



Penn Institute for Economic Research
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6297
pier@econ.upenn.edu
<http://economics.sas.upenn.edu/pier>

PIER Working Paper 14-018

“Uniform Inference in Nonlinear Models
with Mixed Identification Strength”

by

Xu Cheng

<http://ssrn.com/abstract=2435179>

Uniform Inference in Nonlinear Models with Mixed Identification Strength*

Xu Cheng[†]

Department of Economics
University of Pennsylvania

May 8, 2014

Abstract

The paper studies inference in nonlinear models where identification loss presents in multiple parts of the parameter space. For uniform inference, we develop a local limit theory that models mixed identification strength. Building on this non-standard asymptotic approximation, we suggest robust tests and confidence intervals in the presence of non-identified and weakly identified nuisance parameters. In particular, this covers applications where some nuisance parameters are non-identified under the null (Davies (1977, 1987)) and some nuisance parameters are subject to a full range of identification strength. The asymptotic results involve both inconsistent estimators that depend on a localization parameter and consistent estimators with different rates of convergence. A sequential argument is used to peel the criterion function based on identification strength of the parameters. The robust test is uniformly valid and non-conservative.

Keywords: Mixed rates, nonlinear regression, robust inference, uniformity, weak identification.

JEL Codes: C12, C15

*This paper supersedes a working paper previously circulated under the title “Robust Confidence Intervals in Nonlinear Regression under Weak Identification”.

[†]Email: xucheng@econ.upenn.edu. Address: 3718 Locust Walk, Philadelphia, PA, 19104.

1 Introduction

Economic theory and empirical studies often suggest nonlinear relationships among economic variables. These relationships are commonly specified in a parametric form involving several nonlinear component functions with unknown transformation parameters and loading coefficients that measure the importance of each component. Inference in such nonlinear models are non-standard due to loss of identification in multiple areas of the parameter space. This paper investigates inference in an additive nonlinear regression model that takes the form

$$Y_t = \sum_{j=1}^p g_j(X_t, \pi_j) \beta_j + Z_t' \zeta + U_t, \quad (1.1)$$

where the nonlinear function $g_j(\cdot, \pi_j)$ is known up to π_j for $j = 1, \dots, p$. For example, in a multiple regime smooth transition model, the unknown parameter π_j characterizes the transition from one regime to the next via an exponential or logistic transformation and the null hypothesis $H_0 : \beta_p = 0$ reduces p regimes to $p - 1$ regimes.¹ Many other types of nonlinear transformations are discussed in Hansen (1996). In this nonlinear model, π_j is not identified if $\beta_j = 0$ for $j = 1, \dots, p$, which yields p different sources of identification failures. In finite-sample estimation, π_j is weakly identified when β_j is close to 0 and the identification strength of π_j varies with the magnitude of β_j . Such non/weak identification from each nonlinear component is spilled over to the estimation of all unknown parameters.

Mixed identification strength brings new challenges to uniform inference. Take the test $H_0 : \beta_p = 0$ for example. In addition to the non-identification of π_p under the null hypothesis, the nuisance parameters π_j for $j = 1, \dots, p - 1$ could be non-identified, weakly identified, or strongly identified, depending on the unknown value of β_j . In consequence, this is a non-standard test that is different from the problem investigated in Davies (1977, 1987), Andrews and Ploberger (1994), and Hansen (1996), among others. These classical results apply to testing the null hypothesis $H_0 : \beta = (\beta_1, \dots, \beta_p)' = 0$, where the nuisance parameter $\pi = (\pi_1, \dots, \pi_p)'$ is non-identified. When the interest is in a sub-vector of β rather than the full vector, a uniformly valid test has not been studied in the literature.

This paper studies uniform inference for sub-vectors of $\theta = (\beta', \zeta', \pi')'$. The result not only covers the test $H_0 : \beta_p = 0$, but also applies to any linear functions of θ and applies to both tests and confidence sets. For confidence set construction, Andrews and Cheng (2012) consider a broad class of models where non-identification occurs at a single point of the parameter space,

¹See van Dijk, Terasvirta, and Franses (2002) for a review of the smooth transition autoregressive model.

including the model in (1.1) with $p = 1$. The main challenge in this paper is the multiple sources of non/weak identification when $p > 1$, as illustrated by the test $H_0 : \beta_p = 0$. As in Andrews and Cheng (2012), this paper considers a full range of identification strength around the crucial point where identification is lost. However, when the number of such crucial points increases from one to multiple, a new asymptotic theory is required for uniform inference with mixed identification strength.

The paper derives a local limit theory for the least squares estimator and the Wald statistic when β_j converges to 0 at various rates or is bounded away from 0. A faster convergence rate models weaker identification strength of π_j . This is analogous to the model of weak instruments in Staiger and Stock (1997) and Stock and Wright (2000). Because the identification strength is unknown, all convergence rates and all combinations across $j = 1, \dots, p$ are considered for uniform inference, following the approach in Andrews and Guggenberger (2009a, 2010).

