MAXIMUM LIKELIHOOD ESTIMATION AND UNIFORM INFERENCE WITH SPORADIC IDENTIFICATION FAILURE

By

Donald W. K. Andrews and Xu Cheng

October 2011

COWLES FOUNDATION DISCUSSION PAPER NO. 1824

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: October 17, 2011

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 θ. 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 (2007a) (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 θ of interest is of the form $\theta = (\beta, \zeta, \pi)$, where π is identified if and only if $\beta \neq 0$, ζ is not related to the identification of π, and $\psi = (\beta, \zeta)$ is always identified. 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 β, ζ, and π 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): $\beta_n = 0 \; \forall n \geq 1$, π is unidentified; Category I(b): $\beta_n \neq 0$ and $n^{1/2}\beta_n \to b \in R^{d_\beta}$, π is weakly identified; Category II: $\beta_n \to 0$ and $n^{1/2}||\beta_n|| \to \infty$, π is semi-strongly identified; and Category III: $\beta_n \to \beta_0 \neq 0$, π is strongly identified.

For Category I sequences, we obtain the following results: the estimator of π 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 χ^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 π 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 β is near the value 0 that yields lack of identification of π, 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.

The numerical results for the STAR and nonlinear binary choice models are summarized as follows. The asymptotic distributions of the estimators of β and π 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 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 β have

---

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

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 π, 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 β have incorrect asymptotic sizes: .68 versus .92, respectively, for nominal .95 CI's. However, the standard t and QLR CI's for π 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.

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 θ. 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.[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 (2007c) (AC1-SM), and Andrews and Cheng (2008) (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 of weak identification, whereas the present paper only considers a single source. However, the present paper applies to a much broader range of models.

Tests of $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ are tests in which a nuisance parameter π only appears under the alternative. Such tests have been considered in the literature

---

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

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 (β, ζ, π) 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 β, which determines the strength of identification of π, 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), Moreira (2003), and Kleibergen and Mavroeidis (2009).

Andrews and Mikusheva (2010) 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.

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 (2011) 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 asymptotic 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. Appendix A provides proofs of the results given in the paper. Appendix B provides some mis-

cellaneous results. Three Supplemental Appendices to this paper are given in Andrews and Cheng (2007b). 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 $n \rightarrow \infty$." 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_β-vector of zeros, where d_β is the dimension of a parameter β. 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 Π.[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

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

where $\{W_i : i \leq n\}$ are the observations and $\rho(w, \theta)$ is a known function that is twice continuously differentiable in θ. 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.

The paper considers the case where θ 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 θ. To model this identification problem, θ is parti-

---

[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 Π. See the Outline of the Supplemental Appendix of AC1 for more details.

tioned 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 π. These conditions allow for a wide range of cases, including cases in which reparametrization 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 π 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 ϕ, 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,

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

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

6

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 π, and π is not identified. $\square$

The true distribution of the observations $\{W_i : i \leq n\}$ is denoted F_γ for some parameter $\gamma \in \Gamma$. We let P_γ and E_γ denote probability and expectation under F_γ. The parameter space Γ for the true parameter, 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 Θ^* is a compact subset of R^{d_θ} and $\Phi^*(\theta) \subset \Phi^* \; \forall \theta \in \Theta^*$ for some compact metric space Φ^* with a metric that induces weak convergence of the bivariate distributions (W_i, W_{i+m}) for all $i, m \geq 1$.[4] In unconditional likelihood scenarios, no parameter ϕ appears. In conditional likelihood scenarios, with conditioning variables $\{X_i : i \geq 1\}$, ϕ indexes the distribution of $\{X_i : i \geq 1\}$. In nonlinear regression models estimated by least squares, θ indexes the regression functions and possibly a finite-dimensional feature of the distribution of the errors, such as its variance, and ϕ indexes the remaining characteristics of the distribution of the errors, which may be infinite dimensional.

By definition, the estimator $\widehat{\theta}_n$ (approximately) minimizes $Q_n(\theta)$ over an "optimization parameter space" Θ:[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 Θ includes the true parameter space Θ^* (see Assumption 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 Θ^*. The focus of this paper is not on boundary effects.

Without loss of generality (wlog), the optimization parameter space Θ can be written

---

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

[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 π 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.

as

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

We allow $\Psi(\pi)$ to depend on π and, hence, Θ need not be a product space between ψ and π. 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 θ 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, $\mathcal{Z}^*$ $(\subset R^{d_\varsigma})$ is compact, and Π^* $(\subset R^{d_\pi})$ is compact.

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

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

$b_1 > b_1^*$, $b_2 > b_2^*$, $\mathcal{Z}$ $(\subset R^{d_\varsigma})$ is compact, Π $(\subset R^{d_\pi})$ 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 θ, 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

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

8

where $P_\gamma(\cdot)$ denotes probability when γ 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:

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

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

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

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 γ 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:

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

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

9

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 θ. 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_{γ_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 $\delta > 0$.

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.

---

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

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 $\gamma_0 \in \Gamma$. $\square$

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 θ on an open set containing Θ^* $\forall w \in \mathcal{W}$.
(ii) $\rho(w, \theta)$ does not depend on π 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 ψ_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 θ_0.
(v) $\Psi(\pi)$ is compact $\forall \pi \in \Pi$, and Π and Θ 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

$$\rho(W_i, \theta) = Y_i \log L(g_i(\theta)) + (1 - Y_i) \log(1 - L(g_i(\theta))) \tag{3.16}$$

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

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

$$\rho(w, \theta) = \rho^*(w, a(x, \beta)h(x, \pi), \zeta), \text{ where } a(x, 0) = 0, \ \forall w \in \mathcal{W}, \tag{3.17}$$

11

x is a sub-vector of w, and $a(x, \beta)$ and $h(x, \pi)$ are known functions. In (3.17), $\rho(w, \theta)$ does not depend on π 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 β is the coefficient of the nonlinear regressor. Example (ii) demonstrates that $a(x, \beta)$ can be nonlinear in β provided $a(x, \beta) = 0$ at $\beta = 0$. Example (iii) covers the weak IV example and the case in which β enters the model through a single index. The form in (3.17) 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 π. When $\Psi(\pi)$ depends on π, the boundary of $\Psi(\pi)$ is often a continuous linear function of π, 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 θ, 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 ψ.

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

$$B(\beta) = \begin{bmatrix} I_{d_\psi} & 0_{d_\psi \times d_\pi} \\ 0_{d_\pi \times d_\psi} & \iota(\beta) I_{d_\pi} \end{bmatrix} \in R^{d_\theta \times d_\theta}, \text{ where } \iota(\beta) = \begin{cases} \beta & \text{if } \beta \text{ is a scalar} \\ ||\beta|| & \text{if } \beta \text{ is a vector} \end{cases}. \quad (3.18)$$

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

For $\beta \neq 0$, let

$$B^{-1}(\beta)\rho_\theta(w, \theta) = \rho_\theta^\dagger(w, \theta) \text{ and}$$
$$B^{-1}(\beta)\rho_{\theta\theta}(w, \theta)B^{-1}(\beta) = \rho_{\theta\theta}^\dagger(w, \theta) + \iota^{-1}(\beta)\varepsilon(w, \theta), \quad (3.19)$$

where $\rho_{\theta\theta}^\dagger(w, \theta)$ is symmetric and $\rho_\theta^\dagger(w, \theta)$, $\rho_{\theta\theta}^\dagger(w, \theta)$, and $\varepsilon(w, \theta)$ satisfy Assumption S3 below. The re-scaling matrix $B^{-1}(\beta)$ in (3.19) 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 β is close to 0. In contrast, the re-scaled quantity $\rho_\theta^\dagger(W_i, \theta)$ has a covariance matrix that is not close to being singular even when β is close to 0. Similarly, $E_{\gamma_0}\rho_{\theta\theta}(W_i, \theta)$ is singular when $\beta = 0$ and close to singular when β is close to 0. Re-scaling of $\rho_{\theta\theta}(W_i, \theta)$ yields a quantity $\rho_{\theta\theta}^\dagger(W_i, \theta)$ whose expectation is not close to singular even when β is close to 0 plus another term $\varepsilon(W_i, \theta)$ that is asymptotically negligible.

Below we illustrate the form of $\rho_\theta^\dagger(w, \theta)$, $\rho_{\theta\theta}^\dagger(w, \theta)$, and $\varepsilon(w, \theta)$ in Example 1 and for $\rho(w, \theta)$ functions as in (3.17), see Section 3.1.4.

Next, define

$$V^\dagger(\theta_1, \theta_2; \gamma_0) = \sum_{m=-\infty}^{\infty} Cov_{\gamma_0}(\rho_\theta^\dagger(W_i, \theta_1), \rho_\theta^\dagger(W_{i+m}, \theta_2)), \qquad (3.20)$$

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 β is a scalar or vector. We state Assumption S3 for the scalar β case first because it is simpler.

Assumption S3 (scalar β). (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 $M_1(w) : \mathcal{W} \rightarrow R_+$ and $M_2(w) : \mathcal{W} \rightarrow R_+$, $||\rho_{\psi\psi}(w, \theta_1) - \rho_{\psi\psi}(w, \theta_2)|| + ||\rho_{\theta\theta}^\dagger(w, \theta_1) - \rho_{\theta\theta}^\dagger(w, \theta_2)|| \leq M_1(w)\delta$ and $||\rho_\theta^\dagger(w, \theta_1) - \rho_\theta^\dagger(w, \theta_2)||$ $+ ||\varepsilon(w, \theta_1) - \varepsilon(w, \theta_2)|| \leq M_2(w)\delta$, $\forall\theta_1, \theta_2 \in \Theta$ with $||\theta_1 - \theta_2|| \leq \delta$, $\forall w \in \mathcal{W}$.
(iii) $E_{\gamma_0} \sup_{\theta \in \Theta}\{|\rho(W_i, \theta)|^{1+\delta} + ||\rho_{\psi\psi}(W_i, \theta)||^{1+\delta} + ||\rho_{\theta\theta}^\dagger(W_i, \theta)||^{1+\delta} + M_1(W_i) + ||\rho_\theta^\dagger(W_i, \theta)||^q$ $+ ||\varepsilon(W_i, \theta)||^q + M_2(W_i)^q\} \leq C$ 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 $E_{\gamma_0}\rho_{\theta\theta}^\dagger(W_i, \theta_0)$ 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 qth moments in order to establish the stochastic equicontinuity of empirical processes based on $\rho_\theta^\dagger(W_i, \theta)$ and $\varepsilon(W_i, \theta)$ using Lemma 9.4 in 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 π 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 π is approximately proportional to $||\beta_0||$ when $||\beta_0||$ is close to 0. Because parts of $B^{-1}(\beta)$ diverge as β converges to 0, the moment conditions for $\rho_\theta^\dagger(W_i, \theta)$ and $\rho_{\theta\theta}^\dagger(W_i, \theta)$ 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 π is small when β 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), $E_{\gamma_0}\rho_{\theta\theta}^\dagger(W_i, \theta_0)$ and $V^\dagger(\theta_0, \theta_0; \gamma_0)$ typically are positive definite due to the re-scaling in (3.19).

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, $\rho_\theta^\dagger(W_i, \theta)$, $\rho_{\theta\theta}^\dagger(W_i, \theta)$, 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

$$d_{\psi,i}(\pi) = (h(X_i, \pi), Z_i')', \ d_i(\pi) = (h(X_i, \pi), Z_i', h_\pi(X_i, \pi)')', \text{ and}$$

$$D_i(\theta) = \begin{bmatrix} 0 & 0_{1 \times d_\zeta} & h_\pi(X_i, \pi)' \\ 0_{d_\zeta \times 1} & 0_{d_\zeta \times d_\zeta} & 0_{d_\zeta \times d_\pi} \\ h_\pi(X_i, \pi) & 0_{d_\pi \times d_\zeta} & h_{\pi\pi}(X_i, \pi)\beta \end{bmatrix}. \tag{3.21}$$

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

$$\rho_\psi(W_i, \theta) = w_{1,i}(\theta)(Y_i - L_i(\theta))d_{\psi,i}(\pi),$$

$$\rho_\theta(W_i, \theta) = w_{1,i}(\theta)(Y_i - L_i(\theta))B(\beta)d_i(\pi),$$

$$\rho_{\psi\psi}(W_i, \theta) = [w_{1,i}^2(\theta)(Y_i - L_i(\theta))^2 + w_{2,i}(\theta)(Y_i - L_i(\theta))]d_{\psi,i}(\pi)d_{\psi,i}(\pi)',$$

$$\rho_{\theta\theta}(W_i, \theta) = [w_{1,i}^2(\theta)(Y_i - L_i(\theta))^2 + w_{2,i}(\theta)(Y_i - L_i(\theta))]B(\beta)d_i(\pi)d_i(\pi)'B(\beta)$$

$$+ w_{1,i}(\theta)(Y_i - L_i(\theta))D_i(\theta), \text{ where}$$

$$w_{1,i}(\theta) = \frac{-L_i'(\theta)}{L_i(\theta)(1 - L_i(\theta))} \text{ and } w_{2,i}(\theta) = \frac{-L_i''(\theta)}{L_i(\theta)(1 - L_i(\theta))}. \tag{3.22}$$

See Section 13.10 for the calculation of these derivatives.

14

The re-scaled partial derivatives in (3.19) take the form

$$\rho_\theta^\dagger(W_i, \theta) = w_{1,i}(\theta)(Y_i - L_i(\theta))d_i(\pi),$$

$$\rho_{\theta\theta}^\dagger(W_i, \theta) = [w_{1,i}^2(\theta)(Y_i - L_i(\theta))^2 + w_{2,i}(\theta)(Y_i - L_i(\theta))]d_i(\pi)d_i(\pi)', \text{ and}$$

$$\varepsilon(w, \theta) = w_{1,i}(\theta)(Y_i - L_i(\theta)) \begin{bmatrix} 0 & 0_{1 \times d_\zeta} & h_\pi(X_i, \pi)' \\ 0_{d_\zeta \times 1} & 0_{d_\zeta \times d_\zeta} & 0_{d_\zeta \times d_\pi} \\ h_\pi(X_i, \pi) & 0_{d_\pi \times d_\zeta} & h_{\pi\pi}(X_i, \pi) \end{bmatrix}. \quad (3.23)$$

To verify Assumption S3(i), note that $E_{\gamma_0}(Y_i|X_i, Z_i) = P_{\gamma_0}(Y_i = 1|X_i, Z_i) = L_i(\theta_0)$ by (2.4). Hence, $E_{\gamma_0}(Y_i - L_i(\theta_0)|X_i, Z_i) = 0$ implies $E_{\gamma_0}\varepsilon(W_i, \theta_0) = 0$ by the law of iterated expectations (LIE).

Let $\overline{L}_i' = \sup_{\theta \in \Theta} |L_i'(\theta)|$ and $\overline{L}_i'' = \sup_{\theta \in \Theta} |L_i''(\theta)|$.

A mean-value expansion of $L_i(\psi_0, \pi)$ wrt π around π_0 yields

$$L_i(\psi_0, \pi) - L_i(\theta_0) = L_i'(\psi_0, \widetilde{\pi})\frac{\partial g_i(\psi_0, \widetilde{\pi})}{\partial \pi'}(\pi - \pi_0)$$

$$= L_i'(\psi_0, \widetilde{\pi})h_\pi(X_i, \psi_0, \widetilde{\pi})'\beta_0(\pi - \pi_0), \quad (3.24)$$

where $\widetilde{\pi}$ is between π and π_0. To verify the second part of Assumption S3(i), we have

$$||E_{\gamma_0}(w_{1,i}(\psi_0, \pi)[Y_i - L_i(\psi_0, \pi)]h_\pi(X_i, \pi))||$$

$$= ||E_{\gamma_0}(w_{1,i}(\psi_0, \pi)[L_i(\theta_0) - L_i(\psi_0, \pi)]h_\pi(X_i, \pi))||$$

$$\leq |\beta_0| \cdot ||\pi - \pi_0||E_{\gamma_0}(\overline{w}_{1,i}\overline{L}_i'\overline{h}_{\pi,i}^2) \leq C|\beta_0| \cdot ||\pi - \pi_0||, \quad (3.25)$$

for some $C < \infty$, where the equality holds by LIE, the first inequality holds using (3.24), and the second inequality holds by the Cauchy-Schwarz inequality and the moment conditions in (3.32).

Similarly, we have

$$||E_{\gamma_0}(w_{1,i}(\psi_0, \pi)[Y_i - L_i(\psi_0, \pi)]h_{\pi\pi}(X_i, \pi))|| \leq C|\beta_0| \cdot ||\pi - \pi_0|| \quad (3.26)$$

for some $C < \infty$. By (3.23), (3.25), and (3.26), the second part of Assumption S3(i) holds.

The rest of Assumption S3 is verified in Supplemental Appendix D for this example. $\square$

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

This new parameterization is needed when β 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 (3.19) and $\beta/||\beta||$ is not continuous in β for $\theta \in \Theta$. In consequence, the Lipschitz conditions in Assumptions S3(ii) and S3(iii) (scalar β) can not be verified when β is a vector. The new parameterization treats $||\beta||$ and $\omega = \beta/||\beta||$ as separate variables. In Assumption S3 (vector β) below, some Lipschitz conditions are specified in terms of $\theta^+ = (||\beta||, \omega, \zeta, \pi)$.

In Assumption S3 (vector β), both the original parameterization with θ and the alternative parameterization with θ^+ 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 θ^+.

Assumption S3 (vector β). (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 $M_2(w) : \mathcal{W} \to R_+$, $||\rho_{\psi\psi}(w, \theta_1) - \rho_{\psi\psi}(w, \theta_2)|| + ||\rho_{\theta\theta}^\dagger(w, \theta_1^+) - \rho_{\theta\theta}^\dagger(w, \theta_2^+)|| \leq M_1(w)\delta$ and $||\rho_\psi(w, \theta_1) - \rho_\psi(w, \theta_2)|| + ||\rho_\theta^\dagger(w, \theta_1^+) - \rho_\theta^\dagger(w, \theta_2^+)|| + ||\varepsilon(w, \theta_1^+) - \varepsilon(w, \theta_2^+)|| \leq M_2(w)\delta$, $\forall \theta_1, \theta_2 \in \Theta$ 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 β) hold with the definitions of $M_1(w)$ and $M_2(w)$ replaced by those given above.

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

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

Now, 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 (3.17) 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 ζ 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 β and π.

16

The first and second order partial derivatives of $\rho(w, \theta)$ wrt to β and π are

$$\rho_\beta(w, \theta) = \rho'(\cdot)a_\beta(x, \beta)h(x, \pi), \ \ \rho_\pi(w, \theta) = \rho'(\cdot)a(x, \beta)h_\pi(x, \pi),$$

$$\rho_{\beta\beta}(w, \theta) = \rho''(\cdot)a_\beta(x, \beta)a_\beta(x, \beta)'h^2(x, \pi) + \rho'(\cdot)a_{\beta\beta}(x, \beta)h(x, \pi),$$

$$\rho_{\beta\pi}(w, \theta) = \rho''(\cdot)a(x, \beta)h(x, \pi)a_\beta(x, \beta)h_\pi(x, \pi)' + \rho'(\cdot)a_\beta(x, \beta)h_\pi(x, \pi)', \ \ \text{and}$$

$$\rho_{\pi\pi}(w, \theta) = \rho''(\cdot)a^2(x, \beta)h_\pi(x, \pi)h_\pi(x, \pi)' + \rho'(\cdot)a(x, \beta)h_{\pi\pi}(x, \pi). \tag{3.27}$$

In this case, we have

$$\rho_\theta^\dagger(w, \theta) = \rho'(\cdot)a^\dagger(x, \theta), \ \ \rho_{\theta\theta}^\dagger(w, \theta) = \rho''(\cdot)a^\dagger(x, \theta)a^\dagger(x, \theta)', \ \ \text{where}$$

$$a^\dagger(x, \theta) = \left(a_\beta(x, \beta)'h(x, \pi), \frac{a(x, \beta)}{\iota(\beta)}h_\pi(x, \pi)'\right)' \ \text{and}$$

$$\varepsilon(w, \theta) = \rho'(\cdot)\begin{bmatrix} a_{\beta\beta}(x, \beta)h(x, \pi) & a_\beta(x, \beta)h_\pi(x, \pi)' \\ h_\pi(x, \pi)a_\beta(x, \beta)' & \frac{a(x, \beta)}{\iota(\beta)}h_{\pi\pi}(x, \pi) \end{bmatrix}. \tag{3.28}$$

Note that $\beta^{-1}a(x, \beta)$ is continuous at $\beta = 0$ in the scalar β 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 β. In the vector β 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 (3.28), 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) $E_{\gamma_0} \sup_{\|\beta\| < \delta, \pi \in \Pi} |\rho''(W_i, a(X_i, \beta)h(X_i, \pi))| \cdot (\|h(X_i, \pi)\| + \|h_\pi(X_i, \pi)\|) \cdot (|h(X_i, \pi)| + \|h_\pi(X_i, \pi)\| + \|h_{\pi\pi}(X_i, \pi)\|) \cdot \sup_{\|\beta\| < \delta} \|a_\beta(X_i, \beta)\| \cdot (\|a_\beta(X_i, \beta)\| + \|a_{\beta\beta}(X_i, \beta)\|) \leq C$ 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 3.1. *Suppose $\rho(w, \theta)$ belongs to the class in (3.17), where $a(x, \beta) \in R$ and $h(x, \pi) \in R$ are twice differentiable wrt β and π, respectively, and no parameter ζ appears. Then, $\varepsilon(w, \theta)$ takes the form in (3.28) and Assumption S3(i) is implied by Assumption S3* in both the scalar and vector β cases.*

17

Comment. When $\rho(w, \theta)$ belongs to the class in (3.17) and a parameter ζ appears, the form of $\varepsilon(w, \theta)$ is the same as in (3.28) but with zeros in the rows and columns that correspond to ζ. 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 ζ, evaluated at ζ_0. See Appendix B for details.

3.1.5. Assumption S4

Next, we state an assumption that controls how the mean $E_{\gamma_0} \rho_{\psi,i}(\theta)$ changes as the true β_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 β value, β_0, to be

$$K(\theta; \gamma_0) = \frac{\partial}{\partial \beta_0'} E_{\gamma_0} \rho_\psi(W_i, \theta). \tag{3.29}$$

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 (θ, γ^*) at $(\theta, \gamma^*) = ((\psi_0, \pi), \gamma_0)$ uniformly over $\pi \in \Pi$ $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$, where ψ_0 is a sub-vector of γ_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 holds with

$$K(\pi; \gamma_0) = K(\psi_0, \pi; \gamma_0) = E_{\gamma_0} \frac{L_i'^2(\theta_0)}{L_i(\theta_0)(1 - L_i(\theta_0))} h(X_i, \pi_0) d_{\psi,i}(\pi) \tag{3.30}$$

for $\gamma_0 = (\theta_0, \pi_0)$. □

---

[7] The constant $\delta > 0$ is as in Assumption B2(iii) stated below. The set Γ_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.

3.2. Parameter Space Assumptions

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

Define $\Theta_\delta^* = \{\theta \in \Theta^* : ||\beta|| < \delta\}$, where Θ^* is the true parameter space for θ, see (2.6). The optimization parameter space Θ 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 Π as in (2.8).
(iii) Π is compact.

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

The true parameter space Γ satisfies:

Assumption B2. (i) Γ 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 Γ is not empty and that there are elements γ of Γ whose β 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 Γ is compatible with the existence of partial derivatives of certain expectations wrt the true parameter β around $\beta = 0$, which arise in (3.29) and Assumption S4.

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

$$\Gamma = \{\gamma = (\theta, \phi) : \theta \in \Theta^*, \phi \in \Phi^*(\theta)\}, \tag{3.31}$$

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 ϕ is

$$
\Phi^*(\theta_0) = \{\phi_0 \in \Phi^* : \ E_{\gamma_0}(\overline{h}_i^{4q} + \overline{h}_{\pi,i}^{4q} + \overline{h}_{\pi\pi,i}^{4q} + ||Z_i||^{4q} + \overline{w}_{1,i}^{4q} + \overline{w}_{2,i}^{2+\delta}) \leq C
$$

$$
||w_{1,i}(\theta_1) - w_{1,i}(\theta_2)|| \leq M_1(W_i)||\pi_1 - \pi_2||, \ ||w_{2,i}(\theta_1) - w_{2,i}(\theta_2)||
$$

$$
\leq M_2(W_i)||\pi_1 - \pi_2||, \ ||h_{\pi\pi}(X_i, \pi_1) - h_{\pi\pi}(X_i, \pi_2)|| \leq M_h(W_i)||\pi_1 - \pi_2||,
$$

$$
\forall \pi_1, \pi_2 \in \Pi \text{ for some functions } M_1(W_i), \ M_2(W_i), \ M_h(W_i),
$$

$$
E_{\gamma_0}(M_1(W_i)^{4q/3} + M_2(W_i)^{4/3} + M_h(W_i)^{4q/3}) \leq C,
$$

$$
E_{\gamma_0} \sup_{\theta \in \Theta} \left(|\log L_i(\theta)|^{1+\delta} + |\log(1 - L_i(\theta))|^{1+\delta}\right) \leq C,
$$

$$
P_{\gamma_0}(a'(h(X_i, \pi_1), h(X_i, \pi_2), Z_i) = 0) < 1, \ \forall \pi_1, \pi_2 \in \Pi \text{ with } \pi_1 \neq \pi_2, \ \forall a \in R^{d_\zeta+2}
$$

$$
\text{with } a \neq 0, \ E_{\gamma_0} d_i(\pi) d_i(\pi)' \text{ is positive definite } \forall \pi \in \Pi\} \tag{3.32}
$$

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

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 ψ out of θ. Define

$$
\begin{aligned}
\Omega(\pi_1, \pi_2; \gamma_0) &= S_\psi V^\dagger((\psi_0, \pi_1), (\psi_0, \pi_2); \gamma_0) S_\psi', \\
H(\pi; \gamma_0) &= E_{\gamma_0} \rho_{\psi\psi}(W_i, \psi_0, \pi), \\
J(\gamma_0) &= E_{\gamma_0} \rho_{\theta\theta}^\dagger(W_i, \theta_0), \text{ and} \\
V(\gamma_0) &= V^\dagger(\theta_0, \theta_0; \gamma_0).
\end{aligned} \tag{3.33}
$$

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

$$
\begin{aligned}
Q(\theta; \gamma_0) &= E_{\gamma_0} \rho(W_i, \theta) = E_{\gamma_0} E_{\gamma_0}(\rho(W_i, \theta)|X_i, Z_i) \\
&= -E_{\gamma_0}[L_i(\theta_0) \log L_i(\theta) + (1 - L_i(\theta_0)) \log(1 - L_i(\theta))]. \tag{3.34}
\end{aligned}
$$

---

[9]In (3.32), the expectation $E_{\gamma_0}(\cdot)$ only depends on ϕ_0. Because θ_0 shows up in some other expectations, we use $E_{\gamma_0}(\cdot)$ throughout the example for notational consistency.

By calculations given in Section 13.1 of Supplemental Appendix D, we have

$$\Omega(\pi_1, \pi_2; \gamma_0) = E_{\gamma_0} \frac{L_i'^2(\theta_0)}{L_i(\theta_0)(1 - L_i(\theta_0))} d_{\psi,i}(\pi_1) d_{\psi,i}(\pi_2)',$$

$$H(\pi; \gamma_0) = E_{\gamma_0} \frac{L_i'^2(\theta_0)}{L_i(\theta_0)(1 - L_i(\theta_0))} d_{\psi,i}(\pi) d_{\psi,i}(\pi)', \text{ and}$$

$$J(\gamma_0) = V(\gamma_0) = E_{\gamma_0} \frac{L_i'^2(\theta_0)}{L_i(\theta_0)(1 - L_i(\theta_0))} d_i(\pi_0) d_i(\pi_0)'. \tag{3.35}$$

$\square$

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 ψ around the point $\psi_{0,n} = (0, \zeta_n)$ for given π for the form:

$$Q_n(\psi, \pi) = Q_n(\psi_{0,n}, \pi) + D_\psi Q_n(\psi_{0,n}, \pi)'(\psi - \psi_{0,n}) +$$
$$\frac{1}{2}(\psi - \psi_{0,n})' D_{\psi\psi} Q_n(\psi_{0,n}, \pi)(\psi - \psi_{0,n}) + R_n(\psi, \pi), \tag{3.36}$$

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 ψ, 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 ψ 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 expansion in θ around the true value θ_n of the form:

$$Q_n(\theta) = Q_n(\theta_n) + DQ_n(\theta_n)'(\theta - \theta_n) + \frac{1}{2}(\theta - \theta_n)D^2 Q_n(\theta_n)(\theta - \theta_n) + R_n^*(\theta), \tag{3.37}$$

where $DQ_n(\theta_n)$ and $D^2 Q_n(\theta_n)$ denote the vector and matrix of first and second partial derivatives of $Q_n(\theta)$ with respect to θ, respectively, evaluated at $\theta = \theta_n$, and $R_n^*(\theta)$ is a

---

[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 Appendix A.

remainder term that is small for θ close to θ_n.[11]

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

$$G_n(\cdot) \Rightarrow G(\cdot; \gamma_0), \text{ where}$$

$$G_n(\pi) = n^{-1/2} \sum_{i=1}^{n} \left(\rho_{\psi,i}(\psi_{0,n}, \pi) - E_{\gamma_n} \rho_{\psi,i}(\psi_{0,n}, \pi) \right) \tag{3.38}$$

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

$$n^{1/2} B^{-1}(\beta_n) DQ_n(\theta_n) \to_d G^*(\gamma_0) \sim N(0_{d_\theta}, V(\gamma_0)) \tag{3.39}$$

and

$$J_n = B^{-1}(\beta_n) D^2 Q_n(\theta_n) B^{-1}(\beta_n) \to_p J(\gamma_0) \in R^{d_\theta \times d_\theta} \; \forall \gamma_0 \in \Gamma. \tag{3.40}$$

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 ψ 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

$$\xi(\pi; \gamma_0, b) = -\frac{1}{2} \left(G(\pi; \gamma_0) + K(\pi; \gamma_0) b \right)' H^{-1}(\pi; \gamma_0) \left(G(\pi; \gamma_0) + K(\pi; \gamma_0) b \right) \text{ and}$$

$$\eta(\pi; \gamma_0, \omega_0) = -\frac{1}{2} \omega_0' K(\pi; \gamma_0)' H^{-1}(\pi; \gamma_0) K(\pi; \gamma_0) \omega_0. \tag{3.41}$$

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.

---

[11]The precise conditions that the remainder $R_n^*(\theta)$ satisfies are specified in Assumption D2 of AC1.

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

Assumption C6. Each sample path of the stochastic process $\{\xi(\pi; \gamma_0, b) : \pi \in \Pi\}$ in some set $A(\gamma_0, b)$ with $P_{\gamma_0}(A(\gamma_0, b)) = 1$ is minimized over Π at a unique point (which 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 β 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 π by Assumption S2(ii) and is denoted by $\rho_\zeta(w, \psi)$. When $d_\beta = 1$ and $\beta_0 = 0$, define

$$\rho_\psi^*(W_i, \psi_0, \pi_1, \pi_2) = (\rho_\beta(W_i, \psi_0, \pi_1), \rho_\beta(W_i, \psi_0, \pi_2), \rho_\zeta(W_i, \psi_0)')' \text{ and}$$

$$\Omega_G(\pi_1, \pi_2; \gamma_0) = \sum_{m=-\infty}^{\infty} Cov_{\gamma_0}(\rho_\psi^*(W_i, \psi_0, \pi_1, \pi_2), \rho_\psi^*(W_{i+m}, \psi_0, \pi_1, \pi_2)). \quad (3.42)$$

Assumption C6†. (i) $d_\beta = 1$ (i.e., β 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 3.2. *Assumptions* S1-S3 *and* C6† *imply Assumption* C6.

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

$$\Omega_G(\pi_1, \pi_2; \gamma_0) = E_{\gamma_0} \frac{L'^2(Z_i'\zeta_0)}{L(Z_i'\zeta_0)(1 - L(Z_i'\zeta_0))} h_{Z,i}(\pi_1, \pi_2) h_{Z,i}(\pi_1, \pi_2)', \text{ where}$$

$$h_{Z,i}(\pi_1, \pi_2) = (h(X_i, \pi_1), h(X_i, \pi_2), Z_i')'. \quad (3.43)$$

$\square$

The following assumption is used in the proof of consistency of $\widehat{\pi}_n$ for the case where the true parameter β_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 π_0 $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$.

In Assumption C7, π_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 (3.35), when $\beta_0 = 0$,

$$H(\pi; \gamma_0) = E_{\gamma_0} \frac{L'^2(Z_i'\zeta_0)}{L(Z_i'\zeta_0)(1 - L(Z_i'\zeta_0))} d_{\psi,i}(\pi) d_{\psi,i}(\pi)'. \tag{3.44}$$

By (2.4) and (3.30), when $\beta_0 = 0$,

$$K(\pi; \gamma_0) = E_{\gamma_0} \frac{L'^2(Z_i'\zeta_0)}{L(Z_i'\zeta_0)(1 - L(Z_i'\zeta_0))} h(X_i, \pi_0) d_{\psi,i}(\pi). \tag{3.45}$$

Hence, when $\beta_0 = 0$,

$$K(\pi; \gamma_0)' H^{-1}(\pi; \gamma_0) K(\pi; \gamma_0) \leq E_{\gamma_0} \frac{L'^2(Z_i'\zeta_0)}{L(Z_i'\zeta_0)(1 - L(Z_i'\zeta_0))} h^2(X_i, \pi_0) \tag{3.46}$$

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 $b \in R^{d_\varsigma+1}$ 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 (3.32). $\square$

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 ψ for given $\pi \in \Pi$ by

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

Let $Q_n^c(\pi)$ denote the concentrated 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.

24

For $\gamma_n = (\beta_n, \zeta_n, \pi_n, \phi_n) \in \Gamma$, let $Q_{0,n} = Q_n(\psi_{0,n}, \pi)$, where $\psi_{0,n} = (0, \zeta_n)$. Note that $Q_{0,n}$ does not depend on π by Assumption 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 (3.41) 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

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

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

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 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 θ obtained by inverting QLR tests. The function $r(\theta)$ is assumed to be smooth and to be of the form

$$r(\theta) = \begin{bmatrix} r_1(\psi) \\ r_2(\pi) \end{bmatrix}, \tag{5.7}$$

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

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

$$\widetilde{\theta}_n(v) \in \Theta, \ r(\widetilde{\theta}_n(v)) = v, \ \text{and} \ Q_n(\widetilde{\theta}_n(v)) = \inf_{\theta \in \Theta : r(\theta) = v} Q_n(\theta) + o(n^{-1}). \tag{5.8}$$

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

$$QLR_n(v) = 2n(Q_n(\widetilde{\theta}_n(v)) - Q_n(\widehat{\theta}_n))/\widehat{s}_n, \tag{5.9}$$

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.

26

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

$$CS_{r,n}^{QLR} = \{v \in r(\Theta) : QLR_n(v) \leq c_{n,1-\alpha}(v)\}. \tag{5.10}$$

5.2. QLR Assumptions

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

$$\Pi_r(v_2) = \{\pi \in \Pi : r_2(\pi) = v_2 \text{ for some } \theta = (\psi, \pi) \in \Theta\}. \tag{5.11}$$

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 Θ.

(ii) $r_\theta(\theta) \ (= (\partial/\partial\theta')r(\theta))$ is full row rank $d_r \ \forall\theta \in \Theta$.

(iii) $r(\theta)$ satisfies (5.7).

(iv) $d_H(\Pi_r(v_2), \Pi_r(v_{0,2})) \to 0$ as $v_2 \to v_{0,2} \ \forall v_{0,2} \in r_2(\Theta^*)$.

(v) $Q(\psi, \pi; \gamma_0)$ is continuous in ψ at ψ_0 uniformly over $\pi \in \Pi$ (i.e., $\sup_{\pi \in \Pi} |Q(\psi, \pi; \gamma_0) - Q(\psi_0, \pi; \gamma_0)| \to 0$ as $\psi \to \psi_0$) $\forall\gamma_0 \in \Gamma$ with $\beta_0 = 0$.

(vi) $Q(\theta; \gamma_0)$ is continuous in θ at $\theta_0 \ \forall\gamma_0 \in \Gamma$ 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 ψ and π. 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.

27

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 σ^2 and $\widehat{s}_n$ equal to a consistent estimator of σ^2, such as the sample variance based on the residuals.

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:

$$QLR_n = QLR_n(v_n) \text{ and } \widetilde{\theta}_n = \widetilde{\theta}_n(v_n), \text{ where } v_n = r(\theta_n) \text{ and } \gamma_n = (\theta_n, \phi_n).^{13} \quad (5.12)$$

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

---

[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.

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

$$\xi_r(\pi; \gamma_0, b) = \xi(\pi; \gamma_0, b) + \frac{1}{2}\tau(\pi; \gamma_0, b)'P_\psi(\pi; \gamma_0)'H(\pi; \gamma_0)P_\psi(\pi; \gamma_0)\tau(\pi; \gamma_0, b), \text{ where}$$

$$P_\psi(\pi; \gamma_0) = H^{-1}(\pi; \gamma_0)r_{1,\psi}(\psi_0)'\left(r_{1,\psi}(\psi_0)H^{-1}(\pi; \gamma_0)r_{1,\psi}(\psi_0)'\right)^{-1} r_{1,\psi}(\psi_0), \qquad (5.13)$$

$r_{1,\psi}(\psi) = (\partial/\partial\psi')r_1(\psi) \in R^{d_{r_1} \times d_\psi}$, 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

$$\lambda_{QLR}(\gamma_0) = G^*(\gamma_0)'J^{-1}(\gamma_0)P_\theta(\gamma_0)'J(\gamma_0)P_\theta(\gamma_0)J^{-1}(\gamma_0)G^*(\gamma_0),$$

$$P_\theta(\gamma_0) = J^{-1}(\gamma_0)r_\theta(\theta_0)'\left(r_\theta(\theta_0)J^{-1}(\gamma_0)r_\theta(\theta_0)'\right)^{-1} r_\theta(\theta_0), \qquad (5.14)$$

$r_\theta(\theta_0) = (\partial/\partial\theta')r(\theta_0)$, and $J(\gamma_0)$ and $G^*(\gamma_0)$ are defined in (3.33) and (3.39), 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* $||b|| < \infty$,
$QLR_n \to_d 2(\inf_{\pi \in \Pi_{r,0}} \xi_r(\pi; \gamma_0, b) - \inf_{\pi \in \Pi} \xi(\pi; \gamma_0, b))/s(\gamma_0)$.
(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,

$$QLR_n \to_d \lambda_{QLR}(\gamma_0)/s(\gamma_0) \sim \chi^2_{d_r}. \qquad (5.15)$$

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 (5.10), 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

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

$$QLR(h) = 2(\inf_{\pi \in \Pi_{r,0}} \xi_r(\pi; \gamma_0, b) - \inf_{\pi \in \Pi} \xi(\pi; \gamma_0, b))/s(\gamma_0) \qquad (5.16)$$

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*

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

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

$$c_{QLR,1-\alpha}^{LF} = \max\{\sup_{h \in H} c_{QLR,1-\alpha}(h), \chi_{d_r,1-\alpha}^2\}. \tag{5.17}$$

The LF critical value can be improved (i.e., made smaller) by exploiting the knowledge of the null hypothesis value of $r(\theta)$. For instance, if the null hypothesis specifies the value of π to be 3, then the supremum in (5.17) 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 γ is unknown under H_0 but can be consistently estimated, then a *plug-in* LF (or plug-in NI-LF) critical value can be used that has correct size asymptotically and is smaller than the LF (or NI-LF) critical value. The plug-in critical value replaces elements of γ with consistent estimators in the formulae in (5.17) and the supremum over H is reduced to a supremum over the resulting subset of H, denoted $\widehat{H}_n$, for which the consistent estimators appear in each vector γ.[16]

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

---

[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 θ_0 denotes the true value. The NI-LF critical value, denoted $c_{QLR,1-\alpha}^{LF}(v)$, is defined by replacing H by $H(v)$ in (5.17) when the null hypothesis value is $r(\theta_0) = v$. Note that v takes values in the set $V_r = \{v_0 : r(\theta_0) = v_0$ 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 ζ is consistently estimated by $\widehat{\zeta}_n$, then H is replaced by $\widehat{H}_n = \{h = (b, \gamma) \in H : \gamma = (\beta, \widehat{\zeta}_n, \pi, \phi)\}$. If a plug-in NI-LF critical value is employed, $H(v)$ is replaced by $H(v) \cap \widehat{H}_n$, where $H(v)$ is defined in a footnote above. 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 β, it is not necesary to use an ICS procedure. Instead, we recommend using a (possibly plug-in) NI-LF critical value, see the footnote above.

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 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

$$\widehat{J}_n(\theta) = \widehat{V}_n(\theta) = n^{-1} \sum_{i=1}^{n} \frac{L_i'^2(\theta)}{L_i(\theta)(1 - L_i(\theta))} d_i(\pi) d_i(\pi)'. \tag{5.18}$$

$\square$

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

$$A_n = \left(n \widehat{\beta}_n' \widehat{\Sigma}_{\beta\beta,n}^{-1} \widehat{\beta}_n / d_\beta \right)^{1/2}, \tag{5.19}$$

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

The type 2 robust critical value is

$$\widehat{c}_{QLR,1-\alpha,n} = \begin{cases} c_B & \text{if } A_n \le \kappa \\ c_S + [c_B - c_S] \cdot s(A_n - \kappa) & \text{if } A_n > \kappa, \text{ where} \end{cases}$$
$$c_B = c_{QLR,1-\alpha}^{LF} + \Delta_1, \ c_S = \chi_{d_r,1-\alpha}^2 + \Delta_2, \tag{5.20}$$

and $\Delta_1 \ge 0$ and $\Delta_2 \ge 0$ are asymptotic size-correction factors that are defined below. Here, "B" denotes Big, and "S" denotes Small. When $A_n \le \kappa$, $\widehat{c}_{QLR,1-\alpha,n}$ equals the LF critical value $c_{QLR,1-\alpha}^{LF}$ plus a size-correction factor Δ_1. When $A_n > \kappa$, $\widehat{c}_{QLR,1-\alpha,n}$

---

[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 (5.20) 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.

is a convex combination of $c_{QLR,1-\alpha}^{LF} + \Delta_1$ and $\chi_{d_r,1-\alpha}^2 + \Delta_2$, where Δ_2 is another size-correction factor and the weight given to the standard critical value $\chi_{d_r,1-\alpha}^2$ increases with the strength of identification, as measured by $A_n - \kappa$.

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

$$A(h) = \left(\tau_\beta(\pi^*; \gamma_0, b)' \Sigma_{\beta\beta}^{-1}(\pi^*; \gamma_0) \tau_\beta(\pi^*; \gamma_0, b)/d_\beta \right)^{1/2}, \tag{5.21}$$

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

$$NRP(\Delta_1, \Delta_2; h) = P(QLR(h) > c_B \ \& \ A(h) \le \kappa) + P(QLR(h) > c_A(h) \ \& \ A(h) > \kappa)$$
$$= P(QLR(h) > c_B) + P(QLR(h) \in (c_A(h), c_B] \ \& \ A(h) > \kappa), \ \text{where}$$
$$c_A(h) = c_S + (c_B - c_S) \cdot s(A(h) - \kappa). \tag{5.22}$$

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

---

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

[20] In the vector β case, $\Sigma_{\beta\beta}^{-1}(\pi^*; \gamma_0)$ is replaced in (5.21) by a slightly different expresssion, see footnote 51 of AC1. When the type 2 robust critical value is considered in the vector β case, h is defined to include $\omega_0 = \lim_{n \to \infty} \beta_n/||\beta_n|| \in R^{d_\beta}$ as an element, i.e., $h = (b, \gamma_0, \omega_0)$ and $H = \{h = (b, \gamma_0, \omega_0) : ||b|| < \infty, \gamma_0 \in \Gamma \text{ with } \beta_0 = 0, ||\omega_0|| = 1\}$ because the true value ω_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 β in place of $\widehat{\beta}_n$ and a different weight matrix. See AC1 for details.

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

[23] When the LF critical value is achieved at $||b|| = \infty$, i.e., $\chi_{d_r,1-\alpha}^2 \ge \sup_{h \in H} c_{QLR,1-\alpha}(h)$, the

tionally convenient.

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

The type 2 robust critical value can be improved by employing NI and/or plug-in versions of it, denoted by $\widehat{c}_{QLR,1-\alpha,n}(v)$. These are defined by replacing $c_{QLR,1-\alpha}^{LF}$ in (5.20) by the NI-LF or plug-in NI-LF critical value and making Δ_1 and Δ_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 10.1 in Appendix B.

For any given value of κ, the type 2 robust CS has correct asymptotic size due to the choice of Δ_1 and Δ_2. In consequence, a good choice of κ 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 κ 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 κ. The reason is that the size-correction constants Δ_1 and Δ_2 have to adjust as κ is changed in order to maintain correct asymptotic size. The adjustments of Δ_1 and Δ_2 offset the effect of changing κ.

One can select κ in a simple way, i.e., by taking $\kappa = 1.5$ or 2.0, or one can select κ in a more sophisticated way that explicitly depends on FCP's. (See 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

---

standard asymptotic critical value $\chi^2_{d_r,1-\alpha}$ yields a test or CI with correct asymptotic size and constants Δ_1 and Δ_2 are not needed. Hence, here we consider the case where $||b_{\max}|| < \infty$. If $\sup_{h \in H} c_{QLR,1-\alpha}(h)$ is not attained at any point $h_{\max}$, then $b_{\max}$ can be taken to be any point such that $c_{QLR,1-\alpha}(h_{\max})$ is arbitrarily close to $\sup_{h \in H} c_{QLR,1-\alpha}(h)$ for some $h_{\max} = (b_{\max}, \gamma_{\max}) \in H$.

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

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

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

---

[24]See Theorems 4.1, 4.4(a), and 5.1(a) in Sections 4.1, 4.7, and 5, respectively, in AC1. Lemma 9.1 of 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.

our identification set-up, we write the model as

$$Y_t = X_t'\zeta + X_t'\beta \cdot m(Z_t, \pi) + U_t, \text{ where}$$
$$X_t = (1, Y_{t-1}, ..., Y_{t-p})', \ Z_t = Y_{t-d}, \tag{7.3}$$

$\{Y_t : t = 1, ..., n\}$ are observed random variables, $\{U_t : t = 1, ..., n\}$ are unobserved innovations, and $m(\cdot, \cdot)$ is a known transition function. 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

$$m(Z_t, \pi) = (1 + \exp[-\pi_1(Z_t - \pi_2)])^{-1} \tag{7.4}$$

and the second one is the exponential function

$$m(Z_t, \pi) = 1 - \exp[-\pi_1(Z_t - \pi_2)^2], \tag{7.5}$$

where $\pi = (\pi_1, \pi_2)' \in R^2$, $\pi_1 > 0$ measures the slope of the transition, and π_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

$$Q_n(\theta) = n^{-1} \sum_{t=1}^n U_t^2(\theta)/2, \text{ where } U_t(\theta) = Y_t - X_t'\zeta - X_t'\beta \cdot m(Z_t, \pi). \tag{7.6}$$

The LS estimator of θ minimizes $Q_n(\theta)$ over $\theta \in \Theta$. The optimization parameter space Θ takes the form

$$\Theta = \{(\beta, \zeta, \pi) : \beta \in \mathcal{B}, \ \zeta \in \mathcal{Z}(\beta), \ \pi \in \Pi\}. \tag{7.7}$$

We show in Supplemental Appendix E that under the assumptions given below Assumptions S1-S4, B1, B2, C6, C7, V1, and V2 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 Φ is a compact metric space with some metric d_Φ that induces weak convergence of the bivariate distributions (Y_t, Y_{t+m}) for all $t, m \geq 1$.[25] In this model, ϕ is an infinite-dimensional nuisance para-

---

[25] For example, the metric d_Φ can be defined as follows. Let $\{u_t\}_1$ and $\{u_t\}_2$ denote two in-

meter. The true value of $\gamma = (\theta, \phi)$, denoted by γ_0, belongs to a compact set Γ. 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 γ_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 ζ_{int} and β_{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 π. 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 γ_0.

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

---

finite $\{u_t : t = ..., -1, 0, 1, ...\}$ 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 θ, 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_Φ 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 Θ^* is the true parameter space for θ and d_2 is some metric on the space of bivariate distributions that induces weak convergence.

continuous sample paths and covariance kernel $\Omega(\pi_1, \pi_2; \gamma_0)$ for $\pi_1, \pi_2 \in \Pi$, where

$$\Omega(\pi_1, \pi_2; \gamma_0) = E_{\gamma_0} U_t^2 d_{\psi,t}(\pi_1) d_{\psi,t}(\pi_2)' \text{ and}$$
$$d_{\psi,t}(\pi) = (X_t' m(Z_t, \pi), X_t')'. \tag{7.8}$$

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

$$\xi(\pi; \gamma_0, b) = -\frac{1}{2} \left(G(\pi; \gamma_0) + K(\pi; \gamma_0) b\right)' H^{-1}(\pi; \gamma_0) \left(G(\pi; \gamma_0) + K(\pi; \gamma_0)b\right) \text{ and}$$
$$\tau_\beta(\pi; \gamma_0, b) = -S_\beta H^{-1}(\pi; \gamma_0) \left(G(\pi; \gamma_0) + K(\pi; \gamma_0)b\right), \text{ where}$$
$$K(\pi; \gamma_0) = -E_{\gamma_0} d_{\psi,t}(\pi) d_{\psi,t}(\pi_0)' \cdot S_\beta', \; S_\beta = [I_{d_\beta} : 0] \in R^{d_\beta \times d_\psi}, \text{ and}$$
$$H(\pi; \gamma_0) = E_{\gamma_0} d_{\psi,t}(\pi) d_{\psi,t}(\pi)'. \tag{7.9}$$

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 Π at a unique point (which may depend on the sample path), denoted $\pi^*(\gamma_0, b)$, $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$, $\forall b$ with $||b|| < \infty$.
(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

$$\Theta^* = \{(\beta, \zeta, \pi) : \beta \in \mathcal{B}^*, \; \zeta \in \mathcal{Z}^*(\beta), \; \pi \in \Pi^*\}. \tag{7.10}$$

In (7.10), Θ^* is not a product space. For any $\beta \in \mathcal{B}^*$, ζ^* 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 ϕ. The true parameter spaces Θ^* and $\Phi(\theta_0)$ are assumed to satisfy Assumption STAR4.

Assumption STAR4. (i) Θ^* 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 Γ 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 Γ 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 π_1 away from 0 because our focus is on the weak identification of π that occurs when β is close to 0, rather than a different sort of weak identification that occurs when π_1 is close to 0. Assumption STAR4(iv) is employed in the verification of Assumption B2(iii).

The optimization parameter space Θ defined in (7.7) 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 Ψ 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).

7.3. Key Quantities

The quantities $G(\cdot; \gamma_0)$, $K(\pi; \gamma_0)$, $H(\pi; \gamma_0)$, and $\xi(\cdot; \gamma_0)$ in (7.9) appear in Theorem 4.1.

In the STAR model, $W_t = (Y_t, X_t', Z_t')'$. The criterion function in (7.6) takes the form

$$Q_n(\theta) = n^{-1} \sum_{t=1}^{n} \rho(W_t, \theta), \text{ where } \rho(W_t, \theta) = \frac{1}{2}U_t^2(\theta). \tag{7.11}$$

The first- and second-order partial derivatives of $\rho(W_t, \theta)$ wrt ψ and θ are

$$\rho_\psi(W_t, \theta) = -U_t(\theta)d_{\psi,t}(\pi), \ \rho_\theta(W_t, \theta) = -U_t(\theta)d_{\theta,t}(\theta),$$

$$\rho_{\psi\psi}(W_t, \theta) = d_{\psi,t}(\pi)d_{\psi,t}(\pi)',$$

$$\rho_{\theta\theta}(W_t, \theta) = d_{\theta,t}(\theta)d_{\theta,t}(\theta)' - U_t(\theta)D_t(\theta), \ \text{where}$$

$$d_{\psi,t}(\pi) = (X_t'm(Z_t, \pi), X_t')', \ d_{\theta,t}(\theta) = (X_t'm(Z_t, \pi), X_t', \beta'X_tm_\pi(Z_t, \pi)')' \ \text{and}$$

$$D_t(\theta) = \begin{bmatrix} 0_{d_\beta \times d_\beta} & 0_{d_\beta \times d_\zeta} & X_tm_\pi(Z_t, \pi)' \\ 0_{d_\zeta \times d_\beta} & 0_{d_\zeta \times d_\zeta} & 0_{d_\zeta \times d_\pi} \\ m_\pi(Z_t, \pi)X_t' & 0_{d_\pi \times d_\zeta} & \beta'X_t \cdot m_{\pi\pi}(Z_t, \pi) \end{bmatrix}. \tag{7.12}$$

Define

$$d_t(\pi, \omega) = (X_t'm(Z_t, \pi), X_t', \omega'X_tm_\pi(Z_t, \pi)')'. \tag{7.13}$$

The re-scaled partial derivatives in (3.19) take the form

$$\rho_\theta^\dagger(W_t, \theta^+) = -U_t(\theta^+)d_t(\pi, \omega), \ \rho_{\theta\theta}^\dagger(W_t, \theta^+) = d_t(\pi, \omega)d_t(\pi, \omega)', \ \text{and}$$

$$\varepsilon(W_t, \theta^+) = -U_t(\theta^+) \begin{bmatrix} 0_{d_\beta \times d_\beta} & 0_{d_\beta \times d_\zeta} & X_tm_\pi(Z_t, \pi)' \\ 0_{d_\zeta \times d_\beta} & 0_{d_\zeta \times d_\zeta} & 0_{d_\zeta \times d_\pi} \\ m_\pi(Z_t, \pi)X_t' & 0_{d_\pi \times d_\zeta} & \omega'X_t \cdot m_{\pi\pi}(Z_t, \pi) \end{bmatrix}, \ \text{where}$$

$$U_t(\theta^+) = Y_t - X_t'\zeta - X_t'\omega||\beta|| \cdot m(Z_t, \pi). \tag{7.14}$$

Let

$$V^\dagger(\theta_0^+, \theta_0^+; \gamma_0) = V(\gamma_0) = E_{\gamma_0}U_t^2 d_t(\pi_0, \omega_0)d_t(\pi_0, \omega_0)' \ \text{and}$$

$$J(\gamma_0) = E_{\gamma_0}d_t(\pi_0, \omega_0)d_t(\pi_0, \omega_0). \tag{7.15}$$

The quantities in (7.12), (7.14), and (7.15) appear in Assumptions S1-S4. The matrices $J(\gamma_0)$ and $V(\gamma_0)$ appear in Theorem 4.2.

7.4. Variance Matrix Estimators

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

$$\widehat{J}_n = \widehat{J}_n(\widehat{\theta}_n^+) \text{ and } \widehat{V}_n = \widehat{V}_n(\widehat{\theta}_n^+), \text{ where} \tag{7.16}$$

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

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

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 $U_t \sim N(0, 1)$. The true values of ζ_1 and ζ_2 are -1 and 0.5, respectively. The true parameter space for π is $[-3.5, -1.5]$ and the optimization space for π is $[-4, -1]$. The number of simulation repetitions is 20,000.[26]

Figures 1 and 2 provide the asymptotic and finite-sample densities of the ML estimators of β and π in the STAR model when the true π value is $\pi_0 = -1.5$. Each Figure gives the densities for $b = 0$, 2, 4, and 10, where b indexes the magnitude of β. 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 $\pi_0 = -3.0$.

---

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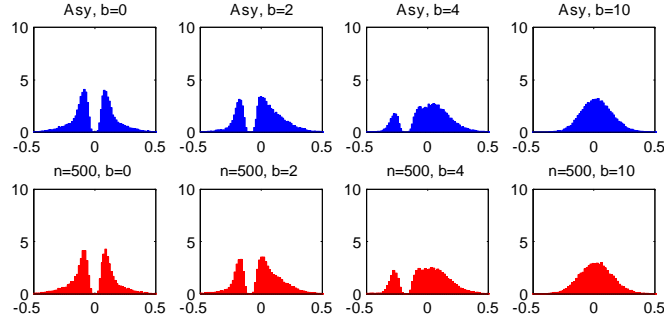
Figure 1. Asymptotic and Finite-Sample ($n = 500$) Densities of the Estimator of β in the STAR Model when $\pi_0 = -1.5$.
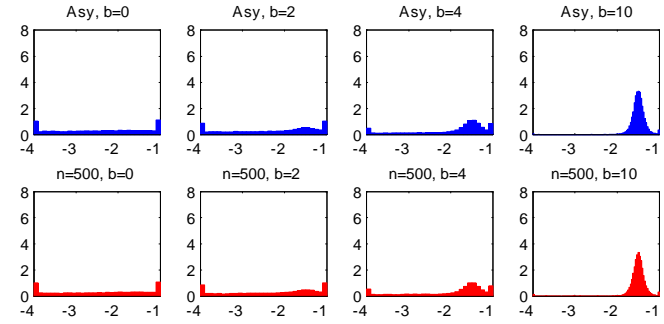

Figure 2. Asymptotic and Finite-Sample ($n = 500$) Densities of the Estimator of π in the STAR Model when $\pi_0 = -1.5$.

Figure 1 shows that the ML estimator of β 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 π in the unidentified and weakly-identified cases. Figures 1 and 2 indicate that the asymptotic approximations developed here work very well.

Figures S-3 to S-6 in Supplemental Appendix C provide the asymptotic and finite-sample ($n = 500$) densities of the t and QLR statistics for β and π 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 χ_1^2 distributions. However, the asymptotic approximations developed here work very well.

Figure 3 provides graphs of the 0.95 asymptotic quantiles of the $|t|$ and QLR statistics concerning β and π 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 β, for small to medium b values, the graphs exceed the 0.95 quantiles under strong identification (given by the horizontal black lines). This
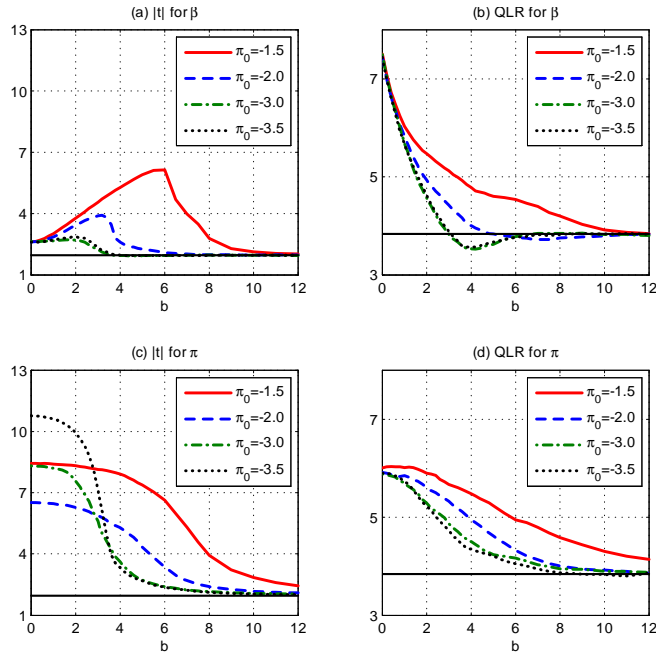
Figure 3. Asymptotic 0.95 Quantiles of the $|t|$ and QLR Statistics for Tests Concerning β and π in the STAR Model.
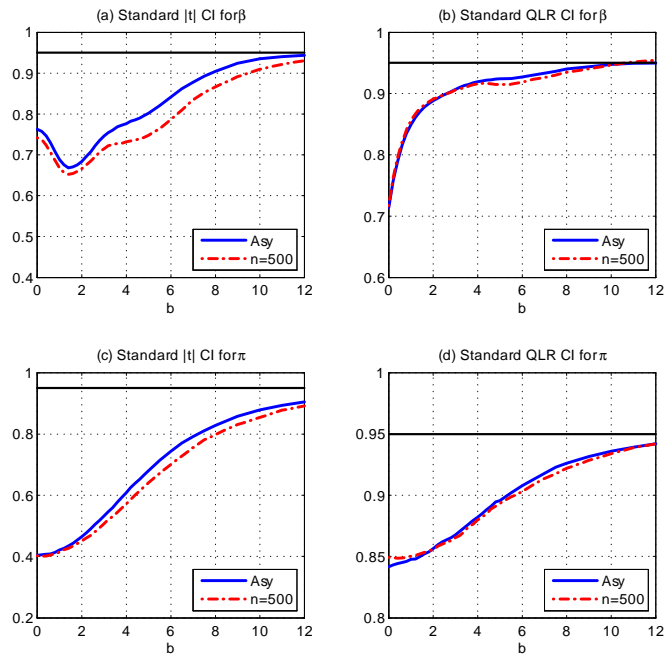
Figure 4. Coverage Probabilities of Standard $|t|$ and QLR CI's for β and π in the STAR Model when $\pi_0 = -1.5$.

implies that tests and CI's that employ the $|t|$ statistic for β and the standard critical value (based on the normal distribution) have incorrect size. The same pattern emerges for the QLR statistic for β (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 β and the standard critical value (based on the χ_1^2 distribution) have incorrect size due to the under-coverage for b 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 π 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 β and π 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 β and 0.40 for the $|t|$ CI for π. The corresponding values for the QLR CI's are 0.72 for β and 0.84 for π. Hence, the size distortions for the standard $|t|$ and QLR CI's for β are similar. But, for the CI's for π, 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 β, we impose the null value of $b = n^{1/2}\beta_0$, where β_0 is the true value of β 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 β is as in (5.17), but with h and H replaced by π and Π, respectively, resulting in a smaller LF critical value. The same simplification applies to the NI-LF critical value for the robust $|t|$ CI for β.

As indicated in Figures 3(a) and 3(b), the NI-LF critical values for both $|t|$ and QLR CI's for β are attained at $\pi_0 = -1.5$ for all b values. In consequence, the robust $|t|$ and QLR CI's for β 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 β. 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 β 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 β are not asymptotically similar
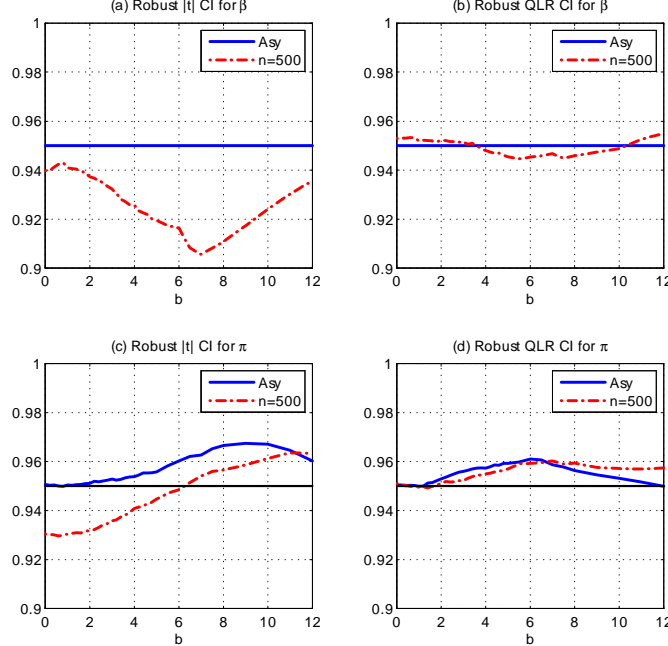
Figure 5. Coverage Probabilities of Robust $|t|$ and QLR CI's for β and π in the STAR Model when $\pi_0 = -1.5$, $\kappa = 2.5$, $D = 1$, and $s(x) = \exp(-x)$.

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 π are constructed with the null value π_0 imposed. Because b is unknown, we apply the smooth transition in (5.20) to obtain critical values for the robust CI's for π. Figures 5(c) and 5(d) report the asymptotic and finite-sample CP's of the robust $|t|$ and QLR CI's for π 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 κ was determined based on minimizing average FCP's. However, a wide range of κ values yield similar results (because the constants Δ_1 and Δ_2 adjust to maintain correct asymptotic size as κ is changed).

Figures 5(c) and 5(d) show that the robust CI's for π 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 π when $\pi_0 = -3.0$ are reported in Figures S-8(c) and S-8(d) in Supplemental Appendix C.

Besides b and π_0, the construction of a robust CI also requires the ζ value in order to obtain the LF (or NI-LF) critical value through simulation. In the STAR model, $\zeta = (\zeta_1, \zeta_2)'$. Because ζ can be consistently estimated, we recommend plugging in the estimator $\widehat{\zeta}_n$ in place of ζ_0 in practice. To ease the computational burden required to simulate the 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 ζ_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 ζ_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 $X_i^* \sim N(3, 1)$, $\mathrm{Corr}(Z_i^*, X_i^*) = 0.5$, and $U_i \sim N(0, 1)$. The true values of ζ_0 and ζ_1 are -2 and 2, respectively. The true parameter space for π is $[1.5, 3.5]$ and the optimization space for π 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.

---

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

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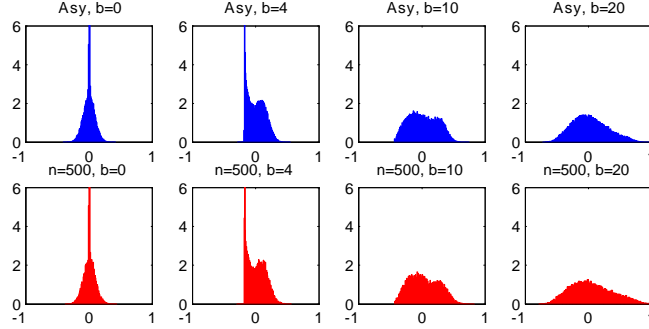
Figure 6. Asymptotic and Finite-Sample (n=500) Densities of the Estimator of β in the Binary Choice Model when $\pi_0 = 1.5$.
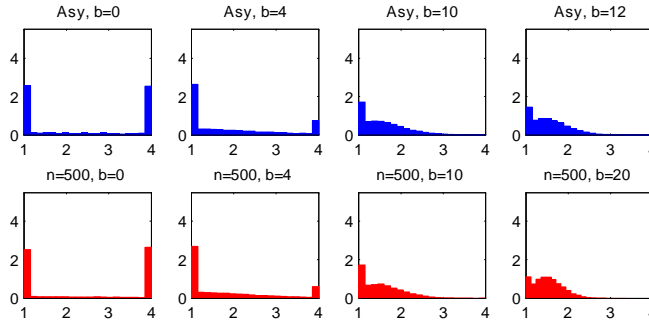

Figure 7. Asymptotic and Finite-Sample (n=500) Densities of the Estimator of π in the Binary Choice Model when $\pi_0 = 1.5$.

The simulation results for the binary choice model are summarized as follows. First, the LS estimators and the $|t|$ and QLR statistics for β and π do not display normal or χ_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]

Second, tests and CI's that employ the $|t|$ and QLR statistics for β 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 β 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 π and the standard critical value have correct asymptotic size and

---

[30] The largest discrepancies between the asymptotic and finite-sample results occur when $\pi_0 = 2.0$ 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 β only under-covers for β very close to zero, which makes it difficult to detect in Figures 8(b) and 9(b).
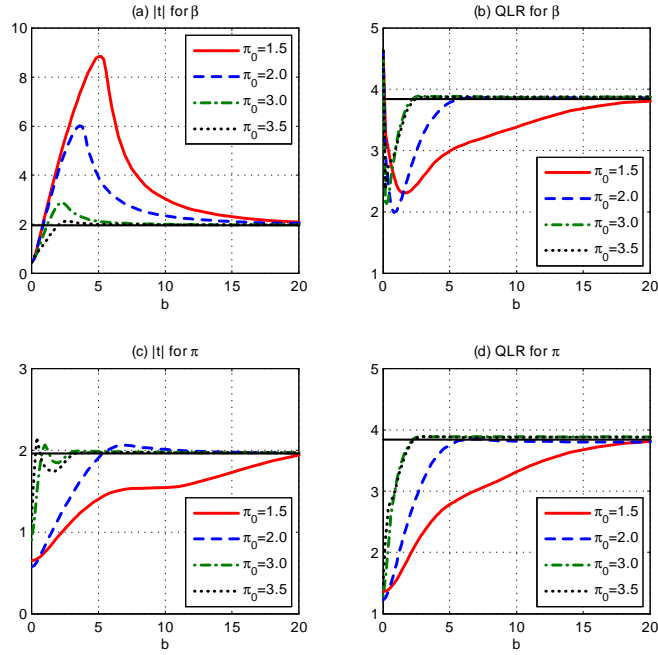
Figure 8. Asymptotic 0.95 Quantiles of the $|t|$ and QLR Statistics for Tests Concerning β and π in the Binary Choice Model.
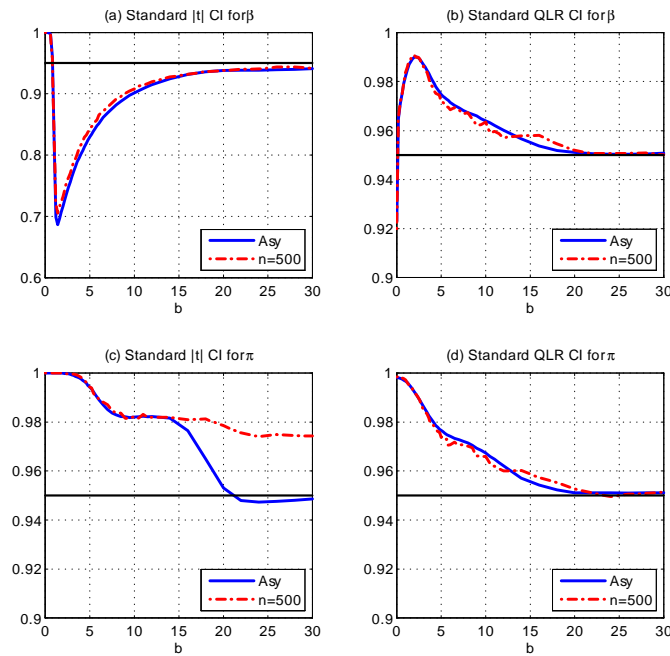

Figure 9. Coverage Probabilities of Standard $|t|$ and QLR CI's for β and π in the Binary Choice Model when $\pi_0 = 1.5$.
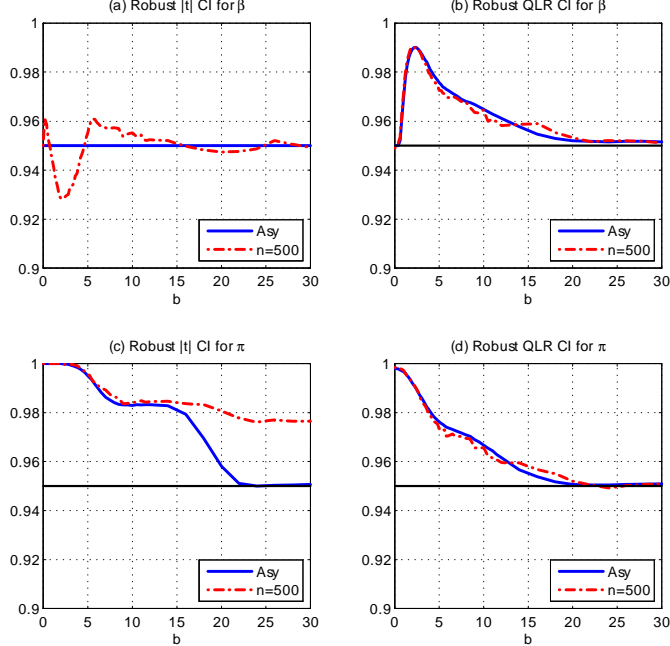
Figure 10. Coverage Probabilities of Robust $|t|$ and QLR CI's for β and π in the Binary Choice Model when $\pi_0 = 1.5$, $\kappa = 1.5$, $D = 1$, and $s(x) = \exp(-x/2)$.

those employ the $|t|$ statistic for π 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 β, 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.

9. Appendix A: Proofs

This Appendix proves the results in Theorems 4.1, 4.2, 5.1, and 5.2. The method of proof is to show that Assumptions B1, B2, and S1-S3 imply certain high-level assumptions in AC1 (specifically, Assumptions A, B3, C1-C4, C8, and D1-D3 of AC1). In addition, it is straightforward that Assumptions S1 and S4 imply Assumption C5 of AC1. Given these results, Theorems 3.1, 3.2, 4.2, 4.3, and 4.4(b) of AC1 imply Theorems 4.1, 4.2, 5.1(a), 5.1(b), and 5.2, respectively, because the results of these theorems are the same, just the assumptions differ.

Lemma 9.1. *Suppose Assumptions* B1 *and* B2 *hold. Assumptions* S1-S3 *imply that Assumptions* A, B3, C1-C4, C8, *and* D1-D3 *of* AC1 *hold with*

$$Q(\theta; \gamma_0) = E_{\gamma_0}\rho(W_i, \theta), \ D_\psi Q_n(\theta) = n^{-1}\sum_{i=1}^n \rho_\psi(W_i, \theta), \ D_{\psi\psi}Q_n(\theta) = n^{-1}\sum_{i=1}^n \rho_{\psi\psi}(W_i, \theta),$$

$$m(W_i, \theta) = \rho_\psi(W_i, \theta), \ \Omega(\pi_1, \pi_2; \gamma_0) = S_\psi V^\dagger((\psi_0, \pi_1), (\psi_0, \pi_2); \gamma_0)S'_\psi,$$

$$H(\pi; \gamma_0) = E_{\gamma_0}\rho_{\psi\psi}(W_i, \psi_0, \pi), \ DQ_n(\theta) = n^{-1}\sum_{i=1}^n \rho_\theta(W_i, \theta),$$

$$D^2 Q_n(\theta) = n^{-1}\sum_{i=1}^n \rho_{\theta\theta}(W_i, \theta), \ J(\gamma_0) = E_{\gamma_0}\rho_{\theta\theta}^\dagger(W_i, \theta_0), \ and \ V(\gamma_0) = V^\dagger(\theta_0, \theta_0; \gamma_0).$$

We start by giving some general results that are useful in the proof of Lemma 9.1. Specifically, Lemma 9.2 is a uniform convergence result for non-stochastic functions, Lemma 9.3 is a uniform LLN, Lemma 9.4 is a stochastic equicontinuity result for empirical processes based on Theorem 3 of Hansen (1996), and Lemma 9.5 is a CLT. All of these results are for strong mixing triangular arrays. The proofs of Lemmas 9.2-9.5 are given below following those of Lemmas 9.1, 3.1, and 3.2.

Lemma 9.2. *Let* $\{q_n(\theta) : n \geq 1\}$ *be non-stochastic functions on* Θ. *Suppose* (i) $q_n(\theta) \to 0 \ \forall \theta \in \Theta$, (ii) $\|q_n(\theta_1) - q_n(\theta_2)\| \leq C\delta \ \forall \theta_1, \theta_2 \in \Theta$ *with* $\|\theta_1 - \theta_2\| \leq \delta$, $\forall n \geq 1$, *for some* $C < \infty$ *and all* $\delta > 0$, *and* (iii) Θ *is compact. Then,* $\sup_{\theta\in\Theta}\|q_n(\theta)\| \to 0$.

Lemma 9.3. *Suppose* (i) *Assumption* S1 *holds,* (ii) *for some function* $M_1(w) : \mathcal{W} \to R^+$ *and all* $\delta > 0$, $\|s(w, \theta_1) - s(w, \theta_2)\| \leq M_1(w)\delta$, $\forall \theta_1, \theta_2 \in \Theta$ *with* $\|\theta_1 - \theta_2\| \leq \delta$, $\forall w \in \mathcal{W}$, (iii) $E_\gamma \sup_{\theta\in\Theta}\|s(W_i, \theta)\|^{1+\varepsilon} + E_\gamma M_1(W_i) \leq C \ \forall \gamma \in \Gamma$ *for some* $C < \infty$ *and* $\varepsilon > 0$, *and* (iv) Θ *is compact. Then,* $\sup_{\theta\in\Theta}\|n^{-1}\sum_{i=1}^n s(W_i, \theta) - E_{\gamma_0}s(W_i, \theta)\| \to_p 0$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0)$ *and* $E_{\gamma_0}s(W_i, \theta)$ *is uniformly continuous on* $\Theta \ \forall \gamma_0 \in \Gamma$.

Comment. Note that the centering term in Lemma 9.3 is $E_{\gamma_0}s(W_i, \theta)$, rather than $E_{\gamma_n}s(W_i, \theta)$.

Lemma 9.4. *Suppose* (i) *Assumption* S1 *holds,* (ii) *for some function* $M_1(w) : \mathcal{W} \to R^+$ *and all* $\delta > 0$, $\|s(w, \theta_1) - s(w, \theta_2)\| \leq M_1(w)\delta$, $\forall \theta_1, \theta_2 \in \Theta$ *with* $\|\theta_1 - \theta_2\| \leq \delta$, $\forall w \in \mathcal{W}$, *and* (iii) $E_\gamma \sup_{\theta\in\Theta}\|s(W_i, \theta)\|^q + E_\gamma M_1(W_i)^q \leq C \ \forall \gamma \in \Gamma$ *for some* $C < \infty$ *and* q *as in Assumption* S1. *Then,* $\nu_n s(\theta) = n^{-1/2}\sum_{i=1}^n(s(W_i, \theta) - E_{\gamma_n}s(W_i, \theta))$ *is stochastically equicontinuous over* $\theta \in \Theta$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0)$, *i.e.,* $\forall \varepsilon > 0$ *and* $\eta > 0$, $\exists \delta > 0$ *such that* $\limsup_{n\to\infty} P[\sup_{\theta_1,\theta_2\in\Theta:\|\theta_1-\theta_2\|<\delta}\|\nu_n s(\theta_1) - \nu_n s(\theta_2)\| > \eta] < \varepsilon \ \forall \gamma_0 \in \Gamma$.

Lemma 9.5. *Suppose* (i) *Assumption* S1 *holds,* (ii) $s(w) \in R$ *and* $E_\gamma |s(W_i)|^q \leq C$ $\forall \gamma \in \Gamma$ *for some* $C < \infty$ *and* q *as in Assumption* S1. *Then,* $n^{-1/2} \sum_{i=1}^n (s(W_i) - E_{\gamma_n} s(W_i)) \to_d N(0, V_s(\gamma_0))$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0)$ $\forall \gamma_0 \in \Gamma$, *where* $V_s(\gamma_0) = \sum_{m=-\infty}^\infty Cov_{\gamma_0}(s(W_i), s(W_{i+m}))$.

Proof of Lemma 9.1. The verification of Assumptions A and B3 are the same for the scalar β and vector β cases. Assumption A follows from Assumptions S2(i) and S2(ii).

Now we verify Assumption B3. In Assumption B3(i), $Q(\theta; \gamma_0) = E_{\gamma_0} \rho(W_i, \theta)$. Assumption B3(i) follows from Lemma 9.3 with $s(w, \theta) = \rho(w, \theta)$ by Assumptions S1, S2(i), S2(v), and S3(iii). Assumptions B3(ii) and B3(iii) can be verified by Assumption B3*. Assumption B3*(i) holds by Assumptions S2(i) and S3(iii). The remaining parts of Assumption B3* follow from Assumption S2 immediately.

We verify the quadratic expansions in Assumptions C1 and D1 using Lemma 11.5 in Appendix A of AC1-SM, which relies on Assumption Q1 of AC1-SM. Assumptions Q1(i) and Q1(ii) follow from Assumption S2(i). Assumption Q1(iii) follows from Lemma 9.3 with $s(w, \theta) = \rho_{\psi\psi}(w, \theta)$.

To verify Assumption Q1(iv) for $\theta \in \Theta_n(\delta_n)$, we write

$$
B^{-1}(\beta_n) n^{-1} \sum_{i=1}^n \rho_{\theta\theta}(W_i, \theta) B^{-1}(\beta_n)
$$

$$
= B(\beta/\iota(\beta_n)) \left(n^{-1} \sum_{i=1}^n \left(\rho_{\theta\theta}^\dagger(W_i, \theta) + \iota^{-1}(\beta)\varepsilon(W_i, \theta) \right) \right) B(\beta/\iota(\beta_n))
$$

$$
= \left(n^{-1} \sum_{i=1}^n \rho_{\theta\theta}^\dagger(W_i, \theta) \right) (1 + o(1)) + \left(n^{-1/2} \sum_{i=1}^n \left(\varepsilon(W_i, \theta) - E_{\gamma_n} \varepsilon(W_i, \theta) \right) \right) \times
$$

$$
(n^{1/2}\iota(\beta_n))^{-1}(1 + o(1)) + \left(E_{\gamma_n} \varepsilon(W_i, \theta)/\iota(\beta_n) \right)(1 + o(1)). \tag{9.1}
$$

In (9.1), the first equality follows from (3.19) and the second equality holds because $B(\beta)$ only depends on β through $\iota(\beta)$, $||\beta|| \leq ||\beta - \beta_n|| + ||\beta_n|| \leq (1 + \delta_n)||\beta_n||$, and $\delta_n = o(1)$. By (9.1) and the fact that $n^{1/2}||\beta_n|| \to \infty$ for $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, to verify Assumption Q1(iv), it suffices to establish the stochastic equicontinuity of $n^{-1} \sum_{i=1}^n \rho_{\theta\theta}^\dagger(W_i, \theta)$ and $n^{-1/2} \sum_{i=1}^n (\varepsilon(W_i, \theta) - E_{\gamma_n} \varepsilon(W_i, \theta))$ over $\theta \in \Theta_n(\delta_n)$ and the equicontinuity of $E_{\gamma_n} \varepsilon(W_i, \theta)/||\beta_n||$ over $\theta \in \Theta_n(\delta_n)$.

When β is a scalar, the stochastic equicontinuity of $n^{-1} \sum_{i=1}^n \rho_{\theta\theta}^\dagger(W_i, \theta)$ follows from Lemma 9.3 using Assumptions S1, S3(ii), and S3(iii). The stochastic equicontinuity

of $n^{-1/2} \sum_{i=1}^{n} (\varepsilon(W_i, \theta) - E_{\gamma_n} \varepsilon(W_i, \theta))$ follows Lemma 9.4 with $s(w, \theta) = \varepsilon(w, \theta)$ using Assumptions S3(ii) and S3(iii).

When β is a vector, the stochastic equicontinuity of $n^{-1} \sum_{i=1}^{n} \rho_{\theta\theta}^\dagger(W_i, \theta^+)$ and $v_n \varepsilon(w, \theta^+) = n^{-1/2} \sum_{i=1}^{n} (\varepsilon(W_i, \theta^+) - E_{\gamma_n} \varepsilon(W_i, \theta^+))$ over $\theta^+ \in \Theta^+$ hold by Lemmas 9.3 and 9.4 using Assumption S3 (vector β). By Andrews (1994, p. 2252), the stochastic equicontinuity of $v_n \varepsilon(w, \theta^+)$ over $\theta^+ \in \Theta^+$ is equivalent to the following: for all sequences of random elements $\{\widehat{\theta}_{1,n}^+ \in \Theta^+ : n \geq 1\}$ and $\{\widehat{\theta}_{2,n}^+ \in \Theta^+ : n \geq 1\}$ that satisfy $||\widehat{\theta}_{1,n}^+ - \widehat{\theta}_{12,n}^+|| \to_p 0$, we have $||v_n \varepsilon(w, \widehat{\theta}_{1,n}^+) - v_n \varepsilon(w, \widehat{\theta}_{2,n}^+)|| \to_p 0$. Note that $v_n \varepsilon(w, \theta^+) = v_n \varepsilon(w, \theta)$, where θ^+ is the reparameterization of θ. Hence, to show the stochastic equicontinuity of $v_n \varepsilon(w, \theta)$ over $\theta \in \Theta_n(\delta_n)$, it is sufficient to show that for all sequences of random elements $\{\widehat{\theta}_{1,n} \in \Theta_n(\delta_n) : n \geq 1\}$ and $\{\widehat{\theta}_{2,n} \in \Theta_n(\delta_n) : n \geq 1\}$, $||\widehat{\theta}_{1,n} - \widehat{\theta}_{12,n}|| \to_p 0$ implies that $||\widehat{\theta}_{1,n}^+ - \widehat{\theta}_{12,n}^+|| \to_p 0$, where $\widehat{\theta}_{i,n}^+$ is the reparameterization of $\widehat{\theta}_{i,n}$ for $i = 1$ and 2. The convergence related to $||\beta||, \zeta$, and π are straightforward. To show $||\widehat{\omega}_{1,n} - \widehat{\omega}_{2,n}|| \to_p 0$, it is sufficient to show that $\widehat{\omega}_n \to_p \omega_0$ for all sequences of random elements $\{\widehat{\theta}_n \in \Theta_n(\delta_n) : n \geq 1\}$ under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ as in Assumption D1. By the definition of $\Theta_n(\delta_n)$, $||\beta_n||^{-1}(\widehat{\beta}_n - \beta_n) = o_p(1)$. This implies that $\widehat{\beta}_n = \beta_n + ||\beta_n|| o_p(1)$ and $||\widehat{\beta}_n||/||\beta_n|| = 1 + o_p(1)$. Hence,

$$\widehat{\omega}_n = \frac{\widehat{\beta}_n}{||\widehat{\beta}_n||} = \frac{\widehat{\beta}_n - \beta_n}{||\beta_n||} \frac{||\beta_n||}{||\widehat{\beta}_n||} + \frac{\beta_n}{||\beta_n||} \frac{||\beta_n||}{||\widehat{\beta}_n||} \to_p \omega_0. \tag{9.2}$$

This completes the verification of the stochastic equicontinuity of $v_n \varepsilon(w, \theta)$ over $\theta \in \Theta_n(\delta_n)$ when β is a vector. The stochastic equicontinuity of $n^{-1} \sum_{i=1}^{n} \rho_{\theta\theta}^\dagger(W_i, \theta)$ over $\theta \in \Theta_n(\delta_n)$ holds by the same reparameterization arguments above in the vector β case.

It remains to show $\sup_{\theta_1, \theta_2 \in \Theta_n(\delta_n)} E_{\gamma_n} (\varepsilon(W_i, \theta_1) - \varepsilon(W_i, \theta_2)) = o(||\beta_n||)$. When β is a scalar, for any $\theta \in \Theta_n(\delta_n)$,

$$\begin{aligned}
&|\beta_n|^{-1} ||E_{\gamma_n} \varepsilon(W_i, \theta)|| \\
&= |\beta_n|^{-1} ||E_{\gamma_n} \varepsilon(W_i, \theta) - E_{\gamma_n} \varepsilon(W_i, \psi_n, \pi) + E_{\gamma_n} \varepsilon(W_i, \psi_n, \pi)|| \\
&\leq C_1 |\beta_n|^{-1} |\beta_n| \delta_n + C_2 ||\pi - \pi_n|| \leq (C_1 + C_2) \delta_n
\end{aligned} \tag{9.3}$$

for some $C_1, C_2 < \infty$ and any constants $\delta_n \to 0$, where the first inequality follows from Assumptions S3(i)-S3(iii) (scalar β) with $\delta = |\beta_n| \delta_n$ in Assumption S3(ii) and $||\theta - (\psi_n, \pi)|| = ||\psi - \psi_n|| \leq |\beta_n| \delta_n$ by the definition of $\Theta_n(\delta_n)$, and the second inequality holds because $\theta \in \Theta_n(\delta_n)$.

When β is a vector, we reparameterize θ as θ^+. For any $\theta \in \Theta_n(\delta_n)$, we have $(1 - \delta_n)||\beta_n|| \leq ||\beta|| \leq (1 + \delta_n)||\beta_n||$ and

$$
\begin{aligned}
||\omega - \omega_n|| &= \left\| \frac{\beta}{||\beta||} - \frac{\beta_n}{||\beta_n||} \right\| = \left\| \frac{\beta - \beta_n}{||\beta_n||} \frac{||\beta_n||}{||\beta||} + \frac{\beta_n}{||\beta_n||} \left(\frac{||\beta_n||}{||\beta||} - 1 \right) \right\| \\
&\leq \delta_n (1 - \delta_n)^{-1} + (1 - \delta_n)^{-1} - 1 = 2(1 - \delta_n)^{-1}\delta_n \leq 4\delta_n,
\end{aligned}
\tag{9.4}
$$

for n large enough that $\delta_n \leq 1/2$, where the first inequality uses the triangle inequality, $||\beta - \beta_n||/||\beta_n|| \leq \delta_n$, and $||\beta_n||/||\beta|| \leq (1 - \delta_n)^{-1}$ and the equalities are straightforward.

Let $\theta^{++} = (||\beta_n||, \omega, \zeta_n, \pi)$. For $\theta \in \Theta_n(\delta_n)$ and large n,

$$
\begin{aligned}
&||\beta_n||^{-1}||E_{\gamma_n}\varepsilon(W_i, \theta)|| \\
&= ||\beta_n||^{-1}||E_{\gamma_n}\varepsilon(W_i, \theta^+)|| \\
&= ||\beta_n||^{-1}||E_{\gamma_n}\varepsilon(W_i, \theta^+) - E_{\gamma_n}\varepsilon(W_i, \theta^{++}) + E_{\gamma_n}\varepsilon(W_i, \theta^{++})|| \\
&\leq C_1||\beta_n||^{-1}||\beta_n||\delta_n + C_2(||\pi - \pi_n|| + ||\omega - \omega_n||) \leq (C_1 + 5C_2)\delta_n
\end{aligned}
\tag{9.5}
$$

for some $C_1, C_2 < \infty$, where the first inequality follows from Assumption S3(i)-S3(iii) (vector β) with $\delta = ||\beta_n||\delta_n$ and $||\theta^+ - \theta^{++}|| \leq ||\beta - \beta_n|| + ||\zeta - \zeta_n|| \leq 2||\beta_n||\delta_n$ by the definition of $\Theta_n(\delta_n)$, and the third inequality holds by $\theta \in \Theta_n(\delta_n)$ and (9.4). This completes the verification of Assumptions C1 and D1 with the stochastic partial derivatives given in Lemma 9.1.

Assumption C2(i) holds with $m(w, \theta) = \rho_\psi(w, \theta)$ by Lemma 11.5(a) of AC1-SM given the verification above of Assumption Q1. Assumptions C2(ii) and C2(iii) follow from Assumptions S2(iii) and S2(iv) given that the true parameter θ_0 lies in the interior of Θ by Assumption B1(i).

The verifications for Assumptions C3, C4, and C8 below are the same for cases with scalar β and vector β because $\rho_\psi(W_i, \theta)$ and $\rho_{\psi\psi}(W_i, \theta)$ do not involve re-scaling with $B(\beta)$.

We now verify Assumption C3. To this end, it is sufficient to show that $\nu_n \rho_\psi(\theta) = n^{-1/2}\sum_{i=1}^n (\rho_\psi(W_i, \theta) - E_{\gamma_n}\rho_\psi(W_i, \theta))$ converges weakly to a Gaussian process on Θ with covariance kernel $S_\psi V^\dagger(\theta_1, \theta_2; \gamma_0)S_\psi'$. The finite-dimensional convergence holds by Lemma 9.5 and the Cramer-Wold device under Assumptions S3(ii) and S3(iii). Note that $\rho_\psi(w, \theta) = S_\psi \rho_\theta^\dagger(w, \theta)$ by the structure of $B(\beta)$. This yields the form of $\Omega(\pi_1, \pi_2; \gamma_0)$ given in Lemma 9.1. The stochastic equicontinuity of $\nu_n \rho_\psi(\theta)$ on $\theta \in \Theta$ follows from Lemma 9.4 with $s(w, \theta) = \rho_\psi(w, \theta)$ under Assumptions S3(ii) and S3(iii). The parameter space

$\Theta \subset R^{d_\theta}$ is compact. Hence, the weak convergence of $v_n \rho_\psi(\theta)$ holds by the Proposition in Andrews (1994, p. 2251).

To show Assumption C4(i) holds with $H(\pi; \gamma_0) = E_{\gamma_0} \rho_{\psi\psi}(W_i, \psi_0, \pi)$, we have

$$\sup_{\pi \in \Pi} ||n^{-1} \sum_{i=1}^{n} \rho_{\psi\psi}(W_i, \psi_{0,n}, \pi) - E_{\gamma_0} \rho_{\psi\psi}(W_i, \psi_0, \pi)||$$

$$\leq \sup_{\theta \in \Theta} ||n^{-1} \sum_{i=1}^{n} \rho_{\psi\psi}(W_i, \theta) - E_{\gamma_0} \rho_{\psi\psi}(W_i, \theta)|| +$$

$$\sup_{\pi \in \Pi} ||E_{\gamma_0} \rho_{\psi\psi}(W_i, \psi_{0,n}, \pi) - E_{\gamma_0} \rho_{\psi\psi}(W_i, \psi_0, \pi)|| \tag{9.6}$$

by the triangle inequality. The first term on the rhs of (9.6) is $o_p(1)$ by Lemma 9.3 with $s(w, \theta) = \rho_{\psi\psi}(w, \theta)$ using Assumptions S1, S2(v), and S3(iii). The second term on the rhs of (9.6) is $o(1)$ because $E_{\gamma_0} \rho_{\psi\psi}(W_i, \theta)$ is continuous in ψ uniformly over $\pi \in \Pi$ by Lemma 9.3. Hence, the rhs of (9.6) is $o_p(1)$, which is the desired result. Assumption C4(ii) holds by Assumptions S3(iii) and S3(iv).

To verify Assumption C8, we have

$$||\frac{\partial}{\partial \psi'} E_{\gamma_n} \rho_\psi(W_i, \psi_n, \pi_n) - E_{\gamma_0} \rho_{\psi\psi}(W_i, \theta_0)|| = ||E_{\gamma_n} \rho_{\psi\psi}(W_i, \psi_n, \pi_n) - E_{\gamma_0} \rho_{\psi\psi}(W_i, \theta_0)||$$

$$\leq \sup_{\theta \in \Theta} ||E_{\gamma_n} \rho_{\psi\psi}(W_i, \theta) - E_{\gamma_0} \rho_{\psi\psi}(W_i, \theta)|| + ||E_{\gamma_0} \rho_{\psi\psi}(W_i, \theta_n) - E_{\gamma_0} \rho_{\psi\psi}(W_i, \theta_0)||, \tag{9.7}$$

where the equality follows from $(\partial/\partial \psi') E_{\gamma_n} \rho_\psi(W_i, \psi_n, \pi_n) = E_{\gamma_n} \rho_{\psi\psi}(W_i, \psi_n, \pi_n)$ using $E_{\gamma_0} \sup_{\theta \in \Theta} ||\rho_{\psi\psi}(W_i, \theta)|| \leq C \ \forall \gamma_0 \in \Gamma$, and the inequality holds by the triangle inequality. The first term in the second line of (9.7) converges to 0 by Lemma 9.2. The conditions for Lemma 9.2 hold by the arguments in the second paragraph of the proof of Lemma 9.3 with $s(w, \theta)$ replaced by $\rho_{\psi\psi}(w, \theta)$. The required conditions are provided in Assumptions S3(ii) and S3(iii). The second term in the second line of (9.7) converges to 0 by the continuity of $E_{\gamma_0} \rho_{\psi\psi}(W_i, \theta)$ in θ, which holds by Lemma 9.3.

To verify Assumption D2, we have

$$J_n = n^{-1} \sum_{i=1}^{n} \rho_{\theta\theta}^\dagger(W_i, \theta_n) + (n^{1/2} \iota(\beta_n))^{-1} n^{-1/2} \sum_{i=1}^{n} \varepsilon(W_i, \theta_n) \tag{9.8}$$

by (3.19). When β is a scalar, by applying Lemma 9.3 with $s(w, \theta) = \rho_{\theta\theta}^{\dagger}(w, \theta)$ and invoking the continuity of $E_{\gamma_0}\rho_{\theta\theta}^{\dagger}(W_i, \theta)$ in θ, we obtain $n^{-1}\sum_{i=1}^{n}\rho_{\theta\theta}^{\dagger}(W_i, \theta_n) \rightarrow_p$ $E_{\gamma_0}\rho_{\theta\theta}^{\dagger}(W_i, \theta_0) = J(\gamma_0)$. Because $n^{1/2}|\beta_n| \rightarrow \infty$, the second summand in (9.8) is $o_p(1)$ provided $n^{-1/2}\sum_{i=1}^{n}\varepsilon(W_i, \theta_n) = O_p(1)$. This is verified by applying the triangular array CLT in Lemma 9.5 with $s(w) = \varepsilon(w, \theta_n)$ using $E_{\gamma_n}\varepsilon(W_i, \theta_n) = 0$ by Assumption S3(i). The above results combine to give $J_n \rightarrow_p J(\gamma_0)$ as desired. The matrix $J(\gamma_0)$ is positive definite by Assumption S3(iv) and symmetric by the construction of $\rho_{\theta\theta}^{\dagger}(W_i, \theta_0)$ in (3.19).

When β is a vector, the verification of Assumption D2 is the same as above by reparameterizing θ as θ^+, replacing Assumption S3 (scalar β) with Assumption S3 (vector β), and using the fact that $\theta_n^+ \rightarrow \theta_0^+$ under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, where θ_n^+ and θ_0^+ are the counterparts of θ_n and θ_0 after reparameterization.

To verify Assumption D3, we have

$$n^{-1/2}B^{-1}(\beta_n)DQ_n(\theta_n) = n^{-1/2}\sum_{i=1}^{n}\rho_{\theta}^{\dagger}(W_i, \theta_n), \qquad (9.9)$$

where the equality follows from (3.19). By Assumptions B1(i), S2(iii), and S2(iv), $(\partial/\partial\theta)$ $E_{\gamma_0}\rho(W_i, \theta_0) = 0 \ \forall \gamma_0 \in \Gamma$. Under Assumption S3(iii), we have $E_{\gamma_0}\sup_{\theta\in\Theta}||\rho_{\theta}(W_i, \theta)|| < \infty$ because the parameter space of β is bounded. Hence, an exchange of ∂ and E yields $E_{\gamma_0}\rho_{\theta}(W_i, \theta_0) = 0$, which implies that $E_{\gamma_0}\rho_{\theta}^{\dagger}(W_i, \theta_0) = 0$ by (3.19). Because $E_{\gamma_n}\rho_{\theta}^{\dagger}(W_i, \theta_n) = 0$, $n^{-1/2}\sum_{i=1}^{n}\rho_{\theta}^{\dagger}(W_i, \theta_n) = \nu_n\rho_{\theta}^{\dagger}(\theta_n)$.

When β is a scalar, $\nu_n\rho_{\theta}^{\dagger}(\theta)$ converges weakly to a Gaussian process with covariance $V^{\dagger}(\theta_1, \theta_2; \gamma_0)$ on $\theta \in \Theta$ by the arguments given in the verification of Assumption C3. Hence, $\nu_n\rho_{\theta}^{\dagger}(\theta_n)$ converges in distribution to a normally distributed random variable with variance $V^{\dagger}(\theta_0, \theta_0; \gamma_0)$. Assumption D3(ii) holds by Assumption S3(v).

When β is a vector, the weak convergence above holds by replacing θ and Θ with θ^+ and Θ^+, respectively, using Assumption S3 (vector β) and the convergence in distribution holds because $\theta_n^+ \rightarrow \theta_0^+$ under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$. This completes the verification of Assumption D3. $\square$

Proof of Lemma 3.1. Assumption S3*(i) and (3.28) imply that $E_{\gamma_0}\varepsilon(W_i, \theta_0) = 0$.

Let $D(x,\theta)$ denote the matrix in the third line of (3.28) so that $\varepsilon(w,\theta) = \rho'(\cdot)D(x,\theta)$. To verify the Lipschitz condition in Assumption S3(i), we have

$$||\beta_0||^{-1}||E_{\gamma_0}\varepsilon(W_i,\psi_0,\pi)|| = ||\beta_0||^{-1}||E_{\gamma_0}\rho'(W_i,a(X_i,\beta_0)h(X_i,\pi))D(X_i,\theta)||$$
$$= ||E_{\gamma_0}\rho''(W_i,a(X_i,\beta_0)h(X_i,\widetilde{\pi}))\frac{a(X_i,\beta_0)}{||\beta_0||}h_\pi(X_i,\widetilde{\pi})'(\pi-\pi_0)D(X_i,\theta)||$$
$$\leq E_{\gamma_0}|\rho''(W_i,a(X_i,\beta_0)h(X_i,\widetilde{\pi}))| \cdot ||a_\beta(X_i,\widetilde{\beta})|| \cdot ||h_\pi(X_i,\widetilde{\pi})|| \cdot ||D(X_i,\theta)|| \cdot ||\pi-\pi_0||$$
$$\leq C||\pi-\pi_0||, \tag{9.10}$$

where the first equality holds by (3.28), the second equality follows from a mean-value expansion of $\rho'(W_i,a(X_i,\beta_0)h(X_i,\pi))$ in π around π_0 with $\widetilde{\pi}$ between π and π_0 and uses Assumption S3*(i), the first inequality follows from a mean-value expansion of $a(X_i,\beta_0)$ in β_0 around 0 with $\widetilde{\beta}$ between β_0 and 0, and the second inequality follows from Assumption S3*(ii) and $|a(X_i,\beta)|/||\beta|| \leq ||a_\beta(X_i,\widetilde{\beta})||$ with $\widetilde{\beta}$ between β and 0 by a mean-value expansion.

This completes the proof when β is a scalar.

When β is a vector, $||E_{\gamma_0}\varepsilon(W_i,\theta^+)|| \leq ||E_{\gamma_0}\varepsilon(W_i,\theta^+) - E_{\gamma_0}\varepsilon(W_i,||\beta_0||,\omega_0,\pi)|| + ||E_{\gamma_0}\varepsilon(W_i,||\beta_0||,\omega_0,\pi)||$, where $\theta^+ = (||\beta_0||,\omega,\pi)$. By (9.10), it is sufficient to show $||E_{\gamma_0}\varepsilon(W_i,||\beta_0||,\omega,\pi) - E_{\gamma_0}\varepsilon(W_i,||\beta_0||,\omega_0,\pi)|| \leq C||\beta_0||(||\omega-\omega_0||+||\pi-\pi_0||)$ for some $C < \infty$. According to (3.28), $\varepsilon(\cdot) = \rho'(\cdot)D(\cdot)$. For notational simplicity, we let $\rho'(\omega)$ and $D(\omega)$ be defined such that $\varepsilon(W_i,||\beta_0||,\omega,\pi) = \rho'(\omega)D(\omega)$. By the triangle inequality,

$$||E_{\gamma_0}\rho'(\omega)D(\omega) - E_{\gamma_0}\rho'(\omega_0)D(\omega_0)||$$
$$\leq ||E_{\gamma_0}(\rho'(\omega) - \rho'(\omega_0))D(\omega)|| + ||E_{\gamma_0}\rho'(\omega_0)(D(\omega) - D(\omega_0))||. \tag{9.11}$$

Because $\rho'(\cdot)$ does not involve $\iota(\beta)$, the reparameterization inside of $\rho'(\cdot)$ simply replaces all β with $||\beta||\omega$. Hence, any partial derivative of $\rho'(\omega)$ wrt ω is equivalent to the partial derivative wrt β in the original parametrization multiplied by $||\beta||$.

The first summand on the right-hand side of (9.11) satisfies

$$||E_{\gamma_0}(\rho'(\omega_0) - \rho'(\omega))D(\omega)||$$
$$\leq E_{\gamma_0}||\rho''(\cdot)a_\beta(\cdot)h(\cdot)|| \cdot ||D(\omega)|| \cdot ||\beta_0|| \cdot ||\omega-\omega_0||$$
$$\leq C||\beta_0|| \cdot ||\omega-\omega_0||, \tag{9.12}$$

where the first inequality holds by a mean-value expansion wrt ω and uses $\beta = ||\beta||\omega$ and the second inequality holds by Assumption S3*(ii). The arguments of the functions are suppressed for brevity.

To bound the second summand on the right-hand side of (9.11), note that $\rho'(\omega_0)$ differs from $\rho'(W_i, a(X_i, \beta_0)h(X_i, \pi_0))$ in Assumption S3*(i) by having π in the place of π_0. Using Assumption S3*(i), we have

$$
\begin{aligned}
&||E_{\gamma_0}\rho'(\omega_0)(D(\omega) - D(\omega_0))|| \\
&\leq E_{\gamma_0}||\rho''(\cdot)a(\cdot)h_\pi(\cdot)(D(\omega) - D(\omega_0))|| \cdot ||\pi - \pi_0|| \\
&= E_{\gamma_0}||\rho''(\cdot)a_\beta(\cdot)h_\pi(\cdot)(D(\omega) - D(\omega_0))|| \cdot ||\beta_0|| \cdot ||\pi - \pi_0|| \\
&\leq C||\beta_0|| \cdot ||\pi - \pi_0||,
\end{aligned}
\tag{9.13}
$$

where the first inequality follows from a mean-value expansion wrt π around π_0, the equality follows from a mean-value expansion wrt β around 0 and uses $a(x, 0) = 0$, and the second inequality follows from the moment conditions in Assumption S3*(ii). The desired result follows from (9.11)-(9.13). $\square$

Proof of Lemma 3.2. We verify Assumption C6 for the sample average estimator using Assumption C6** and Lemma 4.1 of AC1. Because β is a scalar, it remains to show Assumption C6**(ii). By Lemma 9.1,

$$
\Omega(\pi_1, \pi_2; \gamma_0) = \sum_{m=-\infty}^{\infty} Cov_{\gamma_0}(\rho_\psi(W_i, \psi_0, \pi_1), \rho_\psi(W_{i+m}, \psi_0, \pi_2)).
\tag{9.14}
$$

This implies that the covariance matrix $\Omega_G(\pi_1, \pi_2; \gamma_0)$ in Assumption C6**(ii) takes the form $\Omega_G(\pi_1, \pi_2; \gamma_0)$ in Assumption C6†(ii). Hence, Assumption C6**(ii) is implied by Assumption C6†(ii). $\square$

Proof of Lemma 9.2. For any given $\varepsilon > 0$, let $\delta^* = \min\{\delta, \frac{\varepsilon}{2C}\}$. Using the compactness of Θ, let $\{B(\theta_j, \delta^*) : j = 1, ..., J\}$ be a finite cover of Θ, where $B(\theta_j, \delta^*)$ denote a closed ball in Θ of radius $\delta^* \geq 0$ centered at θ_j. Because $q_n(\theta)$ converges to 0 $\forall \theta \in \Theta$, there exists N_j such that $||q_n(\theta_j)|| \leq \varepsilon/2$ for any $n \geq N_j$, for $j = 1, ..., J$. Let $N = \max_{j \leq J} N_j$.

Then, $\max_{j \leq J} ||q_n(\theta_j)|| \leq \varepsilon/2$ for any $n \geq N$. For any $n \geq N$,

$$
\begin{aligned}
\sup_{\theta \in \Theta} ||q_n(\theta)|| &\leq \max_{j \leq J} \left(\sup_{\theta' \in B(\theta_j, \delta^*)} ||q_n(\theta') - q_n(\theta_j)|| + ||q_n(\theta_j)|| \right) \\
&\leq \sup_{\theta \in \Theta} \sup_{\theta' \in B(\theta, \delta^*)} ||q_n(\theta') - q_n(\theta)|| + \max_{j \leq J} ||q_n(\theta_j)|| \leq C\delta^* + \varepsilon/2 \leq \varepsilon, \quad (9.15)
\end{aligned}
$$

where the first inequality uses the property of the finite cover and the triangle inequality, the second inequality is straightforward, the third inequality uses condition (ii) of Lemma 9.2 and $\delta^* \leq \delta$, and the fourth inequality follows from $\delta^* \leq \varepsilon/(2C)$. $\square$

Proof of Lemma 9.3. First, we establish the result of the lemma with $E_{\gamma_n} s(W_i, \theta)$ in place of $E_{\gamma_0} s(W_i, \theta)$. We use the uniform LLN given in Theorem 4 of Andrews (1992) employing Assumption TSE-1B with $q_t(z, \theta) = s(w, \theta)$. Now we verify Assumptions TSE-1B, DM, BD, and P-WLLN of Andrews (1992). Assumption TSE-1B(a) holds because $s(w, \theta)$ is continuous in θ and Θ is compact. Assumption TSE-1B(b) holds because $\{W_i : i \leq n\}$ is strictly stationary and $E_{\gamma_n} 1(W_i \in A) \to E_{\gamma_0} 1(W_i \in A)$ for all measurable sets $A \subset \mathcal{W}$ under $\{\gamma_n\} \in \Gamma(\gamma_0)$ by $\gamma_n \to \gamma_0$ and the weak convergence of W_i under γ_n to W_i under γ_0, which holds by the definition of the metric on Γ, see Section 2.1. Assumption DM holds by condition (iii) of Lemma 9.3. Assumption BD holds because Θ is compact. Assumption P-WLLN holds, i.e., $n^{-1} \sum_{i=1}^n s(W_i, \theta) - E_{\gamma_n} s(W_i, \theta) \to_p 0 \; \forall \theta \in \Theta$ under $\{\gamma_n\} \in \Gamma(\gamma_0)$, by the WLLN for dependent triangular arrays of strong mixing random variables in Example 4 of Andrews (1988) given that $\sup_{\gamma \in \Gamma} E_\gamma ||s(W_i, \theta)||^{1+\delta} < \infty$ for some $\delta > 0$. Theorem 4 of Andrews (1992) gives $\sup_{\theta \in \Theta} ||n^{-1} \sum_{i=1}^n s(W_i, \theta) - E_{\gamma_n} s(W_i, \theta)|| \to 0$ under $\{\gamma_n\} \in \Gamma(\gamma_0)$. Note that the same proof holds whether $\{W_i : i \geq 1\}$ are strong mixing or i.i.d. in Assumption S1.

To obtain the desired result, it remains to show $\sup_{\theta \in \Theta} ||E_{\gamma_n} s(W_i, \theta) - E_{\gamma_0} s(W_i, \theta)|| \to 0$ under $\{\gamma_n\} \in \Gamma(\gamma_0)$. The pointwise convergence holds for any $\theta \in \Theta$ by (i) the weak convergence induced by $\gamma_n \to \gamma_0$ and the definition of the metric on Γ and (ii) $E_\gamma \sup_{\theta \in \Theta} ||s(w, \theta)||^{1+\delta} \leq C \; \forall \gamma \in \Gamma$ for some $\delta > 0$. Because Θ is compact and pointwise convergence holds, we apply Lemma 9.2 with $q_n(\theta) = E_{\gamma_n} s(W_i, \theta) - E_{\gamma_0} s(W_i, \theta)$.

Condition (ii) of Lemma 9.2 holds because for any $\theta_1, \theta_2 \in \Theta$ with $||\theta_1 - \theta_2|| \leq \delta$,

$$
\begin{aligned}
||q_n(\theta_1) - q_n(\theta_2)|| &= ||E_{\gamma_n}\left(s(W_i, \theta_1) - s(W_i, \theta_2)\right) - E_{\gamma_0}\left(s(W_i, \theta_1) - s(W_i, \theta_2)\right)|| \\
&\leq E_{\gamma_n}||s(W_i, \theta_1) - s(W_i, \theta_2)|| + E_{\gamma_0}||s(W_i, \theta_1) - s(W_i, \theta_2)|| \\
&\leq (E_{\gamma_n} M_1(W_i) + E_{\gamma_0} M_1(W_i))\delta \leq C\delta, \qquad\qquad (9.16)
\end{aligned}
$$

where the first inequality follows from the triangle inequality and Jensen's inequality, the second inequality holds by condition (ii) of Lemma 9.3, and the third inequality holds by condition (iii) of Lemma 9.3.

The uniform continuity of $E_{\gamma_0} s(W_i, \theta)$ on Θ holds by the dominated convergence theorem and the compactness of Θ. This completes the proof. $\square$

Proof of Lemma 9.4. For the case that Assumption S1 holds with $\{W_i : i \geq 1\}$ being strong mixing, we show the stochastic equicontinuity (SE) of the empirical process $v_n s(\theta)$ using Theorem 3 of Hansen (1996), which is suitable for strong mixing arrays. When $s(w, \theta)$ is a vector, the SE of $\nu_n s(\theta)$ is implied by the SE of each entry of $\nu_n s(\theta)$. Hence, without loss of generality, we assume $s(w, \theta) \in R$ as in Hansen (1996). We now verify (11)-(13) in Assumption 4 of Hansen (1996). The condition in (11) holds provided $\alpha_m \leq Cm^{-A}$ for some $A > (1/p - 1/q)^{-1}$ and $d_\theta < p < q$. This is implied by Assumption S1. Conditions (12) and (13) hold because $E_\gamma \sup_{\theta \in \Theta} ||s(W_i, \theta)||^q \leq C$ and $E_\gamma M_1(W_i)^q \leq C \; \forall \gamma \in \Gamma$ and $\{W_i : i \geq 1\}$ is strictly stationary. Applying Theorem 3 of Hansen (1996) with $a = d_\theta$ and $\lambda = 1$ yields: for each $\varepsilon > 0$, there exists a $\delta_1 > 0$ such that

$$
\limsup_{n \to \infty} \left\| \sup_{\rho(\theta_1, \theta_2) < \delta_1} |\nu_n s(\theta_1) - \nu_n s(\theta_2)| \right\|_p < \varepsilon, \qquad\qquad (9.17)
$$

where $\rho(\theta_1, \theta_2) = \limsup_{n \to \infty} (E_{\gamma_n} |s(W_i, \theta_1) - s(W_i, \theta_2)|^q)^{1/q}$ and $||\cdot||_p$ is the L^p-norm for some $d_\theta < p < q$. By conditions (ii) and (iii) of Lemma 9.4, for each $\delta_1 > 0$, there exists $\delta > 0$, such that $||\theta_1 - \theta_2|| < \delta$ implies that $\rho(\theta_1, \theta_2) < \delta_1$. This, (9.17), and Markov's inequality yield the SE of $\nu_n s(\theta)$ over $\theta \in \Theta$.

For the case that Assumption S1 holds with $\{W_i : i \geq 1\}$ being i.i.d., the stochastic equicontinuity (SE) of the empirical process $v_n s(\theta)$ holds by Theorems 1 and 2 of Andrews (1994) using the type II class. For this result, the envelope function and the Lipschitz function must have $q = 2 + \delta$ moments finite, which holds by Assumption S1 and condition (iii) of the Lemma. $\square$

Proof of Lemma 9.5. First, we consider the case in which $\{W_i : i \geq 1\}$ are strong mixing. We show that $V_s(\gamma_0)$ exists and

$$\lim_{n\to\infty} Var_{\gamma_n}(n^{-1/2}\sum_{i=1}^n s(W_i)) = V_s(\gamma_0) \tag{9.18}$$

under $\{\gamma_n\} \in \Gamma(\gamma_0)$. By change of variables, we have

$$Var_{\gamma_n}(n^{-1/2}\sum_{i=1}^n s(W_i)) \tag{9.19}$$

$$= \sum_{m=-n+1}^{n-1} Cov_{\gamma_n}(s(W_i), s(W_{i+m})) - \sum_{m=-n+1}^{n-1} \frac{|m|}{n} Cov_{\gamma_n}(s(W_i), s(W_{i+m})).$$

By a standard strong mixing inequality, e.g., see Davidson (1994, p. 212), and Assumption S1,

$$|Cov_\gamma(s(W_i), s(W_{i+m}))| \leq C_1\alpha_m^{1-2/q} \leq C_2 m^{-A(1-2/q)}, \text{ where } A(1-2/q) > d_\theta \frac{q-2}{q-d_\theta} \geq 2 \tag{9.20}$$

using $d_\theta \geq 2$, for some $C_1, C_2 < \infty$ $\forall\gamma \in \Gamma$. Hence, $V_s(\gamma_0)$ exists and the second term on the rhs of (9.19) converges to 0.

It remains to show that the first term on the rhs of (9.19) converges to $V_s(\gamma_0)$. Because the metric on Γ induces weak convergence under $\gamma_n \to \gamma_0$ and $E_\gamma|s(W_i)|^{2+\delta} \leq C$ $\forall\gamma \in \Gamma$ for some $\delta > 0$, we have

$$Cov_{\gamma_n}(s(W_i), s(W_{i+m})) \to Cov_{\gamma_0}(s(W_i), s(W_{i+m})) \tag{9.21}$$

under $\gamma_n \in \Gamma(\gamma_0)$ for any $m \in R$ (e.g., see Theorem 2.20 and Example 2.21 of van der Vaart (1998)). By the DCT, (9.20), and (9.21), we have

$$\lim_{n\to\infty} \sum_{m=-n+1}^{n-1} Cov_{\gamma_n}(s(W_i), s(W_{i+m})) = \lim_{n\to\infty} \sum_{m=-\infty}^{\infty} 1(|m| < n-1)Cov_{\gamma_n}(s(W_i), s(W_{i+m}))$$

$$\to V_s(\gamma_0). \tag{9.22}$$

This and (9.19) yield (9.18).

When $V_s(\gamma_0) = 0$, we have $\lim_{n\to\infty} Var_{\gamma_n}(n^{-1/2}\sum_{i=1}^n s(W_i)) = 0$, which implies that $n^{-1/2}\sum_{i=1}^n(s(W_i) - E_{\gamma_n}s(W_i)) \to_d N(0, V_s(\gamma_0)) = 0$. When $V_s(\gamma_0) > 0$, we assume

60

$Var_{\gamma_n}(n^{-1/2} \sum_{i=1}^n s(W_i)) > 0 \ \forall n \geq 1$ without loss of generality. To show the triangular array CLT in Lemma 9.5, we apply Corollary 1 of de Jong (1997) with $\beta = \gamma = 0$, $c_{ni} = n^{-1/2}(\|\sum_{i=1}^n n^{-1/2}\Delta_n s(W_i)\|_2)^{-1}$, and $X_{ni} = n^{-1/2}\Delta_n s(W_i)(\|n^{-1/2}\sum_{i=1}^n \Delta_n s(W_i)\|_2)^{-1}$, where $\Delta_n s(W_i) = s(W_i) - E_{\gamma_n} s(W_i)$. Now we verify conditions (a)-(c) of Assumption 2 of de Jong (1997). Condition (a) holds automatically. Condition (b) holds because $c_{ni} > 0$ and $E_{\gamma_n}|X_{ni}/c_{ni}|^q = E_{\gamma_n}|\Delta_n s(W_i)|^q \leq C \ \forall \gamma_n \in \Gamma$ for some $C < \infty$. Condition (c) holds by taking $V_{ni} = X_{ni}$, $d_{ni} = 0$, and using Assumption S1 because $\alpha_m \leq Cm^{-A}$ and $A > q/(q-2)$. By Corollary 1 of de Jong (1997), we have $X_{ni} \to_d N(0,1)$. This and (9.18) lead to the desired result.

When Assumption S1 holds with $\{W_i : i \geq 1\}$ being i.i.d. under $\gamma_0 \in \Gamma$, a standard triangular array CLT gives the desired result because $2 + \delta$ moments of $s(W_i)$ are finite and uniformly bounded over $\gamma_0 \in \Gamma$ by Assumption S1. $\square$

10. Appendix B: Miscellaneous Results

This Appendix provides (i) the asymptotic size results for the robust QLR CS's, (ii) a sophisticated method for choosing κ for type 2 robust CS's, (iii) statements of Assumptions V1 and V2, which concern the estimator of the variance matrix of $\widehat{\theta}_n$, and (iv) an extension of the sufficient conditions for Assumption S3*(i) given in Section 3.1.4 for $\rho(w, \theta)$ functions of the form $\rho^*(w, a(x, \beta)h(x, \pi), \zeta)$. The extension is to the case where a parameter ζ appears.

10.1. Asymptotic Size of Robust QLR CS's

Here, we show that the LF and type 2 robust QLR CS's defined in the text of the paper have correct asymptotic size.

For the null-imposed (NI) critical values, we use the following notation: $H(v) = \{h = (b, \gamma_0) \in H : \|b\| < \infty, r(\theta_0) = v\}$, $V_r = \{v_0 : r(\theta_0) = v_0$ for some $h = (b, \gamma_0) \in H\}$, and the NI-LF critical value is $c_{QLR,1-\alpha}^{LF}(v) = \max\{\sup_{h \in H(v)} c_{QLR,1-\alpha}(h), \chi_{d_r,1-\alpha}^2\}$.

The asymptotic size results for the LF QLR CS's rely on the following df continuity conditions, which are not restrictive in most examples.

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

Assumption NI-LF. (i) The df of $QLR(h)$ is continuous at $c_{QLR,1-\alpha}(h)$ $\forall h \in H(v)$, $\forall v \in V_r$.

(ii) For some $v \in V_r$, $c_{QLR,1-\alpha}^{LF}(v) = \chi_{d_r,1-\alpha}^2$ or $c_{QLR,1-\alpha}^{LF}(v)$ is attained at some $h_{\max} \in H$.

For $h \in H$, define

$$
\widehat{c}_{QLR,1-\alpha}(h) \tag{10.23}
$$
$$
= \begin{cases} c_{QLR,1-\alpha}^{LF} + \Delta_1 & \text{if } A(h) \leq \kappa \\ \chi_{d_r,1-\alpha}^2 + \Delta_2 + [c_{QLR,1-\alpha}^{LF} + \Delta_1 - \chi_{d_r,1-\alpha}^2 - \Delta_2] \cdot s(A(h) - \kappa) & \text{if } A(h) > \kappa. \end{cases}
$$

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

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

The asymptotic size results for the type 2 robust QLR CS's rely on the following df continuity conditions, which are not restrictive in most examples.

Assumption Rob2. (i) $P(QLR(h) = \widehat{c}_{QLR,1-\alpha}(h)) = 0$ $\forall h \in H$.
(ii) If $\Delta_2 > 0$, $NRP(\Delta_1, \Delta_2; h^*) = \alpha$ for some point $h^* \in H$.

Assumption NI-Rob2. (i) $P(QLR(h) = \widehat{c}_{QLR,1-\alpha}(h, v)) = 0$ $\forall h \in H(v)$, $\forall v \in V_r$.
(ii) For some $v \in V_r$, $\Delta_2(v) = 0$ or $NRP(\Delta_1(v), \Delta_2(v); h^*) = \alpha$ for some point $h^* \in H(v)$.

The correct asymptotic size properties of LF and robust type 2 QLR CS's are established in the following Theorem.

Theorem 10.1. *Suppose Assumptions* S1-S4, B1, B2, C7, RQ1-RQ3, *and* RQ4(i) *hold. Then, the nominal* $1-\alpha$ *robust* QLR *CS has* $AsySz = 1-\alpha$ *when based on the following critical values*: (a) LF, (b) NI-LF, (c) *type 2 robust, and* (d) *type 2 NI robust, provided the following additional Assumptions hold, respectively*: (a) LF, (b) NI-LF, (c) C6, Rob2, V1, *and* V2, *and* (d) C6, NI-Rob2, V1, *and* V2.

Comments. 1. Plug-in versions of the robust QLR CS's considered in Theorem 10.1 also have asymptotically correct size under continuity assumptions on $c_{QLR,1-\alpha}(h)$ that typically are not restrictive. For brevity, we do not provide formal results here.

2. If part (ii) of Assumption LF, NI-LF, Rob2, or NI-Rob2 does not hold, then the corresponding part of Theorem 10.1 still holds, but with $AsySz \geq 1 - \alpha$.

3. The proof of Theorem 10.1 is as follows. Theorem 10.1 holds by Theorem 5.1(b) of AC1, plus the proof given in Appendix A that Assumptions B1, B2, and S1-S4 imply Assumptions A, B3, C1-C5, C8, and D1-D3 of AC1. The reason is that the results of the Theorem 10.1 and Theorem 5.1(b) of AC1 are the same, just the assumptions differ.

10.2. Choice of κ for Type 2 Robust Confidence Sets

For type 2 robust CS's, a sophisticated method for choosing κ is to minimize the average asymptotic FCP of the robust CS at a chosen set of points.[32] Of interest is a robust CS for $r(\theta)$. Let $\mathcal{K}$ denote the set of κ values from which one selects. First, for given $h \in H$, one chooses a null value $v_{H_0}(h)$ that differs from the true value $v_0 = r(\theta_0)$ (where $h = (b, \gamma_0)$ and $\gamma_0 = (\theta_0, \phi_0)$). The null value $v_{H_0}(h)$ is selected such that the robust CS based on a reasonable choice of κ, such as $\kappa = 1.5$ or 2, has a FCP that is in a range of interest, such as close to 0.50.[33,34] Second, one computes the FCP of the value $v_{H_0}(h)$ for each robust CS with $\kappa \in \mathcal{K}$. Third, one repeats steps one and two for each $h \in \mathcal{H}$, where $\mathcal{H}$ is a representative subset of H.[35] The optimal choice of κ is the value that minimizes over $\mathcal{K}$ the average FCP at $v_{H_0}(h)$ over $h \in \mathcal{H}$.

10.3. Assumptions V1 and V2

Here we state Assumptions V1 and V2, which concern estimators of the asymptotic variance matrix of $\widehat{\theta}_n$. These assumptions are used with the standard t tests and CS's, as well as with the robust t and QLR CS's, which employ variance matrix estimators in the identification category selection procedure. These assumptions are not very restrictive.

---

[32] For t and Wald CS's, asymptotic FCP's follow from the results in this paper, AC1, and/or Andrews and Cheng (2008). For QLR CI's, asymptotic FCP results only cover restrictions involving π, see Comment 5 to Theorem 4.2 of AC1. For other restrictions, one can use a large finite sample size when determining κ.

[33] For reasonable choices, the value of κ used to obtain $v_{H_0}(h)$ typically has very little effect on the final comparison across different values of κ. For example, this is true in the binary choice and STAR models considered here, and in the ARMA(1, 1) model considered in AC1.

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

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

Assumption V1 has two forms depending on whether β is a scalar or a vector.

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

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

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

Assumption V1 (vector $\boldsymbol{\beta}$). (i) $\widehat{J}_n = \widehat{J}_n(\widehat{\theta}_n^+)$ and $\widehat{V}_n = \widehat{V}_n(\widehat{\theta}_n^+)$ for some (stochastic) $d_\theta \times d_\theta$ matrix-valued functions $\widehat{J}_n(\theta^+)$ and $\widehat{V}_n(\theta^+)$ on Θ^+ that satisfy $\sup_{\theta^+ \in \Theta^+} ||\widehat{J}_n(\theta^+) - J(\theta^+; \gamma_0)|| \to_p 0$ and $\sup_{\theta^+ \in \Theta^+} ||\widehat{V}_n(\theta^+) - V(\theta^+; \gamma_0)|| \to_p 0$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$.[36]

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

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

(iv) $P(\tau_\beta(\pi^*(\gamma_0, b); \gamma_0, b) = 0) = 0$ $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$ and $\forall b$ with $||b|| < \infty$.[37]

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

10.4. Adjustment for ζ

Here we provide sufficient conditions for Assumption S3*(i) when $\rho(w, \theta) = \rho^*(w, a(x, \beta)h(x, \pi), \zeta)$, as in (3.17), and a parameter ζ appears. (Section 3.1.4 provides analogous results when no parameter ζ appears.) For simplicity, we assume $a(x, \beta)$ and $h(x, \pi)$ are both scalars. Let $\rho'(\cdot)$ and $\rho''(\cdot)$ denote the first and second order partial derivatives of $\rho^*(w, a(x, \beta)h(x, \pi), \zeta)$ wrt $a(x, \beta)h(x, \pi)$. Let $\rho_\zeta(\cdot)$ and $\rho_{\zeta\zeta}(\cdot)$ denote the first and second order partial derivatives of $\rho^*(w, a(x, \beta)h(x, \pi), \zeta)$ wrt ζ. Let $\rho_{12}(\cdot) \in R^{d_\zeta}$ denote the partial derivative of $\rho'(\cdot)$ wrt ζ. The partial derivatives in (3.27) are the same when

---

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

[37] Assumption V1 (vector β) differs from Assumption V1 (scalar β) because in the vector β case Assumption V1(ii) (scalar β) (i.e., continuity in θ) often fails, but Assumption V1(ii) (vector β) (i.e., continuity in θ^+) holds.

ζ appears in $\rho(w, \theta)$. The partial derivatives wrt ζ are

$$\rho_\zeta(w, \theta) = \rho_\zeta(\cdot) \in R^{d_\zeta \times 1}, \; \rho_{\zeta\zeta}(w, \theta) = \rho_{\zeta\zeta}(\cdot) \in R^{d_\zeta \times d_\zeta},$$
$$\rho_{\beta\zeta}(w, \theta) = a_\beta(x, \beta)h(x, \pi)\rho_{12}(\cdot) \in R^{d_\beta \times d_\zeta},$$
$$\rho_{\pi\zeta}(w, \theta) = a(x, \beta)h_\pi(x, \pi)\rho_{12}(\cdot) \in R^{d_\pi \times d_\zeta}. \tag{10.1}$$

In this case, we define

$$\rho_\theta^\dagger(w, \theta) = \rho'(\cdot)a^\dagger(x, \theta) + \overline{\rho}_\zeta(\cdot), \; \rho_{\theta\theta}^\dagger(w, \theta) = \rho''(\cdot)a^\dagger(x, \theta)a^\dagger(x, \theta)' + \overline{\rho}_{\zeta\zeta}(\cdot), \; \text{where}$$
$$a^\dagger(x, \theta) = \left(a_\beta(x, \beta)'h(x, \pi), 0_{d_\zeta}, \frac{a(x, \beta)}{\iota(\beta)}h_\pi(x, \pi)'\right)',$$
$$\overline{\rho}_\zeta(\cdot) = (0_{d_\beta}, \rho_\zeta(\cdot), 0_{d_\pi})',$$
$$\overline{\rho}_{\zeta\zeta}(\cdot) = \begin{bmatrix} 0_{d_\beta \times d_\beta} & a_\beta(x, \beta)h(x, \pi)\rho_{12}(\cdot) & 0_{d_\beta \times d_\pi} \\ (a_\beta(x, \beta)h(x, \pi)\rho_{12}(\cdot))' & \rho_{\zeta\zeta}(\cdot) & \left(\frac{a(x, \beta)}{\iota(\beta)}h_\pi(x, \pi)\rho_{12}(\cdot)\right)' \\ 0_{d_\pi \times d_\beta} & \frac{a(x, \beta)}{\iota(\beta)}h_\pi(x, \pi)\rho_{12}(\cdot) & 0_{d_\pi \times d_\pi} \end{bmatrix},$$
$$\varepsilon(w, \theta) = \rho'(\cdot)\begin{bmatrix} a_{\beta\beta}(x, \beta)h(x, \pi) & 0_{d_\beta \times d_\zeta} & a_\beta(x, \beta)h_\pi(x, \pi)' \\ 0_{d_\zeta \times d_\beta} & 0_{d_\zeta \times d_\zeta} & 0_{d_\zeta \times d_\pi} \\ h_\pi(x, \pi)a_\beta(x, \beta)' & 0_{d_\pi \times d_\zeta} & \frac{a(x, \beta)}{\iota(\beta)}h_{\pi\pi}(x, \pi) \end{bmatrix}. \tag{10.2}$$

Comparing the definition of $\varepsilon(w, \theta)$ in (10.2) with that in (3.28), it is clear that, if $\rho(w, \theta)$ takes the form in (3.17) and a parameter ζ appears, then Assumption S3* still implies Assumption S3(i) provided $\rho'(\cdot)$ and $\rho''(\cdot)$ in Assumption S3* are adjusted to include ζ, evaluated at ζ_0.

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., 1992. Generic uniform convergence. Econometric Theory 8, 241-257.

Andrews, D. W. K., 1994. Empirical process methods in econometrics. In: Engle, R. F., McFadden, D. (Eds), Handbook of Econometrics, Vol. IV. North-Holland, Amsterdam.

Andrews, D. W. K., Cheng, X., 2007a. Estimation and inference with weak, semi-strong, and strong identification. Cowles Foundation Discussion Paper No. 1773R, Yale University.

Andrews, D. W. K., Cheng, X., 2007b. Supplemental appendices for "Maximum likelihood estimation and uniform inference with sporadic identification failure." Addendum to Cowles Foundation Discussion Paper No. 1824, Yale University.

Andrews, D. W. K., Cheng, X., 2007c. Supplemental material for "Estimation and inference with weak, semi-strong, and strong identification." Cowles Foundation Discussion Paper No. 1773R, Yale University.

Andrews, D. W. K., Cheng, X., 2008. GMM estimation and inference with weak, semi-strong, and strong identification. Unpublished manuscript, Cowles Foundation, Yale University.

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.

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

Davidson, J., 1994. Stochastic Limit Theory. Oxford University Press, Oxford.

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.

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.

Kleibergen, F., Mavroeidis, S., 2009. Inference on subsets of parameters in GMM without assuming identification. Unpublished manuscript, Department of Economics, Brown University.

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., 2011. Nonlinear cointegrating regression under weak identification. Econometric Theory, forthcoming.

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.

van der Vaart, A. W., 1998. Asymptotic Statistics. Cambridge University Press, Cambridge, UK.

Supplemental Appendices

for

Maximum Likelihood Estimation
and Uniform Inference
with Sporadic Identification Failure

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

Xu Cheng
Department of Economics
University of Pennsylvania

11. Outline

This Supplement includes three Supplemental Appendices (denoted C, D, and E) to the paper "Maximum Likelihood Estimation and Uniform Inference with Sporadic Identification Failure." Supplemental Appendix C provides additional numerical results to those provided in the main paper for both the smooth transition autoregressive (STAR) model and the nonlinear binary choice model. Supplemental Appendix D verifies Assumptions S1-S4, B1, B2, C6, V1, and V2 for the nonlinear binary choice model. Supplemental Appendix E does likewise for the STAR model.

We let AC1 abbreviate the paper Andrews and Cheng (2007) "Estimation and Inference with Weak, Semi-strong, and Strong Identification."

12. Supplemental Appendix C: Numerical Results

Table S-1 compares the finite-sample ($n = 500$) coverage probabilities of the null-imposed robust CI's for β in the STAR model with true and estimated values of ζ. (See the end of the STAR-model numerical-results section in the main paper for further discussion.)

Figures S-1 and S-2 report asymptotic and finite-sample ($n = 500$) densities of the estimators for β and π in the STAR model when $\pi_0 = -3.0$. Figures S-3 to S-6 report asymptotic and finite-sample (n=500) densities of the t and QLR statistics for β and π in the STAR model when $\pi_0 = -1.5$. Figures S-7 and S-8 report CP's of nominal 0.95 standard and robust $|t|$ and QLR CI's for β and π in the STAR model when $\pi_0 = -3.0$.

Figures S-9 to S-16 are analogous to Figures S-1 to S-8 but for the binary choice model. The true values of π considered are $\pi_0 = 1.5$ and $\pi_0 = 2.0$.

Table S-1. Finite-Sample Coverage Probabilities of Null-Imposed Robust CI's for β in the STAR Model with True and Estimated Values of ζ, $n = 500$, $\pi_0 = -1.5$[38]

b	0	1	2	3	4	5	6	7	8	9	10	11	12
t_ζ_0	0.939	0.950	0.946	0.947	0.948	0.947	0.944	0.946	0.949	0.949	0.950	0.947	0.956
$t\_\widehat{\zeta}$	0.936	0.951	0.946	0.947	0.947	0.947	0.949	0.944	0.947	0.947	0.947	0.947	0.957
QLR_ζ_0	0.923	0.932	0.930	0.925	0.923	0.924	0.916	0.921	0.923	0.926	0.929	0.932	0.933
$QLR\_\widehat{\zeta}$	0.920	0.935	0.927	0.924	0.926	0.919	0.915	0.908	0.915	0.922	0.925	0.924	0.929

---

[38] The simulation is conducted with the null value of b and the true value of π imposed so that the
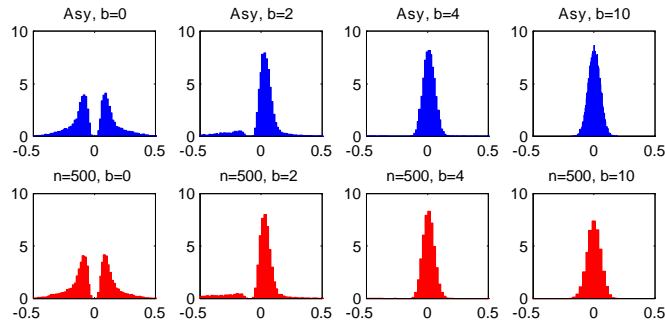
Figure S-1. Asymptotic and Finite-Sample ($n = 500$) Densities of the Estimator of β in the STAR Model when $\pi_0 = -3.0$.
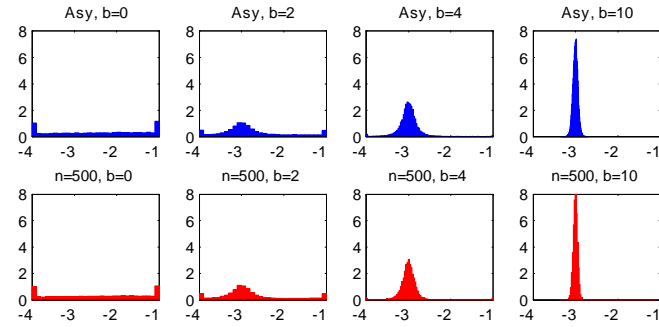

Figure S-2. Asymptotic and Finite-Sample ($n = 500$) Densities of the Estimator of π in the STAR Model when $\pi_0 = -3.0$.
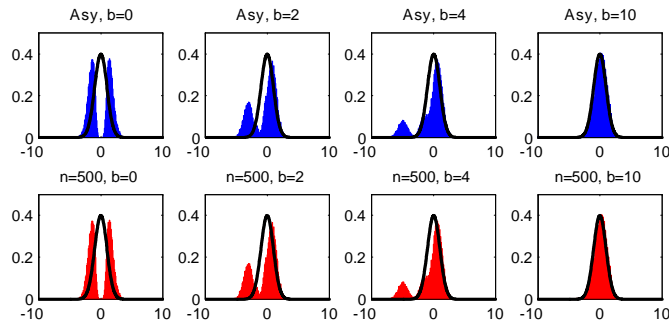

Figure S-3. Asymptotic and Finite-Sample ($n = 500$) Densities of the t Statistic for β in the STAR Model when $\pi_0 = -1.5$ and the Standard Normal Density (Black Line).
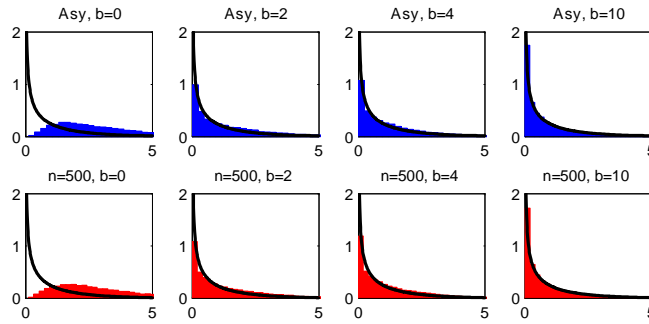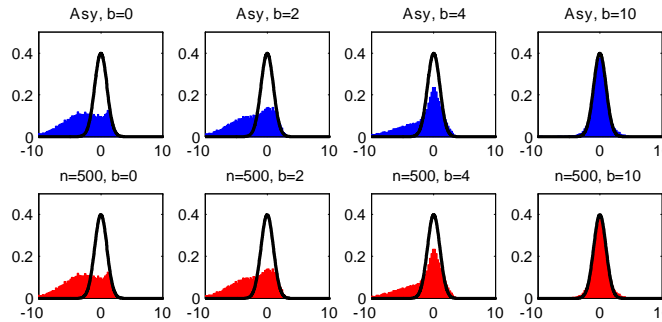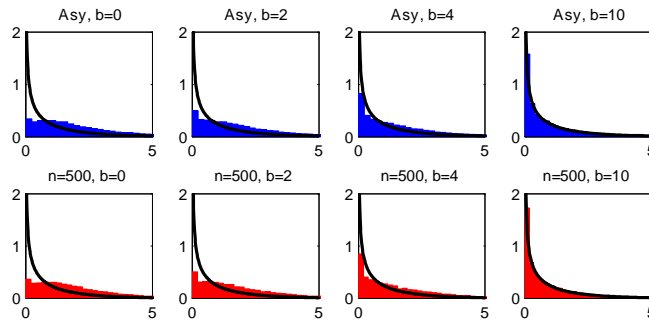
Figure S-4. Asymptotic and Finite-Sample (n=500) Densities of the QLR Statistic for β in the STAR Model when $\pi_0 = -1.5$ and the χ_1^2 Density (Black Line).
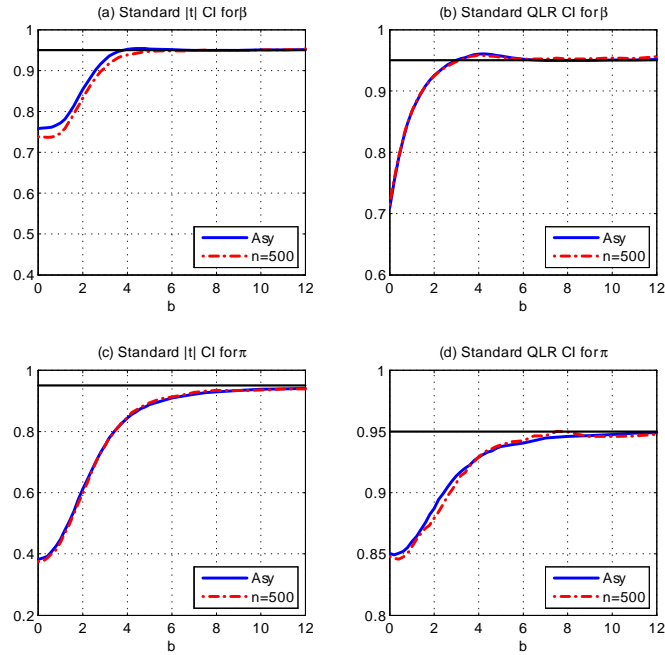

Figure S-5. Asymptotic and Finite-Sample ($n = 500$) Densities of the t Statistic for π in the STAR Model when $\pi_0 = -1.5$ and the Standard Normal Density (Black Line).

Figure S-6. Asymptotic and Finite-Sample (n=500) Densities of the QLR Statistic for π in the STAR Model when $\pi_0 = -1.5$ and the χ_1^2 Density (Black Line).

3

Figure S-7. Coverage Probabilities of Standard $|t|$ and QLR CI's for β and π in the STAR Model when $\pi_0 = -3.0$.
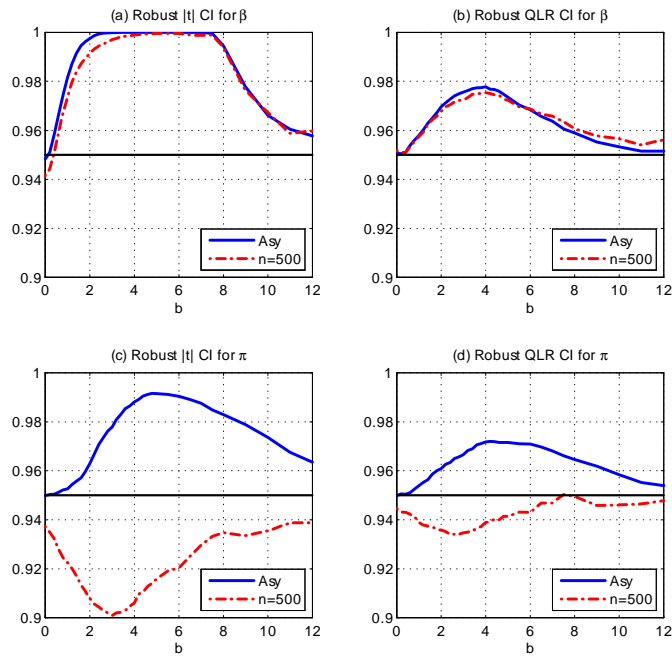
Figure S-8. Coverage Probabilities of Robust $|t|$ and QLR CI's for β and π in the STAR Model when $\pi_0 = -3.0$, $\kappa = 2.5$, $D = 1$, and $s(x) = \exp(-2x)$.
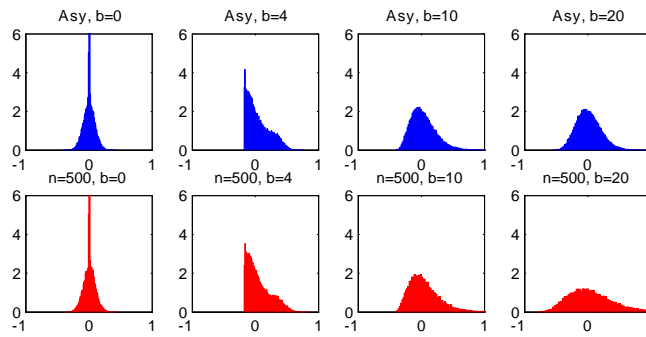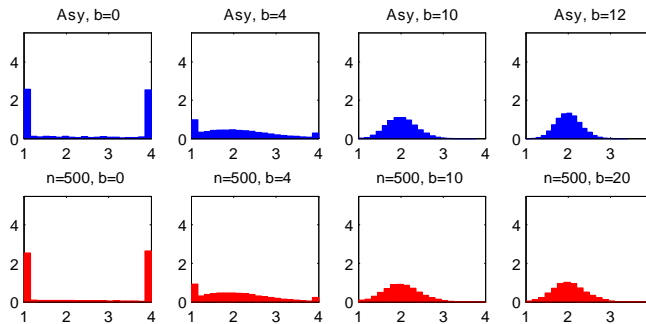
4

Figure S-9. Asymptotic and Finite-Sample (n=500) Densities of the Estimator of β in the Binary Choice Model when $\pi_0 = 2.0$.

Figure S-10. Asymptotic and Finite-Sample (n=500) Densities of the Estimator of π in the Binary Choice Model when $\pi_0 = 2.0$.
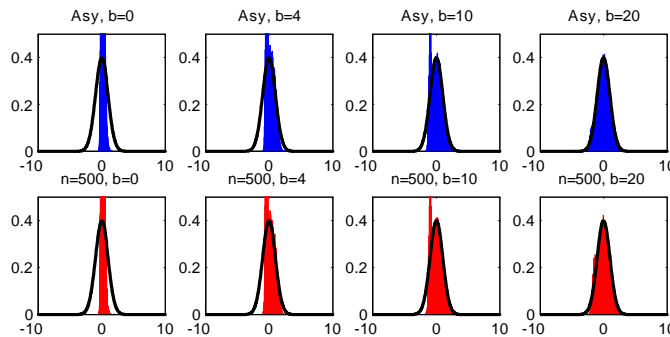

Figure S-11. Asymptotic and Finite-Sample (n=500) Densities of the t Statistic for π in the Binary Choice Model when $\pi_0 = 1.5$ and the Standard Normal Density (Black Line).
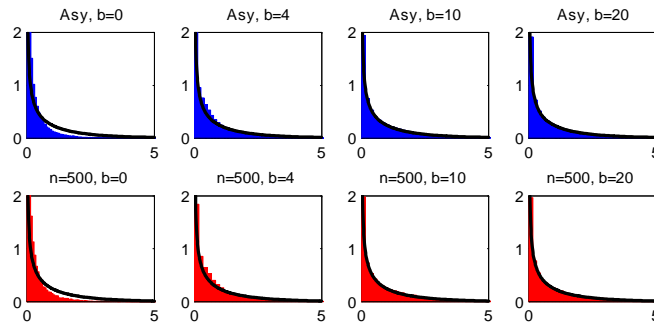
Figure S-12. Asymptotic and Finite-Sample (n=500) Densities of the QLR Statistic for π in the Binary Choice Model when $\pi_0 = 1.5$ and the χ_1^2 Density (Black Line).
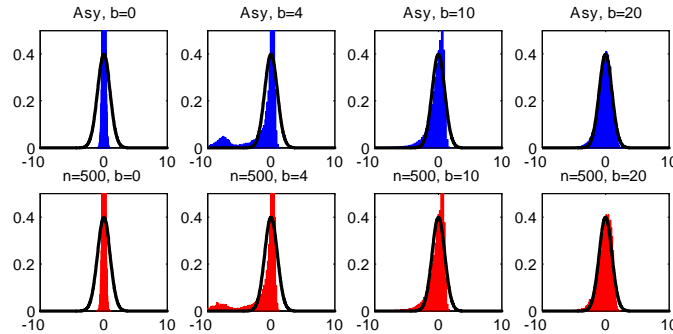

Figure S-13. Asymptotic and Finite-Sample (n=500) Densities of the t Statistic for β in the Binary Choice Model when $\pi_0 = 1.5$ and the Standard Normal Density (Black Line).
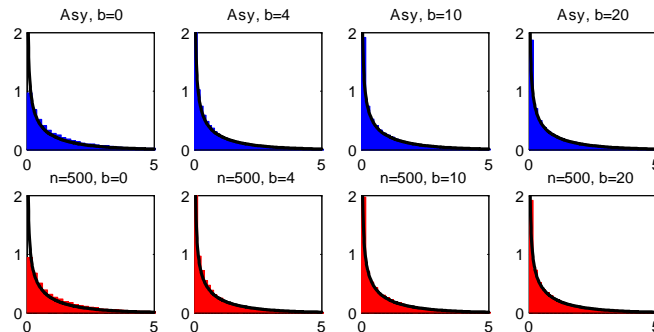

Figure S-14. Asymptotic and Finite-Sample (n=500) Densities of the QLR Statistic for β in the Binary Choice Model when $\pi_0 = 1.5$ and the χ_1^2 Density (Black Line).
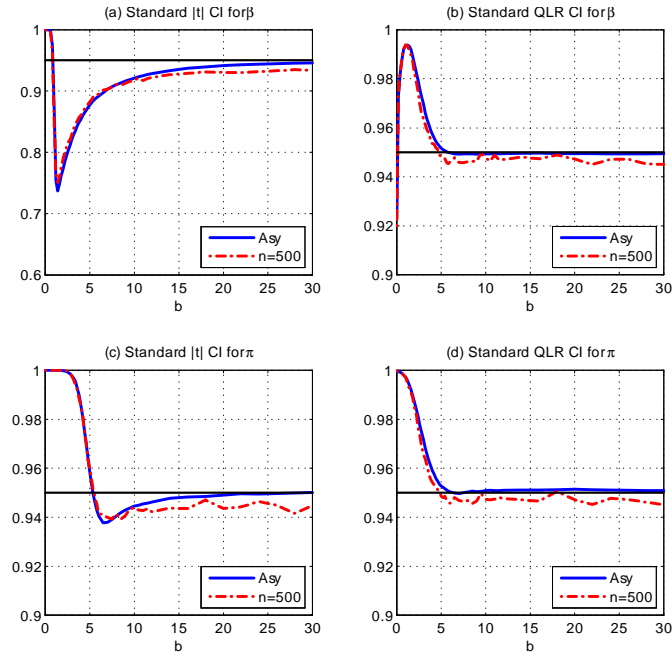
Figure S-15. Coverage Probabilities of Standard $|t|$ and QLR CI's for β and π in the Binary Choice Model when $\pi_0 = 2.0$.
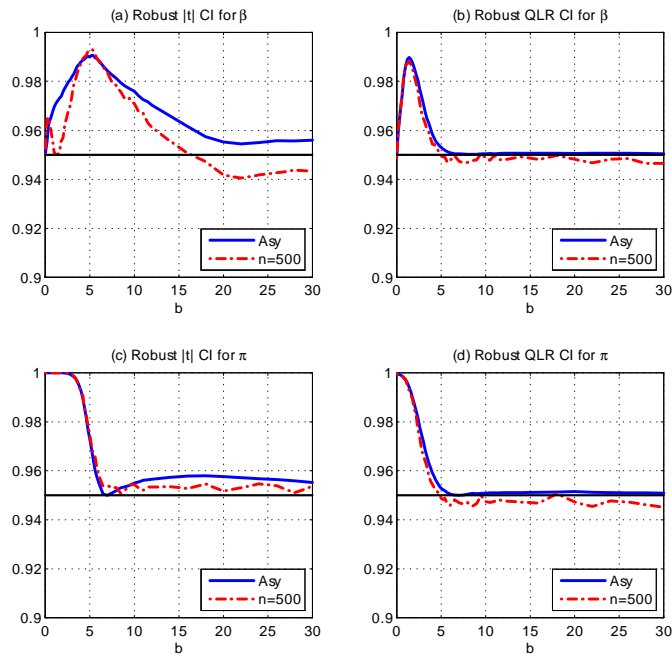

Figure S-16. Coverage Probabilities of Robust $|t|$ and QLR CI's for β and π in the Binary Choice Model when $\pi_0 = 2.0$, $\kappa = 1.5$, $D = 1$, and $s(x) = \exp(-x/2)$.

13. Supplemental Appendix D: Nonlinear Binary Choice Model, Verification of Assumptions

We start by deriving the formulae for the key quantities specified in (3.33). Next, we verify Assumptions S1-S4. Then, we verify Assumptions B1 and B2. Finally, we verify the remaining Assumptions C6, V1, and V2. (Note that Assumption C7 is verified in Section 3.5.)

13.1. Derivation of Key Quantities

Here we calculate the key quantities $\Omega(\pi_1, \pi_2; \gamma_0)$, $H(\pi; \gamma_0)$, $J(\gamma_0)$, and $V(\gamma_0)$ that are specified in (3.33).

By (2.4),

$$E_{\gamma_0}(Y_i - L_i(\theta_0)|X_i, Z_i) = 0 \text{ a.s. and}$$
$$E_{\gamma_0}((Y_i - L_i(\theta_0))^2|X_i, Z_i) = L_i(\theta_0)(1 - L_i(\theta_0)) \text{ a.s.} \tag{13.1}$$

For γ_0 with $\beta_0 = 0$, we have $g_i(\psi_0, \pi) = g_i(\theta_0)$, $L_i(\psi_0, \pi) = L_i(\theta_0)$, $L_i'(\psi_0, \pi) = L_i'(\theta_0)$, and $w_{j,i}(\psi_0, \pi) = w_{j,i}(\theta_0)$ for $j = 1, 2$, $\forall \pi \in \Pi$. In consequence,

$$\Omega(\pi_1, \pi_2; \gamma_0) = S_\psi V^\dagger((\psi_0, \pi_1), (\psi_0, \pi_2); \gamma_0) S_\psi'$$
$$= E_{\gamma_0} \frac{L_i'^2(\theta_0)}{L_i(\theta_0)(1 - L_i(\theta_0))} d_{\psi,i}(\pi_1) d_{\psi,i}(\pi_2)', \tag{13.2}$$

where $S_\psi = [I_{d_\psi} : 0_{d_\psi \times d_\pi}]$, the first equality holds by Lemma 9.1, and the second equality holds by independence across i of $\{W_i : i \leq n\}$ and (13.1).

Now, we have

$$\rho_{\psi\psi}(W_i, \psi_0, \pi) = [w_{1,i}^2(\theta_0)(Y_i - L_i(\theta_0))^2 + w_{2,i}(\theta_0)(Y_i - L_i(\theta_0))] d_{\psi,i}(\pi) d_{\psi,i}(\pi)' \text{ and}$$
$$H(\pi; \gamma_0) = E_{\gamma_0} \rho_{\psi\psi}(W_i, \psi_0, \pi) = E_{\gamma_0} \frac{L_i'^2(\theta_0)}{L_i(\theta_0)(1 - L_i(\theta_0))} d_{\psi,i}(\pi) d_{\psi,i}(\pi)', \tag{13.3}$$

where the first equality uses (3.22), the second equality holds by Lemma 9.1, and the

third equality uses (13.1).

In addition, we have

$$
\begin{aligned}
V(\gamma_0) &= V^\dagger(\theta_0, \theta_0; \gamma_0) = Var_{\gamma_0}(\rho_\theta^\dagger(W_i, \theta_0)) \\
&= E_{\gamma_0} w_{1,i}^2(\theta_0)(Y_i - L_i(\theta_0))^2 d_i(\pi_0) d_i(\pi_0)' \\
&= E_{\gamma_0} \frac{L_i'^2(\theta_0)}{L_i(\theta_0)(1 - L_i(\theta_0))} d_i(\pi_0) d_i(\pi_0)',
\end{aligned}
\tag{13.4}
$$

where the first equality holds by (3.20) and the second equality holds by independence across i of $\{W_i : i \leq n\}$ and (13.1).

Next, we have

$$
\begin{aligned}
J(\gamma_0) &= E_{\gamma_0} \rho_{\theta\theta}^\dagger(W_i, \theta_0) \\
&= E_{\gamma_0}[w_{1,i}^2(\theta_0)(Y_i - L_i(\theta_0))^2 + w_{2,i}(\theta_0)(Y_i - L_i(\theta_0))] d_i(\pi) d_i(\pi)' \\
&= E_{\gamma_0} \frac{L_i'^2(\theta_0)}{L_i(\theta_0)(1 - L_i(\theta_0))} d_i(\pi) d_i(\pi)',
\end{aligned}
\tag{13.5}
$$

where the first equality holds by Lemma 9.1, the second equality holds using (3.23), and the third equality holds by (13.1).

The matrix $K(\pi; \gamma_0)$ is derived in Section 13.6 below.

13.2. Verification of Assumptions S1 and S2

Given that $\{W_i : i \geq 1\}$ are i.i.d. under $\gamma_0 \; \forall \gamma_0 \in \Gamma$, Assumption S1 holds with $q = 2 + \delta$ for $\delta > 0$.

Assumption S2(i) holds with

$$
\rho(W_i, \theta) = -[Y_i \log L_i(\theta) + (1 - Y_i) \log(1 - L_i(\theta))].
\tag{13.6}
$$

When $\beta = 0$, $L_i(\theta) = L(\beta h(X_i, \pi) + Z_i'\zeta)$ does not depend on π and, hence, $\rho(W_i, \theta)$ does not depend on π. This verifies Assumption S2(ii).

To verify Assumptions S2(iii) and S2(iv), we have

$$
E_{\gamma_0}(\rho(W_i, \theta)|X_i, Z_i) = -[L_i(\theta_0) \log L_i(\theta) + (1 - L_i(\theta_0)) \log(1 - L_i(\theta))]
\tag{13.7}
$$

because $E_{\gamma_0}(Y_i|X_i, Z_i) = L_i(\theta_0)$ by (2.4). Now we view $E_{\gamma_0}(\rho(W_i, \theta)|X_i, Z_i)$ as a function

of $L_i(\theta)$. The first- and second-order derivatives of $E_{\gamma_0}(\rho(W_i, \theta)|X_i, Z_i)$ wrt $L_i(\theta)$ are

$$\frac{\partial}{\partial L_i(\theta)} E_{\gamma_0}(\rho(W_i, \theta)|X_i, Z_i) = \frac{L_i(\theta) - L_i(\theta_0)}{L_i(\theta)(1 - L_i(\theta))} \text{ and}$$

$$\frac{\partial^2}{\partial L_i^2(\theta)} E_{\gamma_0}(\rho(W_i, \theta)|X_i, Z_i) = \frac{L_i(\theta_0) + L_i^2(\theta) - 2L_i(\theta)L_i(\theta_0)}{L_i^2(\theta)(1 - L_i(\theta))^2}, \tag{13.8}$$

see (13.45) below. The second-order derivative is positive for all $\theta \in \Theta$ because its numerator is greater than $(L_i(\theta_0) - L_i(\theta))^2 \geq 0$. When $L_i(\theta) = L_i(\theta_0)$, the first-order derivative is 0. Hence, $E_{\gamma_0}(\rho(W_i, \theta)|X_i, Z_i)$, viewed as a function of $L_i(\theta)$, has a unique global minima at $L_i(\theta_0)$. Because $L'(u) > 0$, $E_{\gamma_0}\rho(W_i, \theta)$ is minimized at θ if and only if $P_{\gamma_0}(g_i(\theta) = g_i(\theta_0)) = 1$.

When $\beta_0 = 0$, $g_i(\theta) - g_i(\theta_0) = \beta h(X_i, \pi) + (\zeta - \zeta_0)'Z_i$. Because $P_{\gamma_0}(a'(h(X_i, \pi), Z_i) = 0) < 1$ for all $a \in R^{d_\zeta+1}$ with $a \neq 0$ (by the definition of Φ^* in (3.32)), $P_{\gamma_0}(g_i(\theta) - g_i(\theta_0) = 0) = 1$ if and only if $\beta = 0$ and $\zeta = \zeta_0$. This implies Assumption S2(iii).

When $\beta_0 \neq 0$, $g_i(\theta) - g_i(\theta_0) = \beta h(X_i, \pi) - \beta_0 h(X_i, \pi_0) + (\zeta - \zeta_0)'Z_i$. Because $P_{\gamma_0}(a'(h(X_i, \pi), h(X_i, \pi_0), Z_i) = 0) < 1$ for all $a \in R^{d_\zeta+2}$ with $a \neq 0$ and $\pi \neq \pi_0$, $P_{\gamma_0}(g_i(\theta) - g_i(\theta_0) = 0) < 1$ when $\pi \neq \pi_0$. When $\pi = \pi_0$, $g_i(\theta) - g_i(\theta_0) = (\beta - \beta_0)h(X_i, \pi v) + (\zeta - \zeta_0)'Z_i$. Because $P_{\gamma_0}(a'(h(X_i, \pi), Z_i) = 0) < 1$ for all $a \in R^{d_\zeta+1}$ with $a \neq 0$, $P_{\gamma_0}(g_i(\theta) - g_i(\theta_0) = 0) = 1$ if and only if $\zeta = \zeta_0$, $\beta = \beta_0$, and $\pi = \pi_0$. This verifies Assumption S2(iv).

Assumption S2(v) holds because $\Psi(\pi)$ does not depend on π and Ψ, Π, and Θ are all compact. Assumption S2(vi) holds automatically because $\Psi(\pi)$ does not depend on π.

13.3. Verification of Assumption S3(ii)

Assumption S3(i) is verified in the text of the paper. Here we verify Assumption S3(ii). We use the following generic results in the calculations below. Let $A = aa'$, where $a = (a_1', ..., a_p')' \in R^{d_a}$ and $a_1, ..., a_p$ are vectors (possibly of different dimensions). Then,

$$||A|| = \left(\sum_{j=1}^p \sum_{k=1}^p ||a_j a_k'||^2 \right)^{1/2} = \sum_{j=1}^p ||a_j||^2, \tag{13.9}$$

where the first equality holds by the definition of $||A||$ and the second equality holds because $||ab'|| = ||a|| \cdot ||b||$ for vectors a and b. Similarly, let $A^* = a^* a^{*'}$, where $a_1^*, ..., a_p^*$

10

are sub-vectors of a^* that are conformable with $a_1, ..., a_p$. Then,

$$||A - A^*|| = ||aa' - a^*a^{*\prime}|| \le ||a(a - a^*)'|| + ||(a - a^*)a^{*\prime}||$$

$$= (||a|| + ||a^*||)||a - a^*|| \le \sum_{j=1}^{p}(||a_j|| + ||a_j^*||)\sum_{k=1}^{p}||a_k - a_k^*||, \qquad (13.10)$$

where the first inequality holds by triangle inequality, the second equality holds because $||ab'|| = ||a|| \cdot ||b||$, and the last inequality holds because $(x^2 + y^2)^{1/2} \le x + y$ for non-negative scalars x and y.

Define $v_{1,i}(\theta) = w_{1,i}(\theta)(Y_i - L_i(\theta))$, $v_{2,i}(\theta) = w_{2,i}(\theta)(Y_i - L_i(\theta))$, and $\overline{\beta} = \max\{b_1, b_2\}$. Below, let $\theta_1, \theta_2 \in \Theta$ with $||\theta_1 - \theta_2|| \le \delta$ for $\delta > 0$.

By the triangle inequality, we have

$$||\rho_{\psi\psi}(W_i, \theta_1) - \rho_{\psi\psi}(W_i, \theta_2)||$$
$$\le \left(||v_{1,i}^2(\theta_1) - v_{1,i}^2(\theta_2)|| + ||v_{2,i}(\theta_1) - v_{2,i}(\theta_2)||\right) \cdot ||d_{\psi,i}(\pi_1)d_{\psi,i}(\pi_1)'||$$
$$+ \left(||v_{1,i}^2(\theta_2)|| + ||v_{2,i}(\theta_2)||\right) \cdot ||d_{\psi,i}(\pi_1)d_{\psi,i}(\pi_1)' - d_{\psi,i}(\pi_2)d_{\psi,i}(\pi_2)'||. \quad (13.11)$$

Note that

$$||v_{1,i}^2(\theta_1) - v_{1,i}^2(\theta_2)|| = ||v_{1,i}(\theta_1) - v_{1,i}(\theta_2)|| \cdot ||v_{1,i}(\theta_1) + v_{1,i}(\theta_2)||, \text{ where}$$
$$||v_{1,i}(\theta_1) - v_{1,i}(\theta_2)|| \le ||w_{1,i}(\theta_1) - w_{1,i}(\theta_2)|| \cdot ||Y_i - L_i(\theta_1)|| + ||w_{1,i}(\theta_2)|| \cdot ||L_i(\theta_1) - L_i(\theta_2)||$$
$$\le \left(M_1(W_i) + \overline{w}_{1,i}\overline{L}_i'(\overline{h}_i + ||Z_i|| + \overline{\beta} \cdot \overline{h}_{\pi,i})\right)\delta, \text{ and}$$
$$||v_{1,i}(\theta_1) + v_{1,i}(\theta_2)|| \le 2\overline{w}_{1,i}, \qquad (13.12)$$

where the first inequality follows from the triangle inequality, the second inequality holds by (i) $||w_{1,i}(\theta_1) - w_{1,i}(\theta_2)|| \le M_1(W_i)\delta$, (ii) $||Y_i - L_i(\theta)|| \le 1$, and (iii) $||L_i(\theta_1) - L_i(\theta_2)|| \le \overline{L}_i'(\overline{h}_i + ||Z_i|| + \overline{\beta} \cdot \overline{h}_{\pi,i})\delta$ by a mean-value expansion of $L_i(\theta) = L(g_i(\theta))$ wrt θ, and the third inequality follows from the triangle inequality and $||Y_i - L_i(\theta)|| \le 1$. Similarly,

$$||v_{2,i}(\theta_1) - v_{2,i}(\theta_2)|| \le \left(M_2(W_i) + \overline{w}_{2,i}\overline{L}_i'(\overline{h}_i + ||Z_i|| + \overline{\beta} \cdot \overline{h}_{\pi,i})\right)\delta,$$
$$||v_{1,i}^2(\theta_2)|| \le \overline{w}_{1,i}^2, \text{ and } ||v_{2,i}(\theta_2)|| \le \overline{w}_{2,i}. \qquad (13.13)$$

11

Applying the inequality in (13.9) with $a = d_{\psi,i}(\pi_1) = (h(X_i, \pi_1), Z_i')'$, we have

$$||d_{\psi,i}(\pi_1)d_{\psi,i}(\pi_1)'|| \leq \overline{h}_i^2 + ||Z_i||^2. \tag{13.14}$$

Applying the inequality in (13.10) with $a = d_{\psi,i}(\pi_1)$, $a^* = d_{\psi,i}(\pi_2)$, $||a_1 - a_1^*|| \leq \overline{h}_{\pi,i}||\pi_1 - \pi_2||$, and $||a_2 - a_2^*|| = 0$, we have

$$||d_{\psi,i}(\pi_1)d_{\psi,i}(\pi_1)' - d_{\psi,i}(\pi_2)d_{\psi,i}(\pi_2)'|| \leq 2(\overline{h}_i + ||Z_i||)\overline{h}_{\pi,i}||\pi_1 - \pi_2||. \tag{13.15}$$

Equations (13.11)-(13.15) combine to yield

$$||\rho_{\psi\psi}(W_i, \theta_1) - \rho_{\psi\psi}(W_i, \theta_2)|| \leq M_\psi(W_i)\delta, \tag{13.16}$$

where

$$M_\psi(W_i) = \left[2\overline{w}_{1,i}\left(M_1(W_i) + \overline{w}_{1,i}\overline{L}_i'(\overline{h}_i + ||Z_i|| + \overline{\beta} \cdot \overline{h}_{\pi,i})\right) + M_2(W_i) \right. \tag{13.17}$$
$$\left. + \overline{w}_{2,i}\overline{L}_i'(\overline{h}_i + ||Z_i|| + \overline{\beta} \cdot \overline{h}_{\pi,i})\right](\overline{h}_i^2 + ||Z_i||^2) + 2\left(\overline{w}_{1,i}^2 + \overline{w}_{2,i}\right)(\overline{h}_i + ||Z_i||)\overline{h}_{\pi,i}.$$

To show $||\rho_{\theta\theta}^\dagger(\theta_1) - \rho_{\theta\theta}^\dagger(\theta_2)|| \leq M_{\theta\theta}(W_i)\delta$ for some function $M_{\theta\theta}(W_i)$, the calculation is the same as that above with $d_{\psi,i}(\pi)$ replaced by $d_i(\pi)$. The inequalities in (13.14) and (13.15) become

$$||d_i(\pi_1)d_i(\pi_1)'|| \leq \overline{h}_i^2 + ||Z_i||^2 + \overline{h}_{\pi,i}^2 \text{ and} \tag{13.18}$$
$$||d_i(\pi_1)d_i(\pi_1)' - d_i(\pi_2)d_i(\pi_2)'|| \leq 2(\overline{h}_i + ||Z_i|| + \overline{h}_{\pi,i})(\overline{h}_{\pi,i} + \overline{h}_{\pi\pi,i}) \cdot ||\pi_1 - \pi_2||.$$

By the same arguments as those used in (13.11)-(13.17), we have

$$M_{\theta\theta}(W_i) = \left[2\overline{w}_{1,i}\left(M_1(W_i) + \overline{w}_{1,i}\overline{L}_i'(\overline{h}_i + ||Z_i|| + \overline{\beta} \cdot \overline{h}_{\pi,i})\right)\right.$$
$$\left. + M_2(W_i) + \overline{w}_{2,i}\overline{L}_i'(\overline{h}_i + ||Z_i|| + \overline{\beta} \cdot \overline{h}_{\pi,i})\right] \times (\overline{h}_i^2 + ||Z_i||^2 + \overline{h}_{\pi,i}^2)$$
$$+ 2\left(\overline{w}_{1,i}^2 + \overline{w}_{2,i}\right)(\overline{h}_i + ||Z_i|| + \overline{h}_{\pi,i})(\overline{h}_{\pi,i} + \overline{h}_{\pi\pi,i}). \tag{13.19}$$

Next, we show $||\rho_\theta^\dagger(\theta_1) - \rho_\theta^\dagger(\theta_2)|| \leq M_\theta(W_i)\delta$ for some function $M_\theta(W_i)$. To this end,

12

note that

$$||\rho_\theta^\dagger(\theta_1) - \rho_\theta^\dagger(\theta_2)||$$
$$\leq ||v_{1,i}(\theta_1) - v_{1,i}(\theta_2)|| \cdot (\overline{h}_i + ||Z_i|| + \overline{h}_{\pi,i}) + ||v_{1,i}(\theta_2)|| \cdot (\overline{h}_{\pi,i} + \overline{h}_{\pi\pi,i})\delta, \quad (13.20)$$

where $||v_{1,i}(\theta_1) - v_{1,i}(\theta_2)||$ satisfies the inequality in (13.12) and $||v_{1,i}(\theta_2)|| \leq \overline{w}_{1,i}$. Hence,

$$M_\theta(W_i) = \left(M_1(W_i) + \overline{w}_{1,i}\overline{L}_i'(\overline{h}_i + ||Z_i|| + \overline{\beta} \cdot \overline{h}_{\pi,i}) \right)(\overline{h}_i + ||Z_i|| + \overline{h}_{\pi,i}) + \overline{w}_{1,i}(\overline{h}_{\pi,i} + \overline{h}_{\pi\pi,i}). \quad (13.21)$$

Next, we show $||\varepsilon(W_i, \theta_1) - \varepsilon(W_i, \theta_2)|| \leq M_\varepsilon(W_i)\delta$ for some function $M_\varepsilon(W_i)$. To this end, note that

$$||\varepsilon(W_i, \theta_1) - \varepsilon(W_i, \theta_2)|| \leq 2||v_{1,i}(\theta_1)h_\pi(X_i, \pi_1) - v_{1,i}(\theta_2)h_\pi(X_i, \pi_2)||$$
$$+ ||v_{1,i}(\theta_1)h_{\pi\pi}(X_i, \pi_1) - v_{1,i}(\theta_2)h_{\pi\pi}(X_i, \pi_2)||$$
$$\leq 2||v_{1,i}(\theta_1) - v_{1,i}(\theta_2)||\overline{h}_{\pi,i} + 2||v_{1,i}(\theta_2)||\overline{h}_{\pi\pi,i}\delta \quad (13.22)$$
$$+ ||v_{1,i}(\theta_1) - v_{1,i}(\theta_2)||\overline{h}_{\pi\pi,i} + ||v_{1,i}(\theta_2)||M_h(W_i)\delta,$$

where $||v_{1,i}(\theta_1) - v_{1,i}(\theta_2)||$ satisfies the inequality in (13.12), $||v_{1,i}(\theta_2)|| \leq \overline{w}_{1,i}$, the first inequality follows from a mean-value expansion of $h_\pi(X_i, \pi)$ wrt π and the second inequality follows from $||h_{\pi\pi}(X_i, \pi_1) - h_{\pi\pi}(X_i, \pi_2)|| \leq M_h(X_i) \cdot ||\pi_1 - \pi_2||$. By (13.22), we have

$$M_\varepsilon(W_i) = \left(M_1(W_i) + \overline{w}_{1,i}\overline{L}_i'(\overline{h}_i + ||Z_i|| + \overline{\beta} \cdot \overline{h}_{\pi,i}) \right)(2\overline{h}_{\pi,i} + \overline{h}_{\pi\pi,i}) + \overline{w}_{1,i}(2\overline{h}_{\pi\pi,i} + M_h(W_i)). \quad (13.23)$$

Hence, Assumption S3(ii) holds with

$$M_1(W_i) = M_\psi(W_i) + M_{\theta\theta}(W_i) \text{ and } M_2(W_i) = M_\theta(W_i) + M_\varepsilon(W_i). \quad (13.24)$$

13.4. Verification of Assumption S3(iii)

The condition $E_{\gamma_0}M_2(W_i)^q \leq C_1$ for some $C_1 < \infty$ holds if $E_{\gamma_0}M_\theta(W_i)^q \leq C_2$ and $E_{\gamma_0}M_\varepsilon(W_i)^q \leq C_2$ for some $C_2 < \infty$. Because $\overline{L}_i' \leq \overline{w}_{1,i}$, $E_{\gamma_0}M_\theta(W_i)^q \leq C_2$ and $E_{\gamma_0}M_\varepsilon(W_i)^q \leq C_2$ hold provided, for some $C < \infty$, (i) $E_{\gamma_0}M_1^q(W_i)(\overline{h}_i^q + ||Z_i||^q + \overline{h}_{\pi,i}^q + \overline{h}_{\pi\pi,i}^q) \leq C$, (ii) $E_{\gamma_0}\overline{w}_{1,i}^{2q}(\overline{h}_i^q + ||Z_i||^q + \overline{h}_{\pi,i}^q)(\overline{h}_i^q + ||Z_i||^q + \overline{h}_{\pi,i}^q + \overline{h}_{\pi\pi,i}^q) \leq C$, and (iii) $E_{\gamma_0}\overline{w}_{1,i}^q(\overline{h}_{\pi,i}^q + \overline{h}_{\pi\pi,i}^q + M_h(W_i)^q) \leq C$. Condition (i) holds by conditions in (3.32) using

Hölder's inequality to give $E_{\gamma_0} M_1^q(W_i) \overline{h}_i^q \leq (E_{\gamma_0} M_1^{4q/3})^{3/4} (E_{\gamma_0} \overline{h}_i^{4q})^{1/4} \leq C$ and likewise with $||Z_i||$, $\overline{h}_{\pi,i}$, and $\overline{h}_{\pi\pi,i}$ in place of $\overline{h}_i$. Condition (ii) holds by $E_{\gamma_0} \overline{w}_{1,i}^{2q} \overline{h}_i^q ||Z_i||^q \leq (E_{\gamma_0} \overline{w}_{1,i}^{4q})^{1/2} (E_{\gamma_0} \overline{h}_i^{4q})^{1/4} (E_{\gamma_0} ||Z_i||^{4q})^{1/4} \leq C$ and likewise with $||Z_i||^q$ and $\overline{h}_{\pi,i}^q$ in place of $\overline{h}_i^q$ and $\overline{h}_i^q$, $\overline{h}_{\pi,i}^q$, and $\overline{h}_{\pi\pi,i}^q$ in place of $||Z_i||^q$. Condition (iii) holds by $E_{\gamma_0} \overline{w}_{1,i}^q M_h(W_i)^q \leq (E_{\gamma_0} \overline{w}_{1,i}^{4q})^{1/4} (E_{\gamma_0} M_h(W_i)^{4q/3})^{3/4} \leq C$ and likewise with $\overline{h}_{\pi,i}^q$ and $\overline{h}_{\pi\pi,i}^q$ in place of $M_h(W_i)^q$.

The condition $E_{\gamma_0} M_1(W_i) \leq C_1$ for some $C_1 < \infty$ holds if $E_{\gamma_0} M_\psi(W_i) \leq C_2$ and $E_{\gamma_0} M_{\theta\theta}(W_i) \leq C_2$ for some $C_2 < \infty$. Because $\overline{L}_i' \leq \overline{w}_{1,i}$, $E_{\gamma_0} M_\psi(W_i) \leq C_2$ and $E_{\gamma_0} M_{\theta\theta}(W_i) \leq C_2$ hold provided, for some $C < \infty$, (i) $E_{\gamma_0} M_1(W_i) \overline{w}_{1,i} (\overline{h}_i^2 + ||Z_i||^2 + \overline{h}_{\pi,i}^2) \leq C$, (ii) $E_{\gamma_0} \overline{w}_{1,i}^3 (\overline{h}_i + ||Z_i|| + \overline{h}_{\pi,i})(\overline{h}_i^2 + ||Z_i||^2 + \overline{h}_{\pi,i}^2) \leq C$, (iii) $E_{\gamma_0} M_2(W_i)(\overline{h}_i^2 + ||Z_i||^2 + \overline{h}_{\pi,i}^2) \leq C$, (iv) $E_{\gamma_0} \overline{w}_{1,i} \overline{w}_{2,i} (\overline{h}_i + ||Z_i|| + \overline{h}_{\pi,i})(\overline{h}_i^2 + ||Z_i||^2 + \overline{h}_{\pi,i}^2) \leq C$, (v) $E_{\gamma_0} (\overline{w}_{1,i}^2 + \overline{w}_{2,i}) (\overline{h}_i + ||Z_i|| + \overline{h}_{\pi,i})(\overline{h}_{\pi,i} + \overline{h}_{\pi\pi,i}) \leq C$. Condition (i) holds by conditions in (3.32) using the Cauchy-Schwarz inequality and $q > 2$ to give $E_{\gamma_0} M_1(W_i) \overline{w}_{1,i} \overline{h}_i^2 \leq (E_{\gamma_0} M_1(W_i)^2)^{1/2} (E_{\gamma_0} \overline{w}_{1,i}^4)^{1/4} (E_{\gamma_0} \overline{h}_i^8)^{1/4} \leq C$ and likewise with $||Z_i||^2$ and $\overline{h}_{\pi,i}^2$ in place of $\overline{h}_i^2$. Condition (ii) holds by $E_{\gamma_0} \overline{w}_{1,i}^3 \overline{h}_i \overline{h}_{\pi,i}^2 \leq (E_{\gamma_0} \overline{w}_{1,i}^6)^{1/2} (E_{\gamma_0} \overline{h}_i^4)^{1/4} (E_{\gamma_0} \overline{h}_{\pi,i}^8)^{1/4} \leq C$ and likewise with $||Z_i||$ and $\overline{h}_{\pi,i}$ in place of $\overline{h}_i$ and with $\overline{h}_i^2$ and $||Z_i||^2$ in place of $\overline{h}_{\pi,i}^2$. Condition (iii) holds by $E_{\gamma_0} M_2(W_i) \overline{h}_i^2 \leq (E_{\gamma_0} M_2(W_i)^{4/3})^{3/4} (E_{\gamma_0} \overline{h}_i^8)^{1/4} \leq C$ and likewise with $||Z_i||^2$ and $\overline{h}_{\pi,i}^2$ in place of $\overline{h}_i^2$. Condition (iv) holds by $E_{\gamma_0} \overline{w}_{1,i} \overline{w}_{2,i} \overline{h}_i \overline{h}_{\pi,i}^2 \leq (E_{\gamma_0} \overline{w}_{1,i}^8)^{1/8} (E_{\gamma_0} \overline{w}_{2,i}^2)^{1/2} (E_{\gamma_0} \overline{h}_i^8)^{1/8} (E_{\gamma_0} \overline{h}_{\pi,i}^8)^{1/4} \leq C$ and likewise with $||Z_i||$ and $\overline{h}_{\pi,i}$ in place of $\overline{h}_i$ and with $\overline{h}_i^2$ and $||Z_i||^2$ in place of $\overline{h}_{\pi,i}^2$. Condition (v) holds by $E_{\gamma_0} \overline{w}_{2,i} \overline{h}_i \overline{h}_{\pi,i} \leq (E_{\gamma_0} \overline{w}_{2,i}^2)^{1/2} (E_{\gamma_0} \overline{h}_i^4)^{1/4} (E_{\gamma_0} \overline{h}_{\pi,i}^4)^{1/4} \leq C$ and likewise with $\overline{w}_{1,i}^2$ in place of $\overline{w}_{2,i}$, $||Z_i||$ and $\overline{h}_{\pi,i}$ in place of $\overline{h}_i$, and $\overline{h}_{\pi\pi,i}$ in place of $\overline{h}_{\pi,i}$.

By (13.6),

$$E_{\gamma_0} \sup_{\theta \in \Theta} |\rho(W_i, \theta)|^{1+\delta} \leq E_{\gamma_0} (\sup_{\theta \in \Theta} |\log L_i(\theta)| + \sup_{\theta \in \Theta} |\log(1 - L_i(\theta))|)^{1+\delta} \leq C, \qquad (13.25)$$

for some $C < \infty$, where the first inequality holds because Y_i is 0 or 1 and the second inequality holds by conditions in (3.32).

By (3.22),

$$E_{\gamma_0} \sup_{\theta \in \Theta} ||\rho_{\psi\psi}(W_i, \theta)||^{1+\delta} \leq E_{\gamma_0} (\overline{w}_{1,i}^2 + \overline{w}_{2,i})^{1+\delta} \sup_{\pi \in \Pi} ||d_{\psi,i}(\pi) d_{\psi,i}(\pi)'||^{1+\delta} \leq C \qquad (13.26)$$

for some $C < \infty$, where the first inequality holds by $|Y_i - L_i(\theta)| \leq 1$ and the triangle inequality and the second inequality holds (13.14) and conditions in (3.32). Similarly, we can show $E_{\gamma_0} \sup_{\theta \in \Theta} ||\rho_{\theta\theta}^\dagger(W_i, \theta)||^{1+\delta} \leq C$ with $d_{\psi,i}(\pi)$ in (13.26) replaced by $d_i(\pi)$

and (13.14) replaced by (13.18).

By (3.23),

$$E_{\gamma_0} \sup_{\theta \in \Theta} ||\rho_\theta^\dagger(W_i, \theta)||^q \le E_{\gamma_0}(\overline{w}_{1,i} \sup_{\pi \in \Pi} ||d_i(\pi)||)^q$$
$$\le E_{\gamma_0} \overline{w}_{1,i}^q (\overline{h}_i + ||Z_i|| + \overline{h}_{\pi,i})^q \le C \qquad (13.27)$$

for some $C < \infty$, where the first inequality holds because $|Y_i - L_i(\theta)| \le 1$, the second inequality holds because $||d_i(\pi)|| \le ||h(X_i, \pi)|| + ||Z_i|| + ||h_\pi(X_i, \pi)||$, and the third inequality holds by conditions in (3.32).

By (3.23),

$$E_{\gamma_0} \sup_{\theta \in \Theta} ||\varepsilon(W_i, \theta)||^q \le E_{\gamma_0} \overline{w}_{1,i}^q (2\overline{h}_{\pi,i} + \overline{h}_{\pi\pi,i})^q \le C \qquad (13.28)$$

for some $C < \infty$, where the the first inequality follows from $|Y_i - L_i(\theta)| \le 1$ and the second inequality holds by conditions in (3.32).

This completes the verification of Assumption S3(iii).

13.5. Verification of Assumptions S3(iv) and S3(v)

To verify Assumption S3(iv), we apply the LIE and obtain

$$E_{\gamma_0} \rho_{\psi\psi}(W_i, \theta) = E_{\gamma_0}[w_{1,i}^2(\theta)e_{1,i}(\theta) + w_{2,i}(\theta)e_{2,i}(\theta)]d_{\psi,i}(\pi)d_{\psi,i}(\pi)', \text{ where} \qquad (13.29)$$
$$e_{1,i}(\theta) = E_{\gamma_0}((Y_i - L_i(\theta))^2 | X_i, Z_i) \text{ and } e_{2,i}(\theta) = E_{\gamma_0}(Y_i - L_i(\theta) | X_i, Z_i).$$

When $\beta_0 = 0$, $g_i(\psi_0, \pi) = Z_i'\zeta_0$ and $L_i(\psi_0, \pi) = L(g_i(\psi_0, \pi)) = L(Z_i'\zeta_0)$, $\forall \pi \in \Pi$. By (2.4),

$$e_{1,i}(\psi_0, \pi) = L(Z_i'\zeta_0)(1 - L(Z_i'\zeta_0)) \text{ and } e_{2,i}(\psi_0, \pi) = 0. \qquad (13.30)$$

Hence, when $\beta_0 = 0$,

$$E_{\gamma_0} \rho_{\psi\psi}(W_i, \psi_0, \pi) = E_{\gamma_0} \frac{L'^2(Z_i'\zeta_0)}{L(Z_i'\zeta_0)(1 - L(Z_i'\zeta_0))} d_{\psi,i}(\pi)d_{\psi,i}(\pi)'. \qquad (13.31)$$

The quantity $E_{\gamma_0}\rho_{\psi\psi}(W_i, \psi_0, \pi)$ is continuous in π on Π by the DCT using (13.16), (13.17), and the discussion following (13.24). Hence, $\lambda_{\min}(E_{\gamma_0}\rho_{\psi\psi}(W_i, \psi_0, \pi))$ also is continuous on the compact set Π and attains its minimum at some point $\pi_{\min} \in \Pi$. Its minimum is zero only if the positive semi-definite matrix $E_{\gamma_0}\rho_{\psi\psi}(W_i, \psi_0, \pi_{\min})$ is not

15

positive definite. The latter is ruled out by the fact that $L'^2(Z_i'\zeta_0)/(L(Z_i'\zeta_0)(1-L(Z_i'\zeta_0)))$ is positive a.s. and the condition in (3.32) that $P_{\gamma_0}(a'(h(X_i,\pi),Z_i)=0)<1,\,\forall\pi\in\Pi$, $\forall a\in R^{d_\zeta+1}$ with $a\neq 0$. Thus, $\inf_{\pi\in\Pi}\lambda_{\min}(E_{\gamma_0}\rho_{\psi\psi}(W_i,\psi_0,\pi))>0$ when $\beta_0=0$ and the first part of Assumption S3(iv) holds.

As in (13.29)-(13.31), we can show

$$E_{\gamma_0}\rho_{\theta\theta}^\dagger(W_i,\theta_0)=E_{\gamma_0}\frac{L_i'^2(\theta_0)}{L_i(\theta_0)(1-L_i(\theta_0))}d_i(\pi_0)d_i(\pi_0)' \qquad (13.32)$$

by replacing (ψ_0,π) with θ_0 and $d_{\psi,i}(\pi)$ with $d_i(\pi_0)$ in the arguments above. Because $L_i'(\theta_0)>0$ and $0<L_i(\theta_0)<1$, $E_{\gamma_0}\rho_{\theta\theta}^\dagger(W_i,\theta_0)$ is positive definite because $E_{\gamma_0}d_i(\pi_0)d_i(\pi_0)'$ is positive definite as specified in (3.32). Hence, the second part of Assumption S3(iv) holds.

By (13.4) and (13.32), $V^\dagger(\theta_0,\theta_0;\gamma_0)=E_{\gamma_0}\rho_{\theta\theta}^\dagger(W_i,\theta_0)$. Hence, $V^\dagger(\theta_0,\theta_0;\gamma_0)$ is positive definite.

13.6. Verification of Assumption S4

Because $m(W_i,\theta)=\rho_\psi(W_i,\theta)$ by Lemma 9.1,

$$E_{\gamma_0}m(W_i,\theta)=E_{\gamma_0}\rho_\psi(W_i,\theta)=E_{\gamma_0}w_{1,i}(\theta)(Y_i-L_i(\theta))d_{\psi,i}(\pi)$$
$$=E_{\gamma_0}w_{1,i}(\theta)(L_i(\theta_0)-L_i(\theta))d_{\psi,i}(\pi), \qquad (13.33)$$

where $\gamma_0=(\beta_0,\zeta_0,\pi_0,\phi_0)$, the second equality holds by (3.22), and the third equality holds by iterated expectations and (2.4). In (13.33), $E_{\gamma_0}m(W_i,\theta)$ depends on β_0 only through $L_i(\theta_0)$. Hence,

$$K(\theta;\gamma_0)=(\partial/\partial\beta_0)E_{\gamma_0}w_{1,i}(\theta)(L_i(\theta_0)-L_i(\theta))d_{\psi,i}(\pi)$$
$$=E_{\gamma_0}w_{1,i}(\theta)L_i'(\theta_0)h(X_i,\pi_0)d_{\psi,i}(\pi), \qquad (13.34)$$

where the first equality holds because the observations are identically distributed and the second equality holds by an exchange of E and ∂ because $E_{\gamma_0}\sup_{\theta\in\Theta,\theta_0\in\Theta_0}||w_{1,i}(\theta)L_i'(\theta_0)$ $h(X_i,\pi_0)d_{\psi,i}(\pi)||<\infty$ by conditions in (3.32) and $(\partial/\partial\beta_0)g_i(\theta_0)=h(X_i,\pi_0)$. Hence, Assumption S4(i) holds.

Now we show that Assumptions S4(ii) holds with

$$K(\pi; \gamma_0) = K(\psi_0, \pi; \gamma_0) = E_{\gamma_0} w_{1,i}(\psi_0, \pi) L_i'(\theta_0) h(X_i, \pi_0) d_{\psi,i}(\pi). \qquad (13.35)$$

Define $a_i(\theta, \theta_0) = w_{1,i}(\theta) L_i'(\theta_0) h(X_i, \pi_0) d_{\psi,i}(\pi)$. It suffices to show that $E_{\gamma_n} a_i(\theta, \theta^*) \to E_{\gamma_0} a_i(\theta, \theta^*)$ uniformly over $(\theta, \theta^*) \in \Theta \times \Theta^*$ as $\gamma_n \to \gamma_0$ and $E_{\gamma_0} a_i(\theta, \theta^*)$ is continuous in (θ, θ^*). The continuity holds by the continuity of $a_i(\theta, \theta^*)$ in (θ, θ^*), $E_{\gamma_0} \sup_{(\theta, \theta^*) \in \Theta \times \Theta^*} ||a_i(\theta, \theta^*)|| < \infty$ by conditions in (3.32), and the dominated convergence theorem. By Lemma 9.3, the uniform convergence follows from the pointwise convergence and the equicontinuity of $E_{\gamma_0} a_i(\theta, \theta^*)$ in (θ, θ^*) over $\gamma_0 \in \Gamma$. The pointwise convergence $E_{\gamma_n} a_i(\theta, \theta^*) \to E_{\gamma_0} a_i(\theta, \theta^*)$ holds because (i) the expectations $E_{\gamma_n} a_i(\theta, \theta^*)$ and $E_{\gamma_0} a_i(\theta, \theta^*)$ depend on ϕ_n and ϕ_0, respectively, but not on θ_n and θ_0, (ii) $\phi_n \to \phi_0$ implies convergence in distribution by the metric on Φ^*, and (iii) the $L^{1+\delta}$ boundedness of $a_i(\theta, \theta^*)$, i.e., $E_{\gamma_0} ||a_i(\theta, \theta^*)||^{1+\delta} \leq C < \infty$ for any $\gamma_0 \in \Gamma$. Equicontinuity holds because for any (θ_1, θ_1^*) and (θ_2, θ_2^*) with $||(\theta_1, \theta_1^*) - (\theta_2, \theta_2^*)|| \leq \delta$,

$$\begin{aligned}
&E_{\gamma_0} ||a_i(\theta_1, \theta_1^*) - a_i(\theta_2, \theta_2^*)|| \\
&\leq E_{\gamma_0} ||w_{1,i}(\theta_1) - w_{1,i}(\theta_2)|| \cdot ||L_i'(\theta_1^*) h(X_i, \pi_1^*) d_{\psi,i}(\pi_1)|| \\
&\quad + E_{\gamma_0} ||w_{1,i}(\theta_2)|| \cdot ||L_i'(\theta_1^*) h(X_i, \pi_1^*) d_{\psi,i}(\pi_1) - L_i'(\theta_2^*) h(X_i, \pi_2^*) d_{\psi,i}(\pi_2)|| \\
&\leq E_{\gamma_0} M_1(W_i) \overline{L}_i' \overline{h}_i \sup_{\pi \in \Pi} ||d_{\psi,i}(\pi)|| \delta \qquad (13.36) \\
&\quad + E_{\gamma_0} \overline{w}_{1,i} \left[\left(\overline{L}_i'' \overline{h}_i + \overline{L}_i' \overline{h}_{\pi,i} \right) \sup_{\pi \in \Pi} ||d_{\psi,i}(\pi)|| + \overline{L}_i' \overline{h}_i || \sup_{\pi \in \Pi} (\partial / \partial \pi') d_{\psi,i}(\pi)|| \right] \delta \leq C\delta
\end{aligned}$$

for some $C < \infty$ for all $\gamma_0 \in \Gamma$, where the first inequality holds by the triangle inequality, the second inequality follows from $||w_{1,i}(\theta_1) - w_{1,i}(\theta_2)|| \leq M_1(W_i)\delta$ and a mean-value expansion of $L_i'(\theta_1^*) h(X_i, \pi_1^*) d_{\psi,i}(\pi_1)$ wrt (θ_1, θ_1^*) around (θ_2, θ_2^*), and the third inequality holds by the Cauchy-Schwarz inequality and conditions in (3.32). This completes the verification of Assumption S4.

13.7. Verification of Assumptions B1 and B2

Given the definitions in Section 3.2, Assumptions B1(i) and B1(iii) follow immediately. Assumption B1(ii) holds by taking $\delta < \min\{b_1^*, b_2^*\}$ and $\mathcal{Z}^0 = int(\mathcal{Z})$.

Given the definitions in Sections 3.2, the true parameter space Γ is of the form in (2.6). Thus, Assumption B2(i) holds immediately. Assumption B2(ii) follows from the

17

form of $\mathcal{B}^*$ given in (2.9). Assumption B2(iii) follows from the form of $\mathcal{B}^*$ and the fact that Θ^* is a product space and $\Phi^*(\theta_0)$ does not depend on β_0. Hence, the true parameter space Γ satisfies Assumption B2.

13.8. Verification of Assumption C6

Assumption C6 holds by Lemma 3.2 under Assumptions S1-S3 and C6†. We now verify Assumption C6. Assumption C6†(i) holds because β is a scalar. To verify Assumption C6†(ii), we have

$$\rho_\beta(W_i, \theta) = w_{1,i}(\theta)(Y_i - L_i(\theta))h(X_i, \pi) \text{ and } \rho_\zeta(W_i, \theta) = w_{1,i}(\theta)(Y_i - L_i(\theta))Z_i. \quad (13.37)$$

When $\beta_0 = 0$,

$$\rho_\psi^*(W_i, \psi_0, \pi_1, \pi_2) = w_{1,i}(\psi_0)(Y_i - L_i(\psi_0))h_{Z,i}(\pi_1, \pi_2), \text{ where}$$
$$w_{1,i}(\psi_0) = \frac{L'(Z_i'\zeta_0)}{L(Z_i'\zeta_0)(1 - L(Z_i'\zeta_0))}, \ L_i(\psi_0) = L(Z_i'\zeta_0), \text{ and}$$
$$h_{Z,i}(\pi_1, \pi_2) = (h(X_i, \pi_1), h(X_i, \pi_2), Z_i')'. \quad (13.38)$$

The covariance matrix in Assumption C6†(ii) is

$$\Omega_G(\pi_1, \pi_2; \gamma_0) = Cov_{\gamma_0}(\rho_\psi^*(W_i, \psi_0, \pi_1, \pi_2), \rho_\psi^*(W_i, \psi_0, \pi_1, \pi_2))$$
$$= E_{\gamma_0} w_{1,i}^2(\psi_0)(Y_i - L_i(\psi_0))^2 h_{Z,i}(\pi_1, \pi_2)h_{Z,i}(\pi_1, \pi_2)'$$
$$= E_{\gamma_0} \frac{L'^2(Z_i'\zeta_0)}{L(Z_i'\zeta_0)(1 - L(Z_i'\zeta_0))} h_{Z,i}(\pi_1, \pi_2)h_{Z,i}(\pi_1, \pi_2)', \quad (13.39)$$

where the first equality holds because the observations are independent and identically distributed, the second equality follows from $E\rho_\psi^*(W_i, \psi_0, \pi_1, \pi_2) = 0$, which in turn holds by the LIE and (2.4), and the third equality holds by (13.1). Because $L'(Z_i'\zeta_0) > 0$ and $0 < L(Z_i'\zeta_0) < 1$, $\Omega_G(\pi_1, \pi_2; \gamma_0)$ is positive definite because $P(a'h_{Z,i}(\pi_1, \pi_2) = 0) < 1$ for all $a \in R^{d_z+2}$ with $a \neq 0$ by the conditions in (3.32).

13.9. Verification of Assumptions V1 and V2

Here we verify Assumptions V1 (scalar β) and V2, which are stated in Appendix B of the main paper.

18

For the binary choice model, the matrices $J(\gamma_0)$ $(= V(\gamma_0))$ and $\widehat{J}_n(\theta)$ $(= \widehat{V}_n(\theta))$ are defined in (3.35) and (5.18), respectively. Define

$$J(\theta; \gamma_0) = E_{\gamma_0} \frac{L_i'^2(\theta)}{L_i(\theta)(1 - L_i(\theta))} d_i(\pi) d_i(\pi)'. \tag{13.40}$$

Under $\{\gamma_n\} \in \Gamma(\gamma_0)$, $\sup_{\theta \in \Theta} ||\widehat{J}_n(\theta) - J(\theta; \gamma_0)|| \to_p 0$ and $J(\theta; \gamma_0)$ is continuous in θ on Θ by the uniform law of large numbers in Lemma 9.3, where the smoothness and moment conditions hold by conditions in (3.32). In addition, $J(\theta_0; \gamma_0) = J(\gamma_0)$. This verifies Assumption V1(i) and V1(ii) (for scalar β).

To verify Assumption V1(iii), note that

$$\Sigma(\theta; \gamma_0) = J^{-1}(\theta; \gamma_0) \text{ and } \Sigma(\pi; \gamma_0) = J^{-1}(\psi_0, \pi; \gamma_0). \tag{13.41}$$

Hence, it suffices to show that (i) $\lambda_{\min}(J(\psi_0, \pi; \gamma_0)) > 0$ and (ii) $\lambda_{\max}(J(\psi_0, \pi; \gamma_0)) < \infty$ for all $\pi \in \Pi$. Property (i) holds by essentially the same argument as in the paragraph following (13.31) with $d_i(\pi)$ in place of $d_{\psi,i}(\pi)$ using the condition in (3.32) that $E_{\gamma_0} d_i(\pi) d_i(\pi)'$ is positive definite $\forall \pi \in \Pi$. Positive definiteness of $E_{\gamma_0} d_i(\pi) d_i(\pi)'$ implies the same for $E_{\gamma_0}[L'^2(Z_i'\zeta_0)/(L(Z_i'\zeta_0)(1 - L(Z_i'\zeta_0)))] d_i(\pi) d_i(\pi)'$ because the latter is well-defined and $L'^2(Z_i'\zeta_0)/(L(Z_i'\zeta_0)(1 - L(Z_i'\zeta_0)))$ is positive a.s. Property (ii) holds by the moment conditions in (3.32). This completes the verification of Assumption V1(iii).

Assumptions V1(i) and V1(ii) hold not only under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, but also under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ in this example. This and $\widehat{\theta}_n \to_p \theta_0$ under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, which holds by Lemma 5.3 of AC1, imply that Assumption V2 holds. Among the assumptions employed in Lemma 5.3 of AC1, Assumptions B1, B2, and C7 are verified directly, Assumptions A, B3, and C1-C5 hold by Lemma 9.1 under Assumptions B1, B2, and S1-S4, and Assumption C6 holds by Lemma 3.2 under Assumptions S1-S3 and C6[†].

13.10. Calculation of Partial Derivatives

Here we calculate the partial derivatives of $\rho(W_i, \theta)$ wrt θ. Let L abbreviate $L(g_i(\theta))$. The first-order derivative wrt θ is

$$
\begin{aligned}
\rho_\theta(W_i, \theta) &= -\left[\frac{Y_i}{L} - \frac{1-Y_i}{1-L}\right] L' \frac{\partial}{\partial \theta} g_i(\theta) \\
&= -\frac{Y_i - L}{L(1-L)} L' \frac{\partial}{\partial \theta} g_i(\theta) = w_{1,i}(\theta)(Y_i - L)B(\beta)d_i(\pi), \text{ where} \\
w_{1,i}(\theta) &= \frac{-L'}{L(1-L)}. \tag{13.42}
\end{aligned}
$$

Now we calculate the second-order derivatives. To this end, we have

$$
\begin{aligned}
\frac{\partial}{\partial \theta'}\left[\frac{Y_i}{L} - \frac{1-Y_i}{1-L}\right] &= \left[\frac{-Y_i}{L^2} + \frac{-(1-Y_i)}{(1-L)^2}\right] L' \frac{\partial}{\partial \theta'} g_i(\theta) \\
&= -\left[\frac{Y_i(1-L)^2 + (1-Y_i)L^2}{L^2(1-L)^2}\right] L' \frac{\partial}{\partial \theta'} g_i(\theta) \\
&= -\left[\frac{Y_i - 2Y_i L + L^2}{L^2(1-L)^2}\right] L' \frac{\partial}{\partial \theta'} g_i(\theta) \\
&= -\frac{(Y_i - L)^2}{L^2(1-L)^2} L' \frac{\partial}{\partial \theta'} g_i(\theta), \\
\frac{\partial}{\partial \theta'} L' &= L'' \frac{\partial}{\partial \theta'} g_i(\theta), \text{ and } \frac{\partial^2}{\partial \theta \partial \theta'} g_i(\theta) = D_i(\theta). \tag{13.43}
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\rho_{\theta\theta}(W_i, \theta) &= \left[\frac{(Y_i - L)^2}{L^2(1-L)^2}(L')^2 - \frac{Y_i - L}{L(1-L)}L''\right]\left(\frac{\partial}{\partial \theta}g_i(\theta)\frac{\partial}{\partial \theta'}g_i(\theta)\right) \\
&\quad - \frac{Y_i - L}{L(1-L)}L'\frac{\partial^2}{\partial \theta \partial \theta'}g_i(\theta) \\
&= [w_{1,i}^2(Y_i - L)^2 + w_{2,i}(Y_i - L)]B(\beta)d_i(\pi)d_i(\pi)'B(\beta) \\
&\quad + w_{1,i}(Y_i - L)D_i(\theta), \text{ where} \\
w_{1,i}(\theta) &= \frac{-L'}{L(1-L)} \text{ and } w_{2,i}(\theta) = \frac{-L''}{L(1-L)}. \tag{13.44}
\end{aligned}
$$

Lastly, we calculate the derivatives in (13.8). Let $L = L_i(\theta)$ and $L_0 = L_i(\theta_0)$. We

have

$$\text{FOC} = -\frac{L_0}{L} + \frac{1 - L_0}{1 - L} = \frac{L - L_0}{L(1 - L)} \text{ and}$$

$$\text{SOC} = \frac{L(1 - L) - (L - L_0)(1 - 2L)}{L^2(1 - L)^2}$$

$$= \frac{L - L^2 - (L - L_0 - 2L^2 + 2LL_0)}{L^2(1 - L)^2}$$

$$= \frac{L_0 + L^2 - 2LL_0}{L^2(1 - L)^2} > \frac{(L_0 - L)^2}{L^2(1 - L)^2} > 0. \tag{13.45}$$

14. Supplemental Appendix E: STAR Example, Verification of Assumptions

14.1. Verification of Assumptions S1 and S2

Assumption S1 holds by Assumption STAR1(ii).

Assumption S2(i) holds with

$$\rho(W_t, \theta) = U_t^2(\theta)/2, \text{ where } U_t(\theta) = Y_t - X_t'\zeta - X_t'\beta \cdot m(Z_t, \pi). \qquad (14.46)$$

The residual $U_t(\theta)$ is twice continuously differentiable in θ for both the logistic function and the exponential function. When $\beta = 0$, $U_t(\theta) = Y_t - X_t'\zeta$, which does not depend on π. This verifies Assumption S2(ii).

To verify Assumptions S2(iii) and S2(iv), we have

$$
\begin{aligned}
E_{\gamma_0}\rho(W_t, \theta) &= E_{\gamma_0}\left[Y_t - X_t'\zeta - X_t'\beta \cdot m(Z_t, \pi)\right]^2 \\
&= E_{\gamma_0}\left(U_t - X_t'(\zeta - \zeta_0) - X_t'[\beta m(Z_t, \pi) - \beta_0 m(Z_t, \pi_0)]\right)^2 \\
&= E_{\gamma_0}U_t^2 + E_{\gamma_0}[X_t'(\zeta - \zeta_0) + X_t'(\beta m(Z_t, \pi) - \beta_0 m(Z_t, \pi_0))]^2. \quad (14.47)
\end{aligned}
$$

To verify Assumption S2(iii), we need that when $\beta_0 = 0$,

$$E_{\gamma_0}\rho(W_t, \psi, \pi) - E_{\gamma_0}\rho(W_t, \psi_0, \pi) = E_{\gamma_0}[X_t'(\zeta - \zeta_0) + X_t'\beta m(Z_t, \pi)]^2 > 0 \qquad (14.48)$$

$\forall \psi \neq \psi_0$ and $\forall \pi \in \Pi$. The inequality in (14.48) holds unless

$$P_{\gamma_0}((X_t' + X_t'm(Z_t, \pi))a = 0) = 1, \qquad (14.49)$$

where $a = ((\zeta - \zeta_0)', \beta')'$. By Assumption STAR2(i), (14.49) does not hold for any $a \neq 0$. Hence, the inequality in (14.48) holds $\forall \psi \neq \psi_0$. This completes the verification of Assumption S2(iii).

To verify Assumption S2(iv), we need that when $\beta_0 \neq 0$,

$$
\begin{aligned}
&E_{\gamma_0}\rho(W_t, \theta) - E_{\gamma_0}\rho(W_t, \theta_0) \\
&= E_{\gamma_0}[X_t'(\zeta - \zeta_0) + X_t'\beta m(Z_t, \pi) - X_t'\beta_0 m(Z_t, \pi_0)]^2 > 0 \qquad (14.50)
\end{aligned}
$$

$\forall \theta \neq \theta_0$. The inequality in (14.50) holds unless

$$P_{\gamma_0} \left(X_t'(\zeta - \zeta_0) + X_t'\beta m(Z_t, \pi) - X_t'\beta_0 m(Z_t, \pi_0) = 0 \right) = 1 \qquad (14.51)$$

for some $\theta \neq \theta_0$. Because $\beta_0 \neq 0$, Assumption STAR2(i) implies that (14.51) does not hold for any $\pi \neq \pi_0$. When $\pi = \pi_0$, (14.51) becomes

$$P_{\gamma_0} \left(X_t'(\zeta - \zeta_0) + X_t'(\beta - \beta_0) m(Z_t, \pi_0) = 0 \right) = 1. \qquad (14.52)$$

Because (14.49) does not hold for any $a \neq 0$ for any $\pi \in \Pi$, (14.52) cannot hold for $(\beta, \zeta) \neq (\beta_0, \zeta_0)$. This completes the verification of Assumption S2(iv).

Assumption S2(v) holds by Assumption STAR5(ii). Assumption S2(vi) holds because Ψ does not depend on π.

14.2. Verification of Assumption S3(i)

Now we verify Assumption S3 (vector β). In the STAR model, Z_t is an element of X_t and the function $\rho(\omega, \theta)$ takes the form in (3.19) with

$$a(X_t, \beta) = X_t'\beta \in R, \ h(X_t, \pi) = m(Z_t, \pi) \in R, \text{ and}$$
$$\rho^*(W_t, a(X_t, \beta)h(X_t, \pi), \zeta) = [Y_t - X_t'\zeta - a(X_t, \beta)h(X_t, \pi)]^2/2. \qquad (14.53)$$

By Lemma 3.1, we verify Assumption S3(i) by showing that Assumption S3* holds.

To verify Assumption S3*(i), we have

$$\rho'(W_t, a(X_t, \beta_0)h(X_t, \pi_0), \zeta_0) = -[Y_t - X_t'\zeta_0 - a(X_t, \beta_0)h(X_t, \pi_0)] = -U_t. \qquad (14.54)$$

Note that $\rho'(\cdot)$ and $\rho''(\cdot)$ in Assumption S3* are partial derivatives of $\rho^*(\cdot)$ wrt $a(X_t, \beta)$ $h(X_t, \pi)$. Assumption S3*(i) holds immediately by Assumption STAR1(i).

To verify Assumption S3*(ii), we first derive the terms that appear in it. By (14.53),

$$\rho''(W_t, a(X_t, \beta)h(X_t, \pi), \zeta) = 1,$$
$$h(X_t, \pi) = m(Z_t, \pi), h_\pi(X_t, \pi) = m_\pi(Z_t, \pi), \ h_{\pi\pi}(X_t, \pi) = m_{\pi\pi}(Z_t, \pi),$$
$$a_\beta(X_t, \beta) = X_t, \ a_{\beta\beta}(X_t, \beta) = 0. \qquad (14.55)$$

Assumption S3*(ii) holds because $E_{\gamma_0} \sup_{\pi \in \Pi}(|m(Z_t, \pi)| + ||m_\pi(Z_t, \pi)||) \cdot (|m(Z_t, \pi)| +$

23

$||m_\pi(Z_t,$

$\pi)|| + ||m_{\pi\pi}(Z_t, \pi)||) \cdot ||X_t||^2 \le C$ for some $C < \infty$ by Assumption STAR2(iii) and the Cauchy-Schwarz inequality.

This completes the verification of Assumption S3(i).

14.3. Verification of Assumption S3(ii)

Next, we verify Assumption S3(ii). We first show some generic results that are used in the calculation below. Let $A = aa'$, where $a = [a_1', ..., a_m']' \in R^{d_a}$ and $a_1, ..., a_m$ are sub-vectors of a. Similarly, $A^* = a^* a^{*\prime}$ and $a_1^*, ..., a_m^*$ are sub-vectors of a^*. Then,

$$
\begin{aligned}
||A - A^*|| = ||aa' - a^* a^{*\prime}|| &\le \sum_{i=1}^{m} \sum_{j=1}^{m} ||a_i a_j' - a_i^* a_j^{*\prime}|| \\
&\le \sum_{i=1}^{m} \sum_{j=1}^{m} \left(||a_i a_j' - a_i a_j^{*\prime}|| + ||a_i a_j^{*\prime} - a_i^* a_j^{*\prime}|| \right) \\
&\le \sum_{i=1}^{m} (||a_i|| + ||a_i^*||) \sum_{j=1}^{m} ||a_i - a_i^*||,
\end{aligned}
\tag{14.56}
$$

where the first inequality holds by the inequality $(x^2 + y^2)^{1/2} \le x + y$ for non-negative scalars x and y, the second inequality holds by the triangle inequality, and the third inequality holds by the inequality $||AB|| \le ||A|| \cdot ||B||$ for matrices A and B.

By (7.12),

$$
\begin{aligned}
||\rho_{\psi\psi}(W_t, \theta_1) - \rho_{\psi\psi}(W_t, \theta_2)|| &\le ||d_{\psi,t}(\pi_1) d_{\psi,t}(\pi_1)' - d_{\psi,t}(\pi_2) d_{\psi,t}(\pi_2)'|| \\
&\le 4||X_t|| \cdot ||X_t' m(Z_t, \pi_1) - X_t' m(Z_t, \pi_2)|| \\
&\le 4||X_t||^2 \cdot \sup_{\pi \in \Pi} ||m_\pi(Z_t, \pi)|| \cdot ||\pi_1 - \pi_2||,
\end{aligned}
\tag{14.57}
$$

where the first inequality holds by applying the inequality in (14.56) to $a = d_{\psi,t}(\pi_1) = (X_t' m(Z_t, \pi_1), X_t')'$ and $a^* = d_{\psi,t}(\pi_2) = (X_t' m(Z_t, \pi_2), X_t')'$ and the second inequality holds by a mean-value expansion of $m(Z_t, \pi)$ wrt π.

Applying the arguments in (14.57) to $\rho_{\theta\theta}^\dagger(W_t, \theta^+)$ with $a = (X_t' m(Z_t, \pi_1), X_t', \omega_1' X_t m_\pi(Z_t,$

24

$\pi_1)')'$ and $a^* = (X_t'm(Z_t, \pi_2), X_t', \omega_2'X_tm_\pi(Z_t, \pi_2)')'$ yields

$$||\rho_{\theta\theta}^\dagger(W_t, \theta_1^+) - \rho_{\theta\theta}^\dagger(W_t, \theta_2^+)|| \leq 2||X_t||^2(2 + \sup_{\pi\in\Pi}||m_\pi(Z_t, \pi)||) \times \qquad (14.58)$$

$$\left(\sup_{\pi\in\Pi}(||m_\pi(Z_t, \pi)|| + ||m_{\pi\pi}(Z_t, \pi)||) \cdot ||\pi_1 - \pi_2|| + \sup_{\pi\in\Pi}||m_\pi(Z_t, \pi)|| \cdot ||\omega_1 - \omega_2||\right).$$

Therefore, the function $M_1(W_t)$ in Assumption S3(ii) takes the form

$$M_1(W_t) = 4||X_t||^2 \cdot \sup_{\pi\in\Pi}||m_\pi(Z_t, \pi)||$$

$$+2||X_t||^2(2 + \sup_{\pi\in\Pi}||m_\pi(Z_t, \pi)||) \cdot \sup_{\pi\in\Pi}(2||m_\pi(Z_t, \pi)|| + ||m_{\pi\pi}(Z_t, \pi)||). \, (14.59)$$

The form of $M_1(W_t)$ is used in the verification of Assumption S3(iii) below.

Next, we show the form of $M_2(W_t)$ in Assumption S3(ii) (vector β). By (7.12),

$$||\rho_\psi(W_t, \theta_1) - \rho_\psi(W_t, \theta_2)|| = ||U_t(\theta_1)d_{\psi,t}(\pi_1) - U_t(\theta_2)d_{\psi,t}(\pi_2)||$$

$$\leq |U_t(\theta_1) - U_t(\theta_2)| \cdot ||d_{\psi,t}(\pi_2)|| + |U_t(\theta_1)| \cdot ||d_{\psi,t}(\pi_1) - d_{\psi,t}(\pi_2)||, \qquad (14.60)$$

where the inequality holds by the triangle inequality and $||aB|| = |a| \cdot ||B||$ when a is a scalar.

Let $\overline{\beta} = \sup_{\theta\in\Theta}||\beta||$ and $\overline{\zeta} = \sup_{\theta\in\Theta}||\zeta||$.

Note that in (14.60), the terms concerning $U_t(\theta)$ satisfy

$$|U_t(\theta_1) - U_t(\theta_2)| \leq \sup_{\theta\in\Theta}||\frac{\partial}{\partial\theta'}U_t(\theta)|| \cdot ||\theta_1 - \theta_2||$$

$$\leq (2||X_t|| + ||X_t|| \cdot \overline{\beta} \cdot \sup_{\pi\in\Pi}||m_\pi(Z_t, \pi)||) \cdot ||\theta_1 - \theta_2||,$$

$$|U_t(\theta_1)| \leq ||Y_t|| + ||X_t||\overline{\zeta} + ||X_t||\overline{\beta}. \qquad (14.61)$$

The terms concerning $d_{\psi,t}(\pi)$ satisfy

$$||d_{\psi,t}(\pi_2)|| \leq 2||X_t|| \text{ and}$$

$$||d_{\psi,t}(\pi_1) - d_{\psi,t}(\pi_2)|| \leq ||X_t|| \cdot \sup_{\pi\in\Pi}||m_\pi(Z_t, \pi)|| \cdot ||\pi_1 - \pi_2||. \qquad (14.62)$$

25

The inequalities in (14.60)-(14.62) imply that

$$||\rho_\psi(W_t, \theta_1) - \rho_\psi(W_t, \theta_2)|| \leq M_\psi(W_t) \cdot ||\theta_1 - \theta_2||, \text{ where}$$
$$M_\psi(W_t) = 2||X_t||^2(2 + \overline{\beta} \cdot \sup_{\pi \in \Pi} ||m_\pi(Z_t, \pi)||)$$
$$+(|Y_t| + ||X_t||\overline{\zeta} + ||X_t||\overline{\beta}) \cdot ||X_t|| \cdot \sup_{\pi \in \Pi} ||m_\pi(Z_t, \pi)||. \tag{14.63}$$

Similarly, (7.14) gives

$$||\rho_\theta^\dagger(W_t, \theta_1^+) - \rho_\theta^\dagger(W_t, \theta_2^+)|| = ||U_t(\theta_1^+)d_t(\pi_1, \omega_1) - U_t(\theta_2^+)d_t(\pi_2, \omega_2)|| \tag{14.64}$$
$$\leq |U_t(\theta_1^+) - U_t(\theta_2^+)| \cdot ||d_t(\pi_2, \omega_2)|| + |U_t(\theta_1^+)| \cdot ||d_t(\pi_1, \omega_1) - d_t(\pi_2, \omega_2)||.$$

In (14.64), the terms concerning $U_t(\theta^+)$ satisfy that

$$U_t(\theta^+) = Y_t - X_t'\zeta - ||\beta||\omega'X_t \cdot m(Z_t, \pi),$$
$$|U_t(\theta_1^+)| \leq ||Y_t|| + ||X_t||\overline{\zeta} + ||X_t||\overline{\beta},$$
$$\frac{\partial}{\partial\theta^{+'}}U(\theta^+) = -(\omega'X_t m(Z_t, \pi), ||\beta||X_t'm(Z_t, \pi), X_t', ||\beta||\omega'X_t m_\pi(Z_t, \pi)'),$$
$$|U_t(\theta_1^+) - U_t(\theta_2^+)| \leq \sup_{\theta^+ \in \Theta^+} ||\frac{\partial}{\partial\theta^+}U(\theta^+)|| \cdot ||\theta_1^+ - \theta_2^+||$$
$$\leq \left(2 + \overline{\beta} \cdot \left(\sup_{\pi \in \Pi} ||m_\pi(Z_t, \pi)|| + 1\right)\right) \cdot ||X_t|| \cdot ||\theta_1^+ - \theta_2^+||. \tag{14.65}$$

In (14.64), the terms concerning $d_t(\pi, \omega)$ satisfy

$$||d_t(\pi, \omega)|| \leq ||X_t||(2 + \sup_{\pi \in \Pi} ||m_\pi(Z_t, \pi)||) \text{ and}$$
$$||d_t(\pi_1, \omega_1) - d_t(\pi_2, \omega_2)|| \leq ||X_t|| \cdot \left(\sup_{\pi \in \Pi} ||m_\pi(Z_t, \pi)|| + \sup_{\pi \in \Pi} ||m_{\pi\pi}(Z_t, \pi)||\right) \cdot ||\pi_1 - \pi_2||$$
$$+||X_t|| \cdot \sup_{\pi \in \Pi} ||m_\pi(Z_t, \pi)|| \cdot ||\omega_1 - \omega_2||. \tag{14.66}$$

By (14.64)-(14.66),

$$||\rho_\theta^\dagger(W_t, \theta_1^+) - \rho_\theta^\dagger(W_t, \theta_2^+)|| \leq M_\rho(W_t) \cdot ||\theta_1^+ - \theta_2^+||, \text{ where}$$
$$M_\rho(W_t) = [2 + \overline{\beta} \cdot (\sup_{\pi \in \Pi} ||m_\pi(Z_t, \pi)|| + 1)] \cdot ||X_t||^2 \cdot (2 + \sup_{\pi \in \Pi} ||m_\pi(Z_t, \pi)||)$$
$$+ \left(||Y_t|| + ||X_t||\overline{\zeta} + ||X_t||\overline{\beta}\right) \cdot ||X_t|| \cdot \sup_{\pi \in \Pi} (2||m_\pi(Z_t, \pi)|| + ||m_{\pi\pi}(Z_t, \pi)||)(14.67)$$

26

Another term in Assumption S3(ii) is $||\varepsilon(W_t, \theta_1^+) - \varepsilon(W_t, \theta_2^+)||$, which satisfies

$$||\varepsilon(W_t, \theta_1^+) - \varepsilon(W_t, \theta_2^+)||$$
$$\leq |U_t(\theta_1^+) - U_t(\theta_2^+)| \cdot ||X_t|| \cdot \sup_{\pi \in \Pi}(2||m_\pi(Z_t, \pi) + ||m_{\pi\pi}(Z_t, \pi)||)$$
$$+ |U_t(\theta_1^+)| \cdot ||X_t|| \cdot \sup_{\pi \in \Pi}(||3m_{\pi\pi}(Z_t, \pi) + M_{\pi\pi}(Z_t)) \cdot ||\theta_1^+ - \theta_2^+||, \quad (14.68)$$

where $M_{\pi\pi}(Z_t)$ is as in Assumption STAR2. This and the inequalities in (14.65) imply that

$$||\varepsilon(W_t, \theta_1^+) - \varepsilon(W_t, \theta_2^+)|| \leq M_\varepsilon(W_t) \cdot ||\theta_1^+ - \theta_2^+||, \text{ where } M_\varepsilon(W_t) =$$
$$\left(2 + \overline{\beta} \cdot \left(\sup_{\pi \in \Pi}||m_\pi(Z_t, \pi)|| + 1\right)\right) \cdot ||X_t||^2 \cdot \sup_{\pi \in \Pi}(2||m_\pi(Z_t, \pi)|| + ||m_{\pi\pi}(Z_t, \pi)||)$$
$$+ \left(|Y_t| + ||X_t||\overline{\zeta} + ||X_t||\overline{\beta}\right) \cdot ||X_t|| \cdot (\sup_{\pi \in \Pi}||3m_{\pi\pi}(Z_t, \pi)|| + M_{\pi\pi}(Z_t)). \quad (14.69)$$

Equations (14.63), (14.67), and (14.69) yield that Assumption S3(ii) holds with

$$M_2(W_t) = M_\psi(W_t) + M_\rho(W_t) + M_\varepsilon(W_t). \quad (14.70)$$

14.4. Verification of Assumption S3(iii)

In the verification of Assumption S3(iii) below, we use

$$E_{\gamma_0} \sup_{\theta \in \Theta} |U_t(\theta)|^{2q} = E_{\gamma_0} \sup_{\theta \in \Theta} |Y_t - X_t'\zeta - X_t'\beta \cdot m(Z_t, \pi)|^{2q}$$
$$\leq C_1 E_{\gamma_0}(|Y_t| + ||X_t||)^{2q} \leq C_2 \quad (14.71)$$

for some $C_1, C_2 < \infty$, where the first inequality holds because the parameter spaces of ζ and β are bounded and $|m(Z_t, \pi)| \in [0, 1]$ and the second inequality holds by Holder's inequality and Assumptions STAR1(ii) and STAR2(iii). Because the value of $U_t(\theta)$ does not change when θ is reparameterized as θ^+, (14.71) is equivalent to

$$E_{\gamma_0} \sup_{\theta^+ \in \Theta^+} |U_t(\theta^+)|^{2q} \leq C \quad (14.72)$$

for some $C < \infty$.

27

By (14.46),

$$E_{\gamma_0} \sup_{\theta \in \Theta} |\rho(W_t, \theta)|^{1+\delta} = \frac{1}{2^{1+\delta}} E_{\gamma_0} \sup_{\theta \in \Theta} |U_t(\theta)|^{2(1+\delta)} \le C \tag{14.73}$$

for some $C < \infty$ by (14.71).

By (7.14),

$$E_{\gamma_0} \sup_{\theta^+ \in \Theta^+} ||\rho_\theta^\dagger(W_t, \theta^+)||^q \le E_{\gamma_0} \sup_{\theta^+ \in \Theta^+} |U_i(\theta^+)|^{2q} E_{\gamma_0} \sup_{\theta^+ \in \Theta^+} ||d_t(\pi, \omega)||^{2q}$$
$$\le C_1 E_{\gamma_0} \sup_{\pi \in \Pi} (2||X_t|| + ||X_t|| \cdot ||m_\pi(Z_t, \pi)||)^{2q} \le C_2 \tag{14.74}$$

for some $C_1, C_2 < \infty$, where the first inequality holds by the Cauchy-Schwarz inequality, the second inequality holds by (14.72) and $||AB|| \le ||A|| \cdot ||B||$, and the third inequality holds by Holder's inequality and Assumptions STAR1(ii) and STAR2(iii).

In the calculation of $E_{\gamma_0} \sup_{\theta \in \Theta} ||\rho_{\psi\psi}(W_i, \theta)||^{1+\delta}$ and $E_{\gamma_0} \sup_{\theta \in \Theta} ||\rho_{\theta\theta}^\dagger(W_i, \theta)||^{1+\delta}$ below, we use the following inequality. Let $A = aa'$, where $a = [a_1', ..., a_m'] \in R^{d_a}$ and $a_1, ..., a_m$ are sub-vectors of a. Then,

$$||A|| \le \sum_{i=1}^m \sum_{j=1}^m ||a_i a_j'|| \le \left(\sum_{i=1}^m ||a_i|| \right)^2, \tag{14.75}$$

by arguments analogous to those in (14.56).

By (7.12),

$$E_{\gamma_0} \sup_{\theta \in \Theta} ||\rho_{\psi\psi}(W_i, \theta)||^{1+\delta} = E_{\gamma_0} \sup_{\theta \in \Theta} ||d_{\psi,t}(\pi) d_{\psi,t}(\pi)'||^{1+\delta} \le E_{\gamma_0} (2||X_t||)^{2(1+\delta)} \le C \tag{14.76}$$

for some $C < \infty$, where the first inequality holds by (14.75) with $a = (X_t' m(Z_t, \pi), X_t')'$ and the second inequality holds by Assumptions STAR1(ii) and STAR2(iii).

Similarly, by (7.14),

$$E_{\gamma_0} \sup_{\theta^+ \in \Theta^+} ||\rho_{\theta\theta}^\dagger(W_i, \theta^+)||^{1+\delta} = E_{\gamma_0} \sup_{\theta^+ \in \Theta^+} ||d_t(\pi, \omega) d_t(\pi, \omega)'||^{1+\delta}$$
$$\le E_{\gamma_0} \sup_{\pi \in \Pi} (2||X_t|| + ||X_t|| \cdot ||m_\pi(Z_t, \pi)||)^{2(1+\delta)} \le C \tag{14.77}$$

for some $C < \infty$, where the first inequality holds by (14.75) with $a = (X_t' m(Z_t, \pi), X_t', \omega' X_t m_\pi(Z_t, \pi))'$ and the second inequality holds by Assumption STAR2.

28

By (7.14),

$$
E_{\gamma_0} \sup_{\theta^+ \in \Theta^+} ||\varepsilon(W_i, \theta^+)||^q \leq E_{\gamma_0} \sup_{\theta^+ \in \Theta^+} ||U_t(\theta^+)||^{2q}
$$

$$
\times E_{\gamma_0} \sup_{\pi \in \Pi} (2||X_t|| \cdot ||m_\pi(Z_t, \pi)|| + ||X_t|| \cdot ||m_{\pi\pi}(Z_t, \pi)||)^{2q} \leq C \qquad (14.78)
$$

for some $C < \infty$, where the first inequality holds by the Cauchy-Schwarz inequality and the inequality $||A|| \leq \sum_{i,j} ||A_{i,j}||$ for any matrix A, where $A_{i,j}$ denotes an element of A, and the second inequality follows from (14.72), Holder's inequality, and Assumptions STAR1(ii) and STAR2(iii).

Finally, $E_{\gamma_0}(M_1(W_t) + M_2(W_t)^q) \leq C$ for some $C < \infty$ by Holder's inequality, (14.59), (14.63), (14.67), (14.69), (14.70), and Assumptions STAR(ii) and STAR2(iii).

This completes the verification of Assumption S3(iii) (vector β).

14.5. Verification of Assumptions S3(iv) and S3(v)

To verify Assumption S3(iv), note that

$$
E_{\gamma_0} \rho_{\psi\psi}(W_t, \psi_0, \pi) = E_{\gamma_0} d_{\psi,t}(\pi) d_{\psi,t}(\pi)' \text{ and}
$$
$$
E_{\gamma_0} \rho_{\theta\theta}^\dagger(W_t, \theta_0) = E_{\gamma_0} d_t(\pi_0, \omega_0) d_t(\pi_0, \omega_0)'. \qquad (14.79)
$$

For any $\lambda = (\lambda_1, \lambda_2) \neq 0$, $\lambda_1, \lambda_2 \in R^{d_\beta}$, and $\forall \pi \in \Pi$,

$$
\lambda' E_{\gamma_0} d_{\psi,t}(\pi) d_{\psi,t}(\pi)' \lambda = E_{\gamma_0} \left(\lambda_1' X_t m(Z_t, \pi) + \lambda_2' X_t \right)^2 > 0, \qquad (14.80)
$$

where the inequality holds by Assumption STAR2(i). This implies that $E_{\gamma_0} d_{\psi,t}(\pi) d_{\psi,t}(\pi)'$ is positive definite $\forall \pi \in \Pi$.

For any $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4) \neq 0$, $\lambda_1, \lambda_2 \in R^{d_\beta}$, $\lambda_3, \lambda_4 \in R$, $\forall \omega$ with $||\omega|| = 1$ and $\forall \pi \in \Pi$,

$$
\lambda' E_{\gamma_0} d_t(\pi, \omega) d_t(\pi, \omega)' \lambda \qquad (14.81)
$$
$$
= E_{\gamma_0} (\lambda_1' X_t m(Z_t, \pi) + \lambda_2' X_t + \lambda_3 \omega' X_t m_{\pi,1}(Z_t, \pi) + \lambda_4 \omega' X_t m_{\pi,2}(Z_t, \pi))^2 > 0,
$$

where the inequality holds by Assumption STAR2(ii) with $a = (\lambda_1, \lambda_2, \lambda_3\omega, \lambda_4\omega)$. Note that $\lambda \neq 0$ implies that $a \neq 0$. The inequality in (14.81) implies that $E_{\gamma_0} \rho_{\theta\theta}^\dagger(W_t, \theta_0)$ is positive definite $\forall \gamma_0 \in \Gamma$.

29

To verify Assumption S3(v), note that $\forall m \neq 0$,

$$Cov_\phi(\rho_\theta^\dagger(W_t, \theta_0), \rho_\theta^\dagger(W_{t+m}, \theta_0)) = E_{\gamma_0} U_t U_{t+m} d_t(\pi_0, \omega_0) d_{t+m}(\omega_0, \pi_0)' = 0 \qquad (14.82)$$

by Assumption STAR1(i). This yields that

$$\begin{aligned} V^\dagger(\theta_0, \theta_0; \gamma_0) &= Cov_{\phi_0} U_t^2 d_t(\pi_0, \omega_0) d_t(\pi_0, \omega_0)' \\ &= E_{\gamma_0} U_t^2 d_t(\pi_0, \omega_0) d_t(\pi_0, \omega_0)', \end{aligned} \qquad (14.83)$$

where the second equality uses $E_{\gamma_0} U_t d_t(\pi_0, \omega_0) = 0$ by Assumption STAR1(i). The matrix $E_{\gamma_0} U_t^2 d_t(\pi_0, \omega_0) d_t(\pi_0, \omega_0)'$ is positive definite by the argument in (14.81) with $d_{\psi,t}(\pi)$ replaced by $U_t d_t(\pi, \omega)$ and using $E_{\gamma_0}(U_t^2 | \mathcal{F}_{t-1}) = \sigma^2 > 0$.

14.6. Verification of Assumption S4

To verify Assumption S4, we have

$$\begin{aligned} E_{\gamma_0} \rho_\psi(W_t, \theta) &= -E_{\gamma_0} U_t(\theta) d_{\psi,t}(\pi) \\ &= -E_{\gamma_0}(U_t + X_t'(\zeta_0 - \zeta) + X_t'[\beta_0 m(Z_t, \pi_0) - \beta m(Z_t, \pi)]) d_{\psi,t}(\pi) \text{ and} \\ K(\theta; \gamma_0) &= -E_{\gamma_0} d_{\psi,t}(\pi) X_t' m(Z_t, \pi_0) \\ &= -E_{\gamma_0} d_{\psi,t}(\pi) d_{\psi,t}(\pi_0)' \cdot S_\beta'. \end{aligned} \qquad (14.84)$$

where $S_\beta = [I_{d_\beta} : 0] \in R^{d_\beta \times (2d_\beta)}$.

Assumption S4(i) holds with $K(\theta; \gamma_0)$ in (14.84) by the moment conditions in Assumption STAR2(iii). To verify Assumption S4(ii), we need to show that $E_{\gamma_0} d_{\psi,t}(\pi) d_{\psi,t}(\pi_0)'$ is continuous in π, π_0, and ϕ. Continuity in π and π_0 follows from the the continuity of $m(Z_t, \pi)$ in π and the moment conditions in Assumption STAR2(iii). Continuity in ϕ holds because $\phi_n \to \phi_0$ under d_Φ implies weak convergence of (Y_t, Y_{t+m}) for all $t, m \geq 1$, which in turn implies the convergence of $E_{\gamma_n} d_{\psi,t}(\pi) d_{\psi,t}(\pi_0)'$ to $E_{\gamma_0} d_{\psi,t}(\pi) d_{\psi,t}(\pi_0)'$ by the moment conditions in Assumption STAR2(iii).

The continuity in π, π_0, and ϕ holds uniformly over $\pi \in \Pi$ by Lemma 9.2 using (i) the pointwise convergence above, (ii) the fact that $E_{\gamma_0} d_{\psi,t}(\pi) d_{\psi,t}(\pi_0)'$ is differentiable in π and the partial derivative is bounded over $\pi \in \Pi$, and (iii) the compactness of Π. This completes the verification of Assumption S4.

14.7. Verification of Assumptions B1 and B2

Now we verify Assumptions B1 and B2. Assumptions B1(i) and B1(iii) hold by Assumptions STAR5(i) and STAR5(ii) immediately. Assumption B1(ii) holds with $\mathcal{Z}^0 = int(\mathcal{Z}_0)$ by Assumptions STAR4(iv) and STAR5(iii). Assumption B2(i) holds immediately because the true parameter space Γ is of the form in (2.6) and Γ is assumed to be compact. Assumption B2(ii) holds by Assumption STAR4(ii). Assumption B2(iii) holds by Assumption STAR4(iv) and the form of the true parameter space in (7.10).

14.8. Verification of Assumptions C6 and C7

Assumption C6 is implied by Assumption STAR3(i).

Now we verify Assumption C7 with $H(\pi; \gamma_0)$ and $K(\pi; \gamma_0)$ given in (7.9). By the matrix Cauchy-Schwarz inequality in Tripathi (1999),

$$K(\pi; \gamma_0)'H^{-1}(\pi; \gamma_0)K(\pi; \gamma_0) \leq E_{\gamma_0} X_t X_t' m^2(Z_t, \pi_0). \qquad (14.85)$$

The matrix "$\leq$" holds as an equality if and only if $X_t m(Z_t, \pi_0)a + (X_t', X_t'm(Z_t, \pi))c = 0$ with probability 1 for some $a \in R^{d_\beta}$ and $c \in R^{2d_\beta}$ with $(a', c')' \neq 0$. The "$\leq$" holds as an equality uniquely at $\pi = \pi_0$ by Assumption STAR2(i).

Proof of Lemma 7.1. We prove Lemma 7.1 by verifying Assumption C6† and using Lemma 3.2. Note that

$$\begin{aligned} \rho_\beta(W_t, \psi_0, \pi) &= U_t X_t m(Z_t, \pi), \\ \rho_\zeta(W_t, \psi_0, \pi) &= U_t X_t, \\ \rho_\psi^*(W_i, \psi_0, \pi_1, \pi_2) &= U_t d_\psi^*(\pi_1, \pi_2), \text{ where} \\ d_\psi^*(\pi_1, \pi_2) &= (X_t'm(Z_t, \pi_1), X_t'm(Z_t, \pi_2), X_t')'. \end{aligned} \qquad (14.86)$$

The matrix $\Omega_G(\pi_1, \pi_2; \gamma_0)$ that appears in Assumption C6† takes the form

$$\Omega_G(\pi_1, \pi_2; \gamma_0) = E_{\gamma_0} U_t^2 d_\psi^*(\pi_1, \pi_2)d_\psi^*(\pi_1, \pi_2)' \qquad (14.87)$$

by Assumption STAR1(i). Assumption C6†(ii) holds by Assumption STAR2(i) and $E_{\gamma_0}(U_t^2|\mathcal{F}_{t-1}) = \sigma^2 > 0$ using arguments analogous to those in (14.81). $\square$

14.9. Verification of Assumptions V1 (vector β) and V2

Here we verify Assumptions V1 (vector β) and V2, which are stated in Appendix B of the main paper.

In the STAR model, Assumption V1(i) holds with

$$
\begin{aligned}
&J(\theta^+; \gamma_0) = E_{\gamma_0} d_t\left(\pi, \omega\right) d_t(\pi, \omega)' \text{ and} \\
&V(\theta^+; \gamma_0) = E_{\gamma_0} U_t^2 d_t\left(\pi, \omega\right) d_t(\pi, \omega)' \\
&+ E_{\gamma_0} [X_t'(\zeta_0 - \zeta) + X_t'(||\beta_0||\omega_0 m(Z_t, \pi_0) - ||\beta||\omega m(Z_t, \pi))]^2 d_t(\pi, \omega) d_t(\pi, \omega)',
\end{aligned}
\tag{14.88}
$$

by the uniform law of large numbers in Lemma 9.3.

Assumption V1(ii) holds by the continuity of $m(z, \pi)$ and $m_\pi(z, \pi)$ in π and Assumption STAR2(iii).

To verify Assumption V1(iii), note that $\Sigma(\pi, \omega; \gamma_0)$ takes the form

$$
\begin{aligned}
&\Sigma(\pi, \omega; \gamma_0) \\
&= \left(E_{\gamma_0} d_t\left(\pi, \omega\right) d_t(\pi, \omega)'\right)^{-1} E_{\gamma_0} U_t^2 d_t\left(\pi, \omega\right) d_t(\pi, \omega)' \left(E_{\gamma_0} d_t\left(\pi, \omega\right) d_t(\pi, \omega)'\right)^{-1}.
\end{aligned}
\tag{14.89}
$$

Given that $E_{\gamma_0} d_t\left(\pi, \omega\right) d_t(\pi, \omega)'$ and $E_{\gamma_0} U_t^2 d_t\left(\pi, \omega\right) d_t(\pi, \omega)'$ are both positive definite, $\Sigma(\pi, \omega; \gamma_0)$ is positive definite $\forall \pi \in \Pi$ and $\forall \omega$ with $||\omega|| = 1$.

Because the determinant of $E_{\gamma_0} d_t\left(\pi, \omega\right) d_t(\pi, \omega)'$ is bounded away from 0 as a function of (π, ω) $\forall \gamma_0 \in \Gamma$ and $||E_{\gamma_0} d_t\left(\pi, \omega\right) d_t(\pi, \omega)'|| \leq C_1$ for some $C_1 < \infty$ $\forall \gamma_0 \in \Gamma$ by Assumption STAR2(iii), we have $|| \left(E_{\gamma_0} d_t\left(\pi, \omega\right) d_t(\pi, \omega)'\right)^{-1} || \leq C_2$ for some $C_2 < \infty$. Hence, $||\Sigma(\pi, \omega; \gamma_0)|| \leq C$ $\forall \pi \in \Pi$ and $\forall \omega$ with $||\omega|| = 1$. This completes the verification of Assumption V1(iii).

Assumption V1(iv) holds by Assumption STAR3(ii).

Assumptions V1(i) and V1(ii) hold not only under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, but also under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ in this example. This and $\widehat{\theta}_n \to_p \theta_0$ under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, which holds by Lemma 5.3 of AC1, imply that Assumption V2 holds. Regarding the assumptions employed in Lemma 5.3 of AC1, Assumptions B1, B2, C6, and C7 are verified above and Assumptions A, B3, and C1-C4 hold by Lemma 9.1 under Assumptions B1, B2, and S1-S4. □

References

Andrews, D. W. K., Cheng, X., 2007. Estimation and inference with weak, semi-strong, and strong identification. Cowles Foundation Discussion Paper No. 1773R, Yale University.

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