


Figure 2. Asymptotic and Finite-Sample ($n = 500$) Densities of the Estimator of $\pi$ in the STAR Model when $\pi_0 = -1.5$.

Figure gives the densities for $b = 0$, 2, 4, and 10, where $b$ indexes the magnitude of $\beta$. Specifically, for the finite-sample results, $b = n^{1/2}\beta$. In these Figures, the finite-sample size considered is $n = 500$. Figures S-1 and S-2 in Supplemental Appendix C provide analogous results for $\pi_0 = -3.0$.

Figure 1 shows that the ML estimator of $\beta$ has a bi-modal distribution that is very far from a normal distribution in the unidentified and weakly-identified cases. Figure 2 shows that there is a build-up of mass at the boundaries of the optimization space for the estimator of $\pi$ 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 C provide the asymptotic and finite-sample ($n = 500$) densities of the $t$ and QLR statistics for $\beta$ and $\pi$ in the STAR model when $\pi_0 = -1.5$. These Figures show that in the case of weak identification the $t$ and QLR statistics are not well approximated by standard normal and $\chi^2_1$ 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 $\beta$ and $\pi$ in the STAR model as a function of $b$ for $\pi_0 = -1.5$, $-2.0$, $-3.0$, and $-3.5$. For the $|t|$ statistic concerning $\beta$, for small to medium $b$ values, the graphs exceed the 0.95 quantiles under strong identification (given by the horizontal black lines). This implies that tests and CI's that employ the $|t|$ statistic for $\beta$ and the standard critical value (based on the normal distribution) have incorrect size. The same pattern emerges for the QLR statistic for $\beta$ (although the quantile graphs are slightly below the black line for a range of $b$ around 4 when $\pi_0 = -3.0$ and $\pi_0 = -3.5$). The graphs in Figure 3(b) imply that tests and CI's that employ the QLR statistic for $\beta$ and the standard critical value (based on the $\chi^2_1$ distribution) have incorrect size due to the under-coverage for $b$

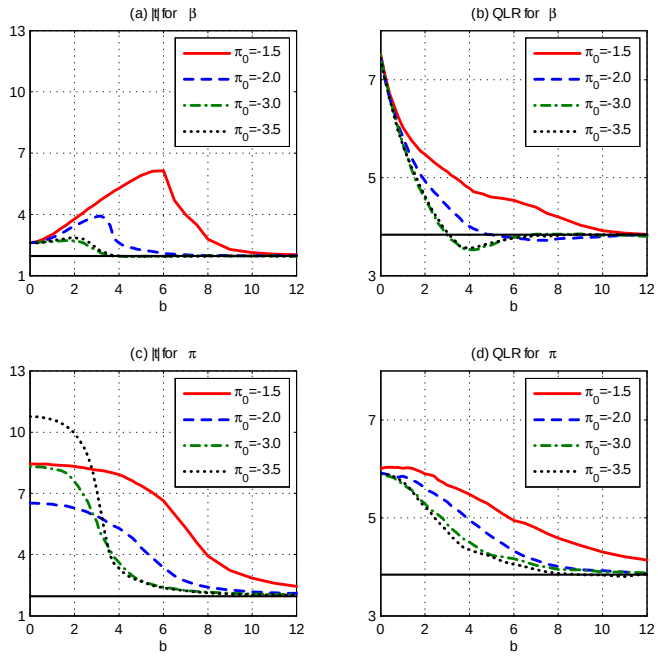


Figure 3. Asymptotic 0.95 Quantiles of the $|t|$ and QLR Statistics for Tests Concerning $\beta$ and $\pi$ in the STAR Model.

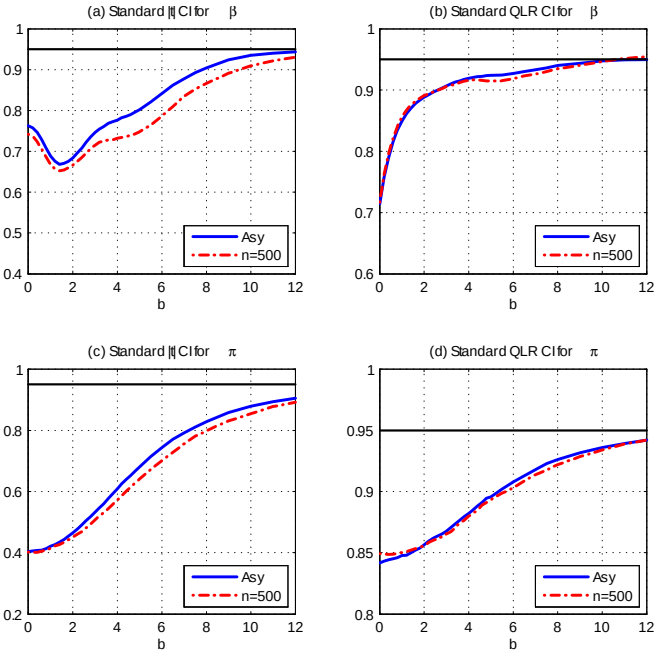


Figure 4. Coverage Probabilities of Standard $|t|$ and QLR CI's for $\beta$ and $\pi$ in the STAR Model when $\pi_0 = -1.5$.

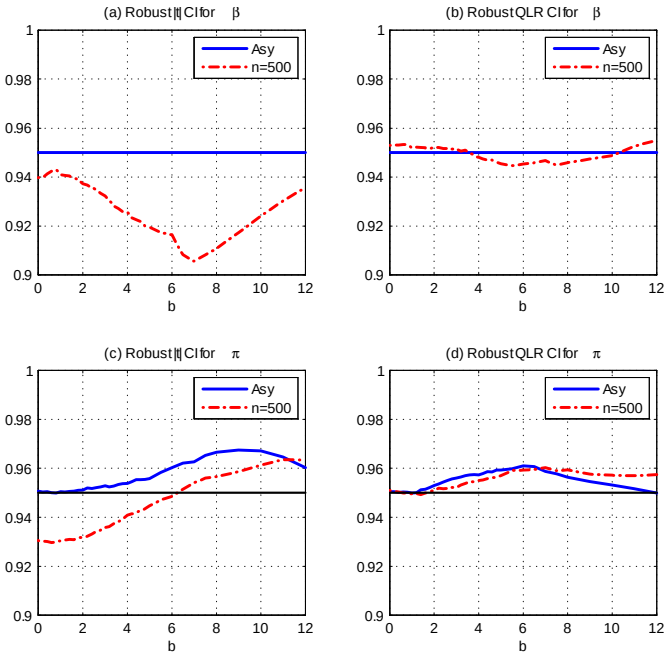


Figure 5. Coverage Probabilities of Robust $|t|$ and QLR CI's for $\beta$ and $\pi$ in the STAR Model when $\pi_0 = -1.5$, $\kappa = 2.5$, $D = 1$, and $s(x) = \exp(-x)$.

values around 0. Given the heights of the graphs in Figure 3(c) and 3(d), tests and CI's that employ the $|t|$ and QLR statistic for $\pi$ and the standard critical value also have incorrect asymptotic size.

Figure 4 reports the asymptotic and finite-sample CP's of nominal 0.95 standard $|t|$ and QLR CI's for $\beta$ and $\pi$ in the STAR model when $\pi_0 = -1.5$. For example, the smallest asymptotic and finite-sample CP's (over $b$) are around 0.67 for the $|t|$ CI for $\beta$ and 0.40 for the $|t|$ CI for $\pi$. The corresponding values for the QLR CI's are 0.72 for $\beta$ and 0.84 for $\pi$. Hence, the size distortions for the standard $|t|$ and QLR CI's for $\beta$ are similar. But, for the CI's for $\pi$, the size distortion of the standard QLR CI (both asymptotic and finite sample) is noticeably smaller than that of the standard $|t|$ CI. Note that the asymptotic CP's provide a very good approximation to the finite-sample CP's. Figure S-7 in Supplemental Appendix C provides analogous results for $\pi_0 = -3.0$.

Next, we consider CI's that are robust to weak identification. For the robust CI for $\beta$, we impose the null value of $b = n^{1/2}\beta_0$, where $\beta_0$ is the true value of $\beta$ under the null. With the knowledge of $b$ under the null, no identification category selection procedure is needed. Furthermore, the NI-LF critical value for the robust QLR CI for $\beta$ is as in (5.11), but with $h$ and $H$ replaced by $\pi$ and $\Pi$, respectively, resulting in a smaller LF

critical value. The same simplification applies to the NI-LF critical value for the robust $|t|$ CI for $\beta$.

As indicated in Figures 3(a) and 3(b), the NI-LF critical values for both $|t|$ and QLR CI's for $\beta$ are attained at $\pi_0 = -1.5$ for all $b$ values. In consequence, the robust $|t|$ and QLR CI's for $\beta$ are asymptotically similar when $\pi_0 = -1.5$, as shown in Figures 5(a) and 5(b). Figures 5(a) and 5(b) also report the finite-sample ($n = 500$) CP's of the robust $|t|$ and QLR CI's for $\beta$. For the former, the finite-sample CP is around 0.91 in the worst case, as opposed to 0.67 for the standard $|t|$ CI. For the latter, the finite-sample CP is around 0.95 for all $b$ values, showing that the robust QLR for $\beta$ has excellent finite-sample performance. Figures S-8(a) and S-8(b) in Supplemental Appendix C provide analogous results for $\pi_0 = -3.0$. The robust CI's for $\beta$ are not asymptotically similar when $\pi_0 = -3.0$, but they have correct asymptotic size and the asymptotic and finite-sample CP's are close for all $b$ values.

The robust CI's for $\pi$ are constructed with the null value $\pi_0$ imposed. Because $b$ is unknown, we apply the smooth transition in (5.14) to obtain critical values for the robust CI's for $\pi$. Figures 5(c) and 5(d) report the asymptotic and finite-sample CP's of the robust $|t|$ and QLR CI's for $\pi$ in the STAR model when $\pi_0 = -1.5$. To construct these robust CI's, we employ the transition function $s(x) = \exp(-x)$ and the constants $\kappa = 2.5$ and $D = 1$. The choices of $s(x)$ and $D$ were determined via some experimentation to be good choices in terms of yielding CP's that are relatively close to the nominal size 0.95 across different values of $b$. Given $s(x)$ and $D$, the choice of $\kappa$ was determined based on minimizing average FCP's. However, a wide range of $\kappa$ values yield similar results (because the constants $\Delta_1$ and $\Delta_2$ adjust to maintain correct asymptotic size as $\kappa$ is changed).

Figures 5(c) and 5(d) show that the robust CI's for $\pi$ have correct asymptotic size and the finite-sample sizes are reasonably close to 0.95 for both the $|t|$ and QLR CI's. Analogous results for the robust CI's for $\pi$ when $\pi_0 = -3.0$ are reported in Figures S-8(c) and S-8(d) in Supplemental Appendix C.

Besides $b$ and $\pi_0$, the construction of a robust CI also requires the $\zeta$ value in order to obtain the LF (or NI-LF) critical value through simulation. In the STAR model, $\zeta = (\zeta_1, \zeta_2)'$. Because $\zeta$ can be consistently estimated, we recommend plugging in the estimator $\widehat{\zeta}_n$ in place of $\zeta_0$ in practice. To ease the computational burden required to simulate the CP's, the finite-sample CP's of the robust CI's reported in Figures 5 and

S-8 are constructed using the true value $\zeta_0$, rather than the estimated value $\widehat{\zeta}_n$.[27] To see how much these robust CI's may differ from their counterparts constructed with $\widehat{\zeta}_n$, which is what one would use in practice, Table S-1 in Supplemental Appendix C compares their CP's in different identification scenarios in a small-scale simulation. The comparison suggests that the robust CI's obtained with $\zeta_0$ and those obtained with $\widehat{\zeta}_n$ are fairly close.[28]

## 8.2. Numerical Results for the Binary Choice Model

The binary choice model considered is

$$Y_i = 1(Y_i^* > 0) \text{ and } Y_i^* = \zeta_0 + \zeta_1 Z_i^* + \beta \cdot h(X_i, \pi) - U_i, \tag{8.2}$$

with $h(x, \pi) = (x^\pi - 1)/\pi$, $Z_i^* \sim N(0,1)$, $X_i = |X_i^*|$ with $X_i^* \sim N(3,1)$, $\text{Corr}(Z_i^*, X_i^*) = 0.5$, and $U_i \sim N(0,1)$. The true values of $\zeta_0$ and $\zeta_1$ are $-2$ and 2, respectively. The true parameter space for $\pi$ is $[1.5, 3.5]$ and the optimization space for $\pi$ is $[1, 4]$. The number of simulation repetitions is $20,000$.[29]

Figures 6-10 provide results analogous to those in Figures 1-5. Figures S-9 to S-16 in Supplemental Appendix C report results analogous to those in Figures S-1 to S-8.

The simulation results for the binary choice model are summarized as follows. First, the LS estimators and the $|t|$ and QLR statistics for $\beta$ and $\pi$ do not display normal or $\chi_1^2$ distributions under non-identification and weak identification. However, the asymptotic approximations developed here work very well in general, as indicated in Figures 6, 7, 9, and 10.[30]

[27] With a single sample, the computational burden is the same whether the true value $\zeta_0$ 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 $\pi_0$ and the single true value of $\zeta_0$ 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.

[28] The comparison is made based on a simulation with $1,000$ samples of size 500 to obtain the finite-sample CP's and $5,000$ simulation repetitions to determine the two LF critical values for each sample. The CI's considered are robust $t$ and QLR CI's for $\beta$. The estimator $\widehat{\zeta}_n$ employed is the null-imposed estimator. For CI's with nomial CP .950, the differences in finite sample CP's for $t$ CI's between using the true $\zeta$ and using $\widehat{\zeta}$ are .003 or less in 12 of the 13 cases and .005 in the other case. For the QLR CI's, differences are .004 or less in 9 of the 13 cases and .005, .008, .008, and .013 in the other four cases.

[29] For the binary choice model, 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 6, a grid of 0.5 for $b$ between 6 and 12, and a grid of 1 for $b$ between 12 and 30.

[30] The largest discrepancies between the asymptotic and finite-sample results occur when $\pi_0 = 2.0$

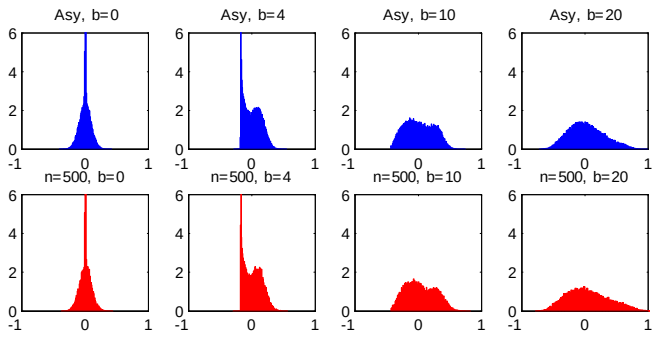


Figure 6. Asymptotic and Finite-Sample (n=500) Densities of the Estimator of $\beta$ in the Binary Choice Model when $\pi_0 = 1.5$.

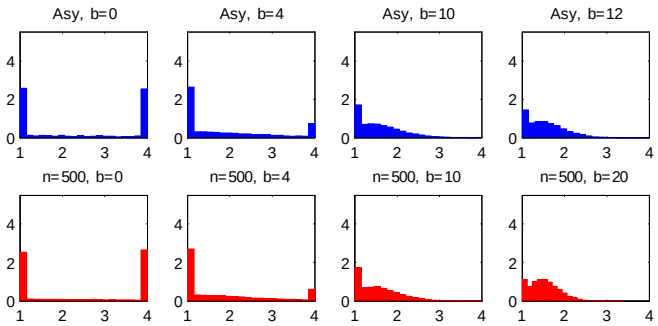


Figure 7. Asymptotic and Finite-Sample (n=500) Densities of the Estimator of $\pi$ in the Binary Choice Model when $\pi_0 = 1.5$.

Second, tests and CI's that employ the $|t|$ and QLR statistics for $\beta$ and the standard critical values have incorrect size, but the size distortion is much smaller for the QLR tests and CI's. For example, the standard $|t|$ and QLR CI's for $\beta$ have asymptotic CP's around 0.70 and 0.92, respectively, when $\pi_0 = 1.5$.[31] Tests and CI's that employ the QLR statistic for $\pi$ and the standard critical value have correct asymptotic size and those employ the $|t|$ statistic for $\pi$ only have small size distortions.

Third, the robust CI's have asymptotic CP's greater than or equal to 0.95 for all $b$. The finite-sample CP's are greater than or equal to 0.95 in all cases except for the robust $|t|$ CI for $\beta$, where the CP's are slightly below 0.95 for a small range of $b$ values and the lowest CP is around 0.93. The finite-sample under-coverage of the robust CI's is much smaller than that of the corresponding standard CI's.

and $b = 20$, see Figures S-9 and S-10, in which case the shape of the asymptotic approximation is good, but its scale is off.

[31] The standard QLR CI for $\beta$ only under-covers for $\beta$ very close to zero, which makes it difficult to detect in Figures 8(b) and 9(b).

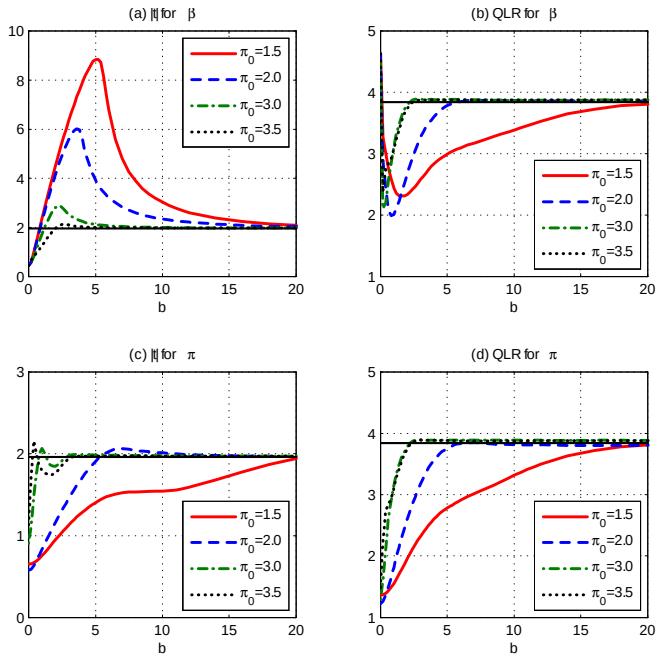


Figure 8. Asymptotic 0.95 Quantiles of the $|t|$ and QLR Statistics for Tests Concerning $\beta$ and $\pi$ in the Binary Choice Model.

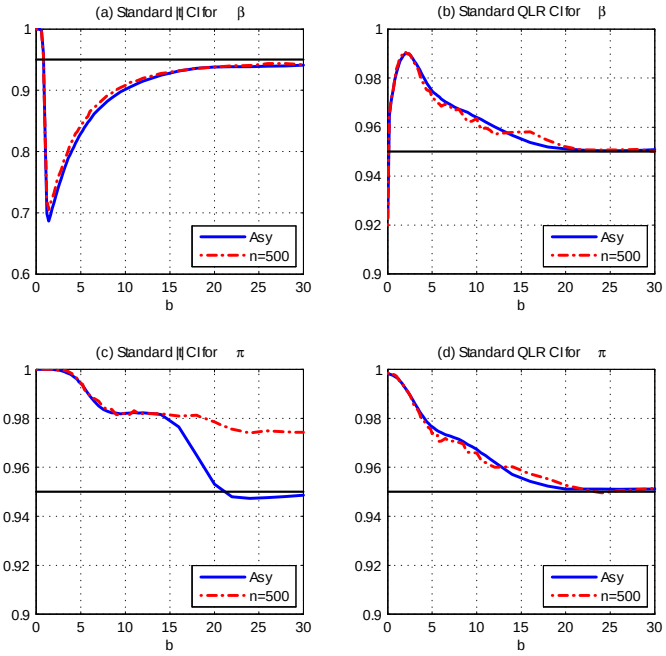


Figure 9. Coverage Probabilities of Standard $|t|$ and QLR CI's for $\beta$ and $\pi$ in the Binary Choice Model when $\pi_0 = 1.5$.

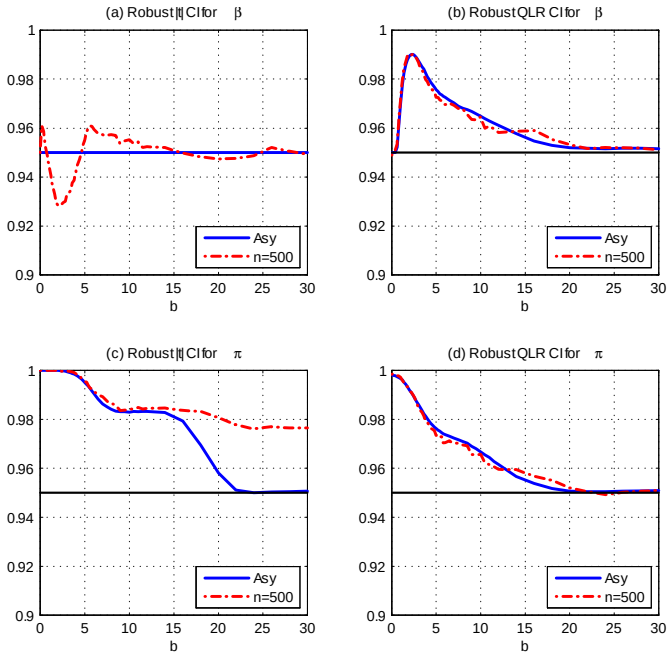


Figure 10. Coverage Probabilities of Robust $|t|$ and QLR CI's for $\beta$ and $\pi$ in the Binary Choice Model when $\pi_0 = 1.5$, $\kappa = 1.5$, $D = 1$, and $s(x) = \exp(-x/2)$.

# 9 Appendix

This Appendix provides the forms $\rho_\theta^\dagger(w,\theta)$, $\rho_{\theta\theta}^\dagger(w,\theta)$, and $\varepsilon(w,\theta)$ when $\rho(w,\theta)$ is as in (3.2) and for the STAR model.

## 9.1. Forms of $\boldsymbol{\rho_\theta^\dagger(\mathbf{w},\theta)}$, $\boldsymbol{\rho_{\theta\theta}^\dagger(\mathbf{w},\theta)}$, and $\boldsymbol{\varepsilon(\mathbf{w},\theta)}$

Here, we illustrate the forms of $\rho_\theta^\dagger(w,\theta)$, $\rho_{\theta\theta}^\dagger(w,\theta)$, and $\varepsilon(w,\theta)$ when $\rho(w,\theta)$ belongs to the class specified in (3.2) and show that Assumption S3(i) holds in this case. For simplicity, we assume $a(x,\beta)$ and $h(x,\pi)$ are both scalars and no parameter $\zeta$ appears. Let $\rho'(\cdot)$ and $\rho''(\cdot)$ abbreviate the first- and second-order derivatives of $\rho^*(w, a(x,\beta)h(x,\pi))$ wrt $a(x,\beta)h(x,\pi)$. Let $a_\beta(x,\beta)$, $a_{\beta\beta}(x,\beta)$, $h_\pi(x,\pi)$, $h_{\pi\pi}(x,\pi)$ denote the first- and second-order partial derivatives of $a(x,\beta)$ and $h(x,\pi)$ wrt $\beta$ and $\pi$. The first and second order

partial derivatives of $\rho(w,\theta)$ wrt to $\beta$ and $\pi$ are

$$\begin{aligned}
\rho_\beta(w,\theta) &= \rho'(\cdot)a_\beta(x,\beta)h(x,\pi),\ \rho_\pi(w,\theta) = \rho'(\cdot)a(x,\beta)h_\pi(x,\pi),\\
\rho_{\beta\beta}(w,\theta) &= \rho''(\cdot)a_\beta(x,\beta)a_\beta(x,\beta)'h^2(x,\pi) + \rho'(\cdot)a_{\beta\beta}(x,\beta)h(x,\pi),\\
\rho_{\beta\pi}(w,\theta) &= \rho''(\cdot)a(x,\beta)h(x,\pi)a_\beta(x,\beta)h_\pi(x,\pi)' + \rho'(\cdot)a_\beta(x,\beta)h_\pi(x,\pi)',\ \text{and}\\
\rho_{\pi\pi}(w,\theta) &= \rho''(\cdot)a^2(x,\beta)h_\pi(x,\pi)h_\pi(x,\pi)' + \rho'(\cdot)a(x,\beta)h_{\pi\pi}(x,\pi). \qquad (9.1)
\end{aligned}$$

In this case, we have

$$\begin{aligned}
\rho^\dagger_\theta(w,\theta) &= \rho'(\cdot)a^\dagger(x,\theta),\ \rho^\dagger_{\theta\theta}(w,\theta) = \rho''(\cdot)a^\dagger(x,\theta)a^\dagger(x,\theta)',\ \text{where}\\
a^\dagger(x,\theta) &= (a_\beta(x,\beta)'h(x,\pi), \frac{a(x,\beta)}{\iota(\beta)}h_\pi(x,\pi)')' \text{ and}\\
\varepsilon(w,\theta) &= \rho'(\cdot)\begin{bmatrix} a_{\beta\beta}(x,\beta)h(x,\pi) & a_\beta(x,\beta)h_\pi(x,\pi)' \\ h_\pi(x,\pi)a_\beta(x,\beta)' & \frac{a(x,\beta)}{\iota(\beta)}h_{\pi\pi}(x,\pi)\end{bmatrix}. \qquad (9.2)
\end{aligned}$$

Note that $\beta^{-1}a(x,\beta)$ is continuous at $\beta = 0$ in the scalar $\beta$ case. In particular, $\lim_{\beta\to 0}\beta^{-1}a(x,\beta) = a_\beta(x,0)$ by a mean-value expansion because $a(x,0) = 0$ and $a(x,\beta)$ is continuously differentiable in $\beta$. In the vector $\beta$ case, $\lim_{\beta\to 0,\beta/||\beta||\to\omega_0}||\beta||^{-1}a(x,\beta) = a_\beta(x,0)\omega_0$.

When $\varepsilon(w,\theta)$ takes the form in (9.2), Assumption S3* below implies Assumption S3(i). In Assumption S3*(i), $X_i$ is a sub-vector of $W_i$ that takes the place of $x$ in $w$.

**Assumption S3*.** (i) $X_i$ is a vector of weakly exogenous variables such that $E_{\gamma_0}(\rho'(W_i, a(X_i,\beta_0)h(X_i,\pi_0))|X_i) = 0$ a.s. $\forall\gamma_0\in\Gamma$.
(ii) $E_{\gamma_0}\sup_{||\beta||<\delta,\pi\in\Pi}|\rho''(W_i,a(X_i,\beta)h(X_i,\pi))|\cdot(||h(X_i,\pi)|| + ||h_\pi(X_i,\pi)||)\cdot(|h(X_i,\pi)| + ||h_\pi(X_i,\pi)|| + ||h_{\pi\pi}(X_i,\pi)||)\cdot\sup_{||\beta||<\delta}||a_\beta(X_i,\beta)||\cdot(||a_\beta(X_i,\beta)|| + ||a_{\beta\beta}(X_i,\beta)||) \le C$ for some $C<\infty$ and $\delta>0$ $\forall\gamma_0\in\Gamma$.

Several of the derivatives in Assumption S3*(ii) are constants in many examples, which makes the moment condition in Assumption S3*(ii) less restrictive than it may appear. For example, when $a(X_i,\beta) = \beta$, $a_\beta(X_i,\beta) = 1$ and $a_{\beta\beta}(X_i,\beta) = 0$.

**Lemma 9.1.** *Suppose $\rho(w,\theta)$ belongs to the class in* (3.2)*, where $a(x,\beta)\in R$ and $h(x,\pi)\in R$ are twice differentiable wrt $\beta$ and $\pi$, respectively, and no parameter $\zeta$ appears. Then, $\varepsilon(w,\theta)$ takes the form in* (9.2) *and Assumption* S3(i) *is implied by Assumption* S3* *in both the scalar and vector $\beta$ cases.*

**Comment.** When $\rho(w,\theta)$ belongs to the class in (3.2) and a parameter $\zeta$ appears, the form of $\varepsilon(w,\theta)$ is the same as in (9.2) but with zeros in the rows and columns that correspond to $\zeta$. In this case, Assumption S3(i) is still implied by Assumption S3$^*$ provided $\rho'(\cdot)$ and $\rho''(\cdot)$ in Assumption S3$^*$ are adjusted to include $\zeta$, evaluated at $\zeta_0$. See Supplemental Appendix B for details.

## 9.2. STAR Model Quantities

For the STAR model, the criterion function in (7.4) is of the form $Q_n(\theta) = n^{-1}\sum_{t=1}^{n} \rho(W_t,\theta)$ with $\rho(W_t,\theta) = (1/2)U_t^2(\theta)$. The first- and second-order partial derivatives of $\rho(W_t,\theta)$ wrt $\psi$ and $\theta$ are

$$\begin{aligned}
\rho_{\psi}(W_t,\theta) &= -U_t(\theta)d_{\psi,t}(\pi),\ \rho_{\theta}(W_t,\theta) = -U_t(\theta)d_{\theta,t}(\theta),\\
\rho_{\psi\psi}(W_t,\theta) &= d_{\psi,t}(\pi)d_{\psi,t}(\pi)',\\
\rho_{\theta\theta}(W_t,\theta) &= d_{\theta,t}(\theta)d_{\theta,t}(\theta)' - U_t(\theta)D_t(\theta),\ \text{where}\\
d_{\psi,t}(\pi) &= (X_t'm(Z_t,\pi), X_t')',\ d_{\theta,t}(\theta) = (X_t'm(Z_t,\pi), X_t', \beta'X_t m_{\pi}(Z_t,\pi)')' \text{ and}\\
D_t(\theta) &= \begin{bmatrix} 0_{d_\beta\times d_\beta} & 0_{d_\beta\times d_\zeta} & X_t m_{\pi}(Z_t,\pi)' \\ 0_{d_\zeta\times d_\beta} & 0_{d_\zeta\times d_\zeta} & 0_{d_\zeta\times d_\pi} \\ m_{\pi}(Z_t,\pi)X_t' & 0_{d_\pi\times d_\zeta} & \beta'X_t\cdot m_{\pi\pi}(Z_t,\pi) \end{bmatrix}. \qquad (9.3)
\end{aligned}$$

Define

$$d_t(\pi,\omega) = (X_t'm(Z_t,\pi), X_t', \omega'X_t m_{\pi}(Z_t,\pi)')'. \qquad (9.4)$$

The rescaled partial derivatives in (3.4) take the form

$$\begin{aligned}
\rho_{\theta}^{\dagger}(W_t,\theta^+) &= -U_t(\theta^+)d_t(\pi,\omega),\ \rho_{\theta\theta}^{\dagger}(W_t,\theta^+) = d_t(\pi,\omega)d_t(\pi,\omega)',\ \text{and}\\
\varepsilon(W_t,\theta^+) &= -U_t(\theta^+)\begin{bmatrix} 0_{d_\beta\times d_\beta} & 0_{d_\beta\times d_\zeta} & X_t m_{\pi}(Z_t,\pi)' \\ 0_{d_\zeta\times d_\beta} & 0_{d_\zeta\times d_\zeta} & 0_{d_\zeta\times d_\pi} \\ m_{\pi}(Z_t,\pi)X_t' & 0_{d_\pi\times d_\zeta} & \omega'X_t\cdot m_{\pi\pi}(Z_t,\pi) \end{bmatrix}, \text{ where}\\
U_t(\theta^+) &= Y_t - X_t'\zeta - X_t'\omega||\beta||\cdot m(Z_t,\pi). \qquad (9.5)
\end{aligned}$$

Let

$$\begin{aligned}
V^{\dagger}(\theta_0^+,\theta_0^+;\gamma_0) &= V(\gamma_0) = E_{\gamma_0}U_t^2 d_t(\pi_0,\omega_0)d_t(\pi_0,\omega_0)' \text{ and}\\
J(\gamma_0) &= E_{\gamma_0}d_t(\pi_0,\omega_0)d_t(\pi_0,\omega_0). \qquad (9.6)
\end{aligned}$$

The quantities in (9.3), (9.5), and (9.6) appear in Assumptions S1-S4. The matrices $J(\gamma_0)$ and $V(\gamma_0)$ appear in Theorem 4.2.

$t$ tests and CI's employ estimators of $J(\gamma_0)$ and $V(\gamma_0)$. We estimate these matrices by

$$\widehat{J}_n = \widehat{J}_n(\widehat{\theta}_n^+) \text{ and } \widehat{V}_n = \widehat{V}_n(\widehat{\theta}_n^+), \text{ where} \tag{9.7}$$
$$\widehat{J}_n(\theta^+) = n^{-1}\sum_{i=1}^{n} d_t(\pi,\omega)d_t(\pi,\omega)' \text{ and } \widehat{V}_n(\theta^+) = n^{-1}\sum_{i=1}^{n} U_t^2(\theta^+)d_t(\pi,\omega)d_t(\pi,\omega)'.$$

These variance matrix estimators also are used to construct the identification-category-selection statistic $A_n$.

# References


Andrews, D. W. K., 1988. Laws of large numbers for dependent non-identically distributed random variables. Econometric Theory 4, 458-467.

Andrews, D. W. K., Cheng, X., 2011a. GMM estimation and uniform subvector inference with possible identification failure. Cowles Foundation Discussion Paper No. 1828, Yale University.

Andrews, D. W. K., Cheng, X., 2011b. Supplemental appendices for "Maximum likelihood estimation and uniform inference with sporadic identification failure." Addendum to Cowles Foundation Discussion Paper No. 1824R, Yale University.

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

Andrews, D. W. K., Cheng, X., 2012b. Supplemental material for "Estimation and inference with weak, semi-strong, and strong identification." Available on the Econometric Society website.

Andrews, D. W. K., Cheng, X., Guggenberger, P., 2009. Generic results for establishing the asymptotic size of confidence intervals and tests. Cowles Foundation Discussion Paper No. 1813, Yale University.

Andrews, D. W. K., Guggenberger, P., 2010. Asymptotic size and a problem with subsampling and with the $m$ out of $n$ bootstrap. Econometric Theory 26, 426-468.

Andrews, I., Mikusheva, A., 2011. Maximum likelihood inference in weakly identified DSGE models. Unpublished manuscript, Department of Economics, MIT.

Andrews, I., Mikusheva, A., 2012. A geometric approach to weakly identified econometric models. Unpublished manuscript, Department of Economics, MIT.

Antoine, B., Renault, E., 2009. Efficient GMM with nearly-weak instruments. Econometrics Journal 12, S135-S171.

Antoine, B., Renault, E., 2010. Efficient inference with poor instruments, a general framework. In: Giles, D., Ullah, A. (Eds.), Handbook of Empirical Economics and Finance. Taylor and Francis, Oxford, UK.

Bhattacharya, R., Lee, C., 1995. On geometric ergodicity of nonlinear autoregressive models. Statistics and Probability Letters 22, 311-315.

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. Unpublished working paper, Department of Economics, Yale University.

Choi, I., Phillips, P. C. B., 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.

de Jong, R. M., 1997. Central limit theorems for dependent heterogeneous random variables. Econometric Theory 13, 353-367.

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

Hansen, B. E., 1996. Stochastic equicontinuity for unbounded dependent heterogeneous arrays. Econometric Theory 12, 347-359.

Hansen, B. E., 2000. Sample splitting and threshold estimation. Econometrica 68, 575-603.

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.

Lundbergh, S., Teräsvirta, T. 2006. A time series model for an exchange rate in a target zone with applications. Journal of Econometrics 131, 579-609.

Luukkonen, R., Saikkonen, P., Teräsvirta, T., 1988. Testing linearity against smooth transition autoregressive models. Biometrika 75, 491-499.

Ma, J., Nelson, C. R., 2008. Valid inference for a class of models where standard inference performs poorly; including nonlinear regression, ARMA, GARCH, and unobserved components. Unpublished manuscript, Department of Economics, U. of Washington.

Moreira, M. J., 2003. A conditional likelihood ratio test for structural models. Econometrica 71, 1027-1048.

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

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

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. Unpublished working paper, Boston University.

Sargan, J. D., 1983. Identification and lack of identification. Econometrica, 51 1605-1633.

Shi, X., Phillips, P. C. B., 2012. Nonlinear cointegrating regression with weak identification. Econometric Theory, 28 1-39.

Staiger, D., Stock, J. H., 1997. Instrumental variables regression with weak instruments. Econometrica 65, 557-586.

Stock, J. H., Wright, J. H., 2000. GMM with weak instruments. Econometrica 68, 1055-1096.

Teräsvirta, T., 1994. Specification, estimation, and evaluation of smooth transition autoregressive models. Journal of the American Statistical Association 89, 208-218.

Teräsvirta, T., Anderson, H. M., 1992. Characterizing nonlinearities in business cycles using smooth transition autoregressive models. Journal of Applied Econometrics 7, S119-S136.

Tripathi, G. T., 1999. A matrix extension of the Cauchy-Schwarz inequality. Economics Letters 63, 1-3.