The main technical innovation of the paper is the use of sequential arguments to develop the asymptotic theory for estimators and test statistics in the presence of mixed identification strength. This asymptotic theory allows for the coexistence of both inconsistent estimators and consistent estimators with different rates of convergence. To implement the sequential arguments, we first concentrate out the loading coefficients β and ζ , which are always strongly identified, then group the nonlinear parameters π_j based on their identification strength. Starting from the most strongly identified group to the most weakly identified group, the sequential procedure concentrates out one group at a time. The most weakly identified group involves inconsistent estimators that are functionals of chi-square processes. The rate of convergence of consistent estimators are derived in a sequential manner. Finally, the process is reversed by plugging the most weakly identified group to other groups and the test statistics. Uniformly valid tests and confidence sets are suggested based on these non-standard asymptotic distributions.

The asymptotic theory in this paper complements the mixed-rate results developed in Lee (2005, 2010), Radchenko (2008), and Antoine and Renault (2012). In particular, a rotation akin to that in Antoine and Renault (2012) is used to develop the asymptotic distribution of the Wald statistic. The asymptotic results also relate to those considered for near weak instruments by Hahn and Kuersteiner (2002), Caner (2010), and Antoine and Renault (2009). In addition, mixed-rate results have a long history for non-stationary time series, such as Phillips and Park (1988), Sims, Stock, and Watson (1990), Kitamura and Phillips (1997), just to name a few. Different from these papers, the present problem is tied to loss of identification and it involves both inconsistent estimators and consistent estimators with different rates of convergence. The Wald statistic does not always have

an asymptotic chi-square distribution. Furthermore, a different proof strategy based on sequential peeling is used for the identification problem at hand.

There is growing interest in robust inference with weakly identified nuisance parameters. The projection method is studied in Dufour and Taamouti (2005, 2007). Recent development with weakly identified nuisance parameters include Chaudhuri and Zivot (2011), Andrews and Cheng (2012, 2013, 2014), Guggenberger, Kleibergen, Mavroeidis, and Chen (2012), Andrews and Mikusheva (2012, 2013), Chen, Ponomareva, and Tamer (2013), among others. In the nonlinear model considered in this paper, the direction of weak identification is known. Making use of this structure, we propose a robust and non-conservative test in the presence of multiple weakly identified nuisance parameters. In a general nonlinear model without such knowledge, the geometric approach in Andrews and Mikusheva (2013) provides an informative robust test.

This paper also broadly relates to many other papers on non-identification and weak identification. The weak instrument literature is related to the weak identification considered in the present paper, e.g., see Nelson and Startz (1990), Dufour (1997), Staiger and Stock (1997), Stock and Wright (2000), Kleibergen (2002, 2005), Moreira (2003), Guggenberger and Smith (2005), Andrews, Moreira, and Stock (2006), Olea (2013), Andrews (2013), and other papers referenced in Andrews and Stock (2007). Guerron-Quintana, Inoue, and Kilian (2013), Andrews and Mikusheva (2012, 2013), Qu (2013) consider weak identification in DSGE models, an important issue discussed in Schorfheide (2013). Nelson and Startz (2007) introduce the zero-information-limit condition, which applies to the models considered in this paper. Ma and Nelson (2010) consider tests based on linearization for nonlinear models under weak identification. Sargan (1983), Phillips (1989), and Choi and Phillips (1992) study simultaneous equations models where some parameters are unidentified. Shi and Phillips (2012) consider weak identification with integrated regressors.

The rest of the paper is organized as follows. Section 2 introduces the drifting sequences of true parameters used to model mixed identification strength. Sections 3 and 4 develop the asymptotic distributions of the least squares estimator and the Wald and t statistic under mixed identification strength. Section 5 proposes a robust test based on this non-standard asymptotic distribution. This robust test has correct asymptotic size and it is as efficient as the standard test under strong identification. Proofs are collected in the Appendix.

2 Uniformity and Drifting Sequences of Distributions

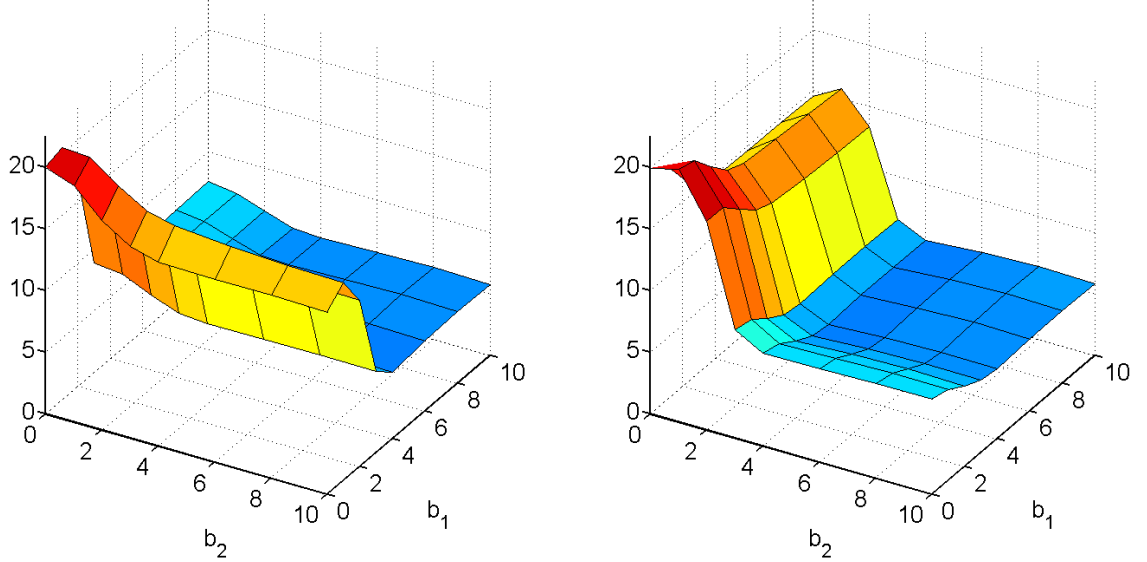
We are interested in a sub-vector of θ , denoted by $R\theta$, where the matrix R has full rank $d_r \leq d_\theta$. The true value of θ belongs to a set Θ^* , which includes a neighborhood around $\beta = 0$. Thus, the area where non/weak identification occurs is part of the parameter space. For a fixed value of v , we test the null hypothesis $H_0 : R\theta = v$ using the test statistic $T_n(R)$ and a critical value $c_{n,1-\alpha}(v)$, where α is the nominal size. This notation allows $c_{n,1-\alpha}(v)$ to depend on both the sample size and the null value, although in standard scenarios it typically is the $1 - \alpha$ quantile of a chi-square distribution or standard normal distribution. A nominal $1 - \alpha$ confidence set for $R\theta$ is $CS_n = \{v : T_n(R) \leq c_{n,1-\alpha}(v)\}$, obtained by inverting tests.

Without knowing the true parameters, we aim to control the maximum null rejection probability of a test over all true parameters consistent with the null, called the finite-sample size of a test. To this end, a reliable critical value should be based on a uniform approximation of the distribution of $T_n(R)$ over the parameter space. However, standard asymptotic results developed under strong identification fail to do so. To illustrate this uniformity issue, Figure 1 takes a simple model with $p = 2$ and plots the finite-sample ($n = 500$) rejection probability of the standard two-sided t test for different true values of β_1 and β_2 . The data generating process (DGP) is specified below where the robust test is introduced and more simulation results are reported. This figure confirms that the standard approximation can be excellent for some true parameters but poor for the rest. Furthermore, the area where standard approximation fails does not disappear even for large samples.

The lack of uniformity also applies to approximations by some non-standard distributions. Use the simple model $p = 2$ for example. To test the null hypothesis $H_0 : \beta_2 = 0$, a non-standard approximation is required due to the loss of identification of π_2 . However, the non-standard distribution that works well for large β_1 may work poorly when β_1 is close to 0. Figure 1 demonstrates that, even when the true value of β_2 is fixed at 0, the distribution of the t statistics vary with the true value of β_1 . To obtain a valid test for $H_0 : \beta_2 = 0$, we should consider all possible identification strength of π_1 as well as the non-identification of π_2 .

To better approximate the finite-sample distribution of the test statistic $T_n(R)$, we consider alternative asymptotic approximations along drifting sequences of true parameters. Let $\beta_{j,n}$ denote the true value of β_j for sample size n , for $j = 1, \dots, p$. Due to the nonlinear structure of the model, π_j is strongly identified only if $\beta_{j,n} \rightarrow \beta_{j,0} \neq 0$. For the rest, the rate at which $\{\beta_{j,n} : n \geq 1\}$ converges to 0 models the identification strength of π_j . To achieve a uniform approximation, we

Figure 1: Standard Two-Sided t Test: Finite-Sample Rejection Probability ($\times 100$) for $H_0 : \beta_1 = \beta_{1,0}$ (left) and $H_0 : \beta_2 = \beta_{2,0}$ (right).



Notes: nominal size $\alpha = 5\%$, sample size $n = 500$, the true values of β_1 and β_2 are $\beta_{1,0} = b_1/\sqrt{500}$ and $\beta_{2,0} = b_2/\sqrt{500}$ for b_1 and b_2 in the X axis and Y axis of the figure.

consider sequences of $\beta_{j,n}$ for $j = 1, \dots, p$ that satisfy one of the following conditions:

- (i) $\beta_{j,n} \rightarrow 0$, $n^{1/2}\beta_{j,n} \rightarrow b_j \in R$, (weak identification) or
- (ii) $\beta_{j,n} \rightarrow 0$, $n^{1/2}|\beta_{j,n}| \rightarrow \infty$, (semi-strong identification) or
- (iii) $\beta_{j,n} \rightarrow \beta_{j,0} \neq 0$ (strong identification). (2.1)

In addition, $\lim_{n \rightarrow \infty} \beta_{j,n}/\beta_{j',n} \in R \cup \{\pm\infty\}$ for sequences in (ii) and (iii)². Following the terminology in Andrews and Cheng (2012), the sequences in (i), (ii), (iii) are associated with weak, semi-strong, and strong identification of π_j , respectively. The semi-strong identification case provides an important link between the two extreme cases and it is crucial for uniform results. In the rest of the paper, we first develop asymptotic distributions of estimators and test statistics along these drifting true parameters, under which the p nonlinear regressors are categorized into different identification groups. The grouping rule is specified in Section 3.1 below. In particular, the semi-strong identification category is further divided into different groups based on the rate at which $\beta_{j,n}$ converges to 0. In practice, the group specification depends on the true parameters and is

²Without loss of generality, we assume $\beta_{j,n} \neq 0 \forall n$ for sequences in (ii) and (iii).

unknown. We show that the class of asymptotic approximations along *all* group specifications is sufficiently large to yield a uniform approximation of the finite-sample size of a test.

3 Asymptotic Distributions of Estimators

The observations $\{W_t = (Y_t, X_t', Z_t')' : t \leq n\}$ are independent and identically distributed (i.i.d.) or strictly stationary. We assume U_t has zero mean conditional on X_t and Z_t . The true value of θ belongs to the set $\Theta^* = \mathcal{B}_1^* \times \cdots \times \mathcal{B}_p^* \times \mathcal{Z}^* \times \Pi^*$, where \mathcal{B}_j^* for $j = 1, \dots, p$ is a closed interval that includes both zero and non-zero values. Thus, the area where non/weak identification occurs is part of the parameter space. Both \mathcal{Z}^* and Π^* are compact sets. For any $\theta \in \Theta^*$, the distribution of $\{W_t : t \leq n\}$ is denoted by F_γ for the parameter $\gamma = (\theta, \phi) \in \Gamma$, where $\phi \in \Phi^*$ denotes an infinite-dimensional nuisance parameter that characterizes the distribution. In parametric models, the finite-dimensional parameter θ fully specifies the distribution of the data and ϕ does not exist. Let \mathbb{P}_γ and \mathbb{E}_γ denote the probability and expectation under the distribution indexed by γ .

In addition to the drifting sequences $\{\beta_{j,n} : n \geq 1\}$, we allow other parameters to change with the sample size, following the approach in Andrews and Guggenberger (2009a, 2010). As such, we not only obtain uniform results over $\mathcal{B}_1^* \times \cdots \times \mathcal{B}_p^*$, but also over $\gamma \in \Gamma$. Specifically, for sample size n , the true parameters are

$$\theta_n = (\beta_n', \zeta_n', \pi_n')', \quad \beta_n = (\beta_{1,n}, \dots, \beta_{p,n})', \quad \pi_n = (\pi_{1,n}, \dots, \pi_{p,n})', \quad \text{and } \gamma_n = (\theta_n, \phi_n) \quad (3.1)$$

where $\theta_n \rightarrow \theta_0 = (\beta_0', \zeta_0', \pi_0')'$, $\gamma_n \rightarrow \gamma_0 \in \Gamma$, and the subscript 0 denotes the limit of true values. We consider rescaling $\beta_{j,n}$ as in (2.1) rather than other parameters because the distributions are non-standard only when some elements of β are close to 0.

The least squares sample criterion function³ is

$$Q_n(\theta) = \frac{1}{2n} \sum_{t=1}^n \left(Y_t - \sum_{j=1}^p g_j(X_t, \pi_j) \beta_j \right)^2. \quad (3.2)$$

The least squares estimator $\hat{\theta}_n$ minimizes $Q_n(\theta)$ over $\theta \in \Theta$, where $\Theta = \mathcal{B}_1 \times \cdots \times \mathcal{B}_p \times \mathcal{Z} \times \Pi$, \mathcal{B}_j for $j = 1, \dots, p$ are closed intervals, and \mathcal{Z} and Π are compact sets. To focus on the identification issue rather than the boundary effect, we assume all true values in Θ^* are in the interior of Θ . We

³The constant 1/2 is added to simplify the asymptotic results presented below.

derive asymptotic distributions along sequences of true parameters $\{\gamma_n \in \Gamma : n \geq 1\}$, assuming the following assumptions hold for any $\gamma \in \Gamma$.

Assumption 1. $g_j(x, \pi)$ is twice continuously differentiable with respect to (wrt) π , $\forall \pi \in \Pi$ and any x in its support. We denote the first and second order derivatives of $g_j(x, \pi_j)$ wrt π_j by $g_{\pi_j}(x, \pi_j)$ and $g_{\pi\pi_j}(x, \pi_j)$, respectively. For some non-stochastic function $M_j(x) \in R$, $|g_{\pi\pi_j}(x, \pi_j) - g_{\pi\pi_j}(x, \bar{\pi}_j)| \leq M_j(x)|\pi_j - \bar{\pi}_j|$, $\forall \pi_j, \bar{\pi}_j \in \Pi_j$.

For time series data, the following assumption holds. Let d_θ denote the dimensional of θ .

Assumption 2. (i) $\{W_t : t \geq 1\}$ is a strictly stationary and strong mixing sequence with mixing coefficients $\alpha_m \leq Cm^{-r}$ for some $r > d_\theta q / (q - d_\theta)$ and some $q > d_\theta \geq 2$.

(ii) $\mathbb{E}_\gamma(U_t | \mathcal{F}_{t-1}) = 0$ and $\mathbb{E}_\gamma |U_t|^{2q} \leq C$, where \mathcal{F}_{t-1} is the sigma field to which X_t, Z_t , and U_{t-1} are adapted.

(iii) $\mathbb{E}_\gamma(\sup_{\pi_j \in \Pi_j} [g_j(X_t, \pi_j)^{2q} + g_{\pi_j}(X_t, \pi_j)^{2q} + g_{\pi\pi_j}(X_t, \pi_j)^{2q}] + M_j(X_t)^{2q}) \leq C$, for $j = 1, \dots, p$.

For i.i.d. data, the following assumption holds in place of Assumption 2 for some $\delta > 0$. In the asymptotic results below, we use Assumption 2 to represent both of them.

Assumption 2*. (i) $\{W_t : t \geq 1\}$ is i.i.d.

(ii) $\mathbb{E}_\gamma(U_t | X_t, Z_t) = 0$, $\mathbb{E}_\gamma |U_t|^{4+\delta} \leq C$.

(iii) $\mathbb{E}_\gamma(\sup_{\pi_j \in \Pi_j} [g_j(X_t, \pi_j)^{4+\delta} + g_{\pi_j}(X_t, \pi_j)^{4+\delta} + g_{\pi\pi_j}(X_t, \pi_j)^{4+\delta}] + M_j(X_t)^{4+\delta}) \leq C$, for $j = 1, \dots, p$.

Let $g(X_t, \pi) = (g_1(X_t, \pi_1), \dots, g_p(X_t, \pi_p))' \in R^p$.

Assumption 3. $\forall \pi, \pi_0 \in \Pi$ and some $\varepsilon > 0$, $\mathbb{P}_\gamma(a' [g(X_t, \pi)', g(X_t, \pi_0)']', Z_t] = 0) \leq 1 - \varepsilon$ for any $a \neq 0$ and $\pi \neq \pi_0$.

Assumptions 1 and 2 are standard regularity assumptions on dependence, smoothness, and moment conditions. In subsequent analysis, they are necessary to obtain the uniform law of large numbers (ULLN) and the weak convergence of some empirical processes. Assumption 3 is for the identification of β and ζ and the identification of π when β is different from 0.

3.1 Grouping Rules and Reparameterization

To derive asymptotic results with mixed identification strength, we first group the nonlinear regressors based on the order of magnitude of $\beta_{j,n}$ for $j = 1, \dots, p$. Without loss of generality, we assume $\beta_{j',n} = O(\beta_{j,n}) \forall j' > j$.

The grouping rule is as follows.

- (i) All $\beta_{j,n}$ that have a non-zero limit are put in the first group. If all $\beta_{j,n}$ have zero limits, the first group is empty.
- (ii) All $\beta_{j,n}$ that are $O(n^{-1/2})$ are put in the last group.
- (iii) For those that converge to 0 but at a rate slower than $n^{-1/2}$, members in group k converge to 0 slower than members in group k' for any $k' > k$ and members in the same group converge to 0 at the same rate.

Following this grouping rule, the first group is associated with strong identification, the last group is associated with weak identification, and the middle groups are associated with semi-strong identification, ordered by the rate of convergence. Note that the group index k for $\beta_{j,n}$ is a property associated with the drifting sequence $\{\beta_{j,n} : n \geq 1\}$. Therefore, the group index k does not change with the sample size n . We call k the group index for β_j .

A reparameterization follows the grouping rule. Suppose there are K groups and $\beta_{k_1}, \dots, \beta_{k_{p_k}}$ are the p_k elements in group k . Let

$$\mathcal{I}_k = \{k_1, \dots, k_{p_k}\} \quad (3.3)$$

denote the indices for group k . For example, suppose $\beta_n = (3, 1, n^{-1/4}, n^{-1/3}, 2n^{-1/3}, n^{-1/2}, n^{-1})'$. The group indices are $\mathcal{I}_1 = \{1, 2\}, \mathcal{I}_2 = \{3\}, \mathcal{I}_3 = \{4, 5\}, \mathcal{I}_4 = \{6, 7\}$, and the number of group is $K = 4$.

Following the group indices in (3.3), we use the subscript \mathcal{I}_k to denote a sub-vector associated with group k , e.g.,

$$\beta_{\mathcal{I}_k} = (\beta_{k_1}, \dots, \beta_{k_{p_k}})' \in R^{p_k}. \quad (3.4)$$

For the drifting sequences, $\beta_{\mathcal{I}_k,n}$ denote the true values of $\beta_{\mathcal{I}_k}$ when the sample size is n and $\beta_{\mathcal{I}_k,0}$ denote its limit. The grouping rule implies that

$$\begin{aligned} \text{for different groups:} & \quad \beta_{\mathcal{I}_{k'},n} = o(\|\beta_{\mathcal{I}_k,n}\|) \text{ for } k' > k, \\ \text{for the same group:} & \quad \beta_{\mathcal{I}_k,n} \asymp \|\beta_{\mathcal{I}_k,n}\| \text{ for } k = 1, \dots, K - 1, \end{aligned} \quad (3.5)$$

where \asymp represents convergence at the same rate.⁴ In the presence of weak identification, $\beta_{\mathcal{I}_k} = O(n^{-1/2})$ for $k = K$. If all regressors are in the semi-strong or strong identification category, the second line of (3.5) also applies to $k = K$.

Example. Consider a two-regressor model where $Y_t = \beta_1 g(X_t, \pi_1) + \beta_2 g(X_t, \pi_2) + U_t$.

- (i) If $\beta_{1,n} \rightarrow \beta_{1,0} \neq 0$ and $\beta_{2,n} \rightarrow \beta_{2,0} \neq 0$, $\mathcal{I}_1 = \{1, 2\}$.
- (ii) If $n^{1/2}\beta_{1,n} \rightarrow b_1 \in R$, $n^{1/2}\beta_{2,n} \rightarrow b_2 \in R$, $\mathcal{I}_1 = \emptyset$, $\mathcal{I}_2 = \{1, 2\}$. Here $\mathcal{I}_1 = \emptyset$ because both $\beta_{1,n}$ and $\beta_{2,n}$ have zero limits.
- (iii) If $\beta_{1,n} \rightarrow 0$, $\|n^{1/2}\beta_{1n}\| \rightarrow \infty$, $\beta_{2,n} \asymp \beta_{1,n}$, $\mathcal{I}_1 = \emptyset$ and $\mathcal{I}_2 = \{1, 2\}$.
- (iv) If $\beta_{1,n} \rightarrow \beta_{1,0} \neq 0$ and $\beta_{2,n} \rightarrow 0$, $\mathcal{I}_1 = \{1\}$, $\mathcal{I}_2 = \{2\}$.
- (v) If $\beta_{1,n} \rightarrow 0$, $\beta_{2,n} \rightarrow 0$, $\|n^{1/2}\beta_{1n}\| \rightarrow \infty$, $\beta_{2n}/\beta_{1n} \rightarrow 0$, $\mathcal{I}_1 = \emptyset$, $\mathcal{I}_2 = \{1\}$, $\mathcal{I}_3 = \{2\}$.

In cases (i), (ii), (iii), π_1 and π_2 have the same identification strength. In case (iv) and (v), the identification strength of π_1 and π_2 is mixed. \square

3.2 Sequential Peeling of the Criterion Function

The minimization of the sample criterion function $Q_n(\theta)$ can be viewed in a sequential way. Apply the grouping results and define

$$\pi_{\mathcal{I}_k} = (\pi_{k_1}, \dots, \pi_{k_{p_k}})' \quad (3.6)$$

for group k . The regressors and their derivatives in group k are collected in

$$\begin{aligned} g_k(x, \pi_{\mathcal{I}_k}) &= (g_{k_1}(x, \pi_{k_1}), \dots, g_{k_{p_k}}(\pi_{k_{p_k}}))' \in R^{p_k}, \\ g_{\pi_k}(x, \pi_{\mathcal{I}_k}) &= (g_{\pi_{k_1}}(\pi_{k_1}), \dots, g_{\pi_{k_{p_k}}}(\pi_{k_{p_k}}))' \in R^{p_k}. \end{aligned} \quad (3.7)$$

When analyzing $\pi_{\mathcal{I}_k}$, we use π_{k^-} to denote elements of π in previous groups and π_{k^+} to denote elements of π in subsequent groups, i.e.,

$$\pi_{k^-} = (\pi'_{\mathcal{I}_1}, \dots, \pi'_{\mathcal{I}_{k-1}})' \text{ and } \pi_{k^+} = (\pi'_{\mathcal{I}_{k+1}}, \dots, \pi'_{\mathcal{I}_K})'. \quad (3.8)$$

It follows that $\pi = (\pi'_{k^-}, \pi'_{\mathcal{I}_k}, \pi'_{k^+})'$. The identification strength of $\pi_{k^-}, \pi_{\mathcal{I}_k}, \pi_{k^+}$ is in a decreasing order by definition.

⁴For two sequences of non-zero constants $\{a_n : n \geq 1\}$ and $\{b_n : n \geq 1\}$, $a_n \asymp b_n$ if and only if $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} \neq 0$ and $\lim_{n \rightarrow \infty} \frac{b_n}{a_n} \neq 0$. When a_n is a vector, we write $a_n \asymp b_n$ if the relationship holds for each element of a_n .

According to the grouping rule, $\pi_{\mathcal{I}_1}$ is strongly identified. We put all strongly identified elements of π in this group because they can be analyzed together with β and ζ , which are also strongly identified. The semi-strongly identified and weakly-identified elements of π are analyzed differently using the sequential procedure proposed below. If no elements of π are strongly identified, $\mathcal{I}_1 = \emptyset$ and $\pi_{\mathcal{I}_1}$ disappears.

We now describe the sequential procedure and introduce some notations.

- (i) For $k = 1$, conditional on π_{1+} , minimizing $Q_n(\theta) = Q_n(\beta, \zeta, \pi_{\mathcal{I}_1}, \pi_{1+})$ over β , ζ , and $\pi_{\mathcal{I}_1}$ yields $\hat{\beta}(\pi_{1+})$, $\hat{\zeta}(\pi_{1+})$, and $\hat{\pi}_{\mathcal{I}_1}(\pi_{1+})$. The concentrated criterion function $Q_n(\hat{\beta}(\pi_{1+}), \hat{\zeta}(\pi_{1+}), \hat{\pi}_{\mathcal{I}_1}(\pi_{1+}), \pi_{1+})$ is written as $Q_n^c(\pi_{1+}) = Q_n^c(\pi_{\mathcal{I}_2}, \pi_{2+})$ because $\pi_{1+} = (\pi_{\mathcal{I}_2}, \pi_{2+})$.
- (ii) Continue the procedure for $k = 2, \dots, K - 1$ sequentially. For each k , conditional on π_{k+} , minimize $Q_n^c(\pi_{\mathcal{I}_k}, \pi_{k+})$ over $\pi_{\mathcal{I}_k}$ to obtain $\hat{\pi}_{\mathcal{I}_k}(\pi_{k+})$. Concentrating out $\pi_{\mathcal{I}_k}$, the criterion function $Q_n^c(\hat{\pi}_{\mathcal{I}_k}(\pi_{k+}), \pi_{k+})$ is written as $Q_n^c(\pi_{k+}) = Q_n^c(\pi_{\mathcal{I}_{k+1}}, \pi_{(k+1)+})$.
- (iii) For $k = K$, the criterion function is $Q_n^c(\pi_{\mathcal{I}_K})$ and its minimizer is $\hat{\pi}_{\mathcal{I}_K}$.
- (iv) Reverse the order of the procedure. Sequentially plug in the estimators from $\hat{\pi}_{\mathcal{I}_K}$ to $\hat{\pi}_{\mathcal{I}_2}$, we obtain $\hat{\pi}_{\mathcal{I}_{K-1}} = \hat{\pi}_{\mathcal{I}_{K-1}}(\hat{\pi}_K)$, \dots , $\hat{\pi}_{\mathcal{I}_1} = \hat{\pi}_{\mathcal{I}_1}(\hat{\pi}_{\mathcal{I}_2}, \dots, \hat{\pi}_{\mathcal{I}_K})$, $\hat{\beta} = \hat{\beta}(\hat{\pi}_{\mathcal{I}_2}, \dots, \hat{\pi}_{\mathcal{I}_K})$, and $\hat{\zeta} = \hat{\zeta}(\hat{\pi}_{\mathcal{I}_2}, \dots, \hat{\pi}_{\mathcal{I}_K})$.

This is an equivalent representation of the standard least squares estimator and

$$\hat{\theta} = (\hat{\beta}', \hat{\zeta}', \hat{\pi}'_{\mathcal{I}_1}, \dots, \hat{\pi}'_{\mathcal{I}_K})'. \quad (3.9)$$

This sequential representation is necessary for deriving the asymptotic results with mixed identification strength.

The asymptotic analysis starts with the uniform consistency of the strongly identified parameters. Roughly speaking, the sample criterion function $Q_n(\theta)$ uniformly converges to its population counterpart $Q(\theta)$, which identifies the true values of $\beta, \zeta, \pi_{\mathcal{I}_1}$ but does not depend on π_{1+} because $\beta_{\mathcal{I}_k, n} \rightarrow 0$ for $k > 1$. By an extension of standard arguments for consistency of extremum estimators, we obtain the uniform consistency for the strongly identified parameters.

Lemma 1 (consistency for strong identification groups)

Suppose Assumption 1-3 hold. Then,

$$\sup_{\pi_1^+ \in \Pi_1^+} \left(\|\hat{\zeta}(\pi_{1+}) - \zeta_n\| + \|\hat{\beta}(\pi_{1+}) - \beta_n\| + \|\hat{\pi}_{\mathcal{I}_1}(\pi_{1+}) - \pi_{\mathcal{I}_1, n}\| \right) \rightarrow_p 0.$$

To obtain consistency for the semi-strong identification groups, we analyze the concentrated criterion function $Q_n^c(\pi_{\mathcal{I}_k}, \pi_{k+})$ sequentially for $k = 2, \dots, K-1$. We show that, after proper re-centering and rescaling, $Q_n^c(\pi_{\mathcal{I}_k}, \pi_{k+})$ has a non-degenerate limit that identifies the true value of $\pi_{\mathcal{I}_k}$. This limit is non-degenerate in $\pi_{\mathcal{I}_k}$ but is degenerate in π_{k+} . In consequence, parameters with different identification strength are analyzed sequentially.

Before presenting asymptotic results for the semi-strong identification groups, we first define some notations. Analogous to π_{k-} and π_{k+} , define

$$\beta_{k-} = (\beta'_{\mathcal{I}_1}, \dots, \beta'_{\mathcal{I}_{k-1}})' \in \mathbb{R}^{d_{k-}} \quad \text{and} \quad \beta_{k+} = (\beta'_{\mathcal{I}_{k+1}}, \dots, \beta'_{\mathcal{I}_K})' \in \mathbb{R}^{d_{k+}}, \quad (3.10)$$

which are associated with the coefficients before and after $\beta_{\mathcal{I}_k}$. When analyzing $Q_n^c(\pi_{\mathcal{I}_k}, \pi_{k+})$, the parameters that have been concentrated out are collected in

$$\psi_{k-} = (\beta', \zeta', \pi'_{k-})'. \quad (3.11)$$

The true value of ψ_{k-} is denoted by $\psi_{k-,n}$. Let $\widehat{\psi}_{k-}(\pi_k, \pi_{k+})$ denote the estimator of ψ_{k-} conditional on (π_k, π_{k+}) . Following the description of the sequential procedure, $Q_n^c(\pi_{\mathcal{I}_k}, \pi_{k+}) = Q_n(\widehat{\psi}_{k-}(\pi_k, \pi_{k+}), \pi_{\mathcal{I}_k}, \pi_{k+})$.

Define

$$\psi_{k-,n}^0 = (\beta'_{k-,n}, \beta_k^{0'}, \beta_{k+}^{0'}, \zeta'_n, \pi'_{k-,n}), \quad \text{with } \beta_k^0 = 0 \text{ and } \beta_{k+}^0 = 0. \quad (3.12)$$

Note that the difference between $\psi_{k-,n}$ and $\psi_{k-,n}^0$ lies in β_k and β_{k+} . To derive the asymptotic distribution of the concentrated criterion function, $Q_n(\widehat{\psi}_{k-}(\pi_k, \pi_{k+}), \pi_{\mathcal{I}_k}, \pi_{k+})$ is centered around $Q_n(\psi_{k-,n}^0, \pi_{\mathcal{I}_k}, \pi_{k+})$. We set $\beta_k^0 = 0$ and $\beta_{k+}^0 = 0$ in $\psi_{k-,n}^0$ so that the centering term $Q_n(\psi_{k-,n}^0, \pi_{\mathcal{I}_k}, \pi_{k+})$ does not depend on (π_k, π_{k+}) . To make it clear, $Q_n(\psi_{k-,n}^0, \pi_{\mathcal{I}_k}, \pi_{k+})$ is abbreviated to $Q_n(\psi_{k-,n}^0)$.

Define a derivative vector

$$d_{\psi_k,t}(\pi) = (g(X_t, \pi)', Z'_t, g_{\pi_{k-}}(X_t, \pi_{k-})')', \quad (3.13)$$

where $g_{\pi_{k-}}(X_t, \pi_{k-}) = (g_{\pi_1}(X_t, \pi_{\mathcal{I}_k})', \dots, g_{\pi_{k-1}}(X_t, \pi_{\mathcal{I}_{k-1}})')'$ is a collection of the first order derivative in groups 1 to $k-1$. For any $\pi_{\mathcal{I}_k}, \widetilde{\pi}_{\mathcal{I}_k} \in \Pi_{\mathcal{I}_k}$, define a covariance matrix

$$H_k(\pi_{\mathcal{I}_k}, \widetilde{\pi}_{\mathcal{I}_k} | \pi_{k+}) = \mathbb{E}_{\gamma_0} d_{\psi_k,t}(\pi_{k-,0}, \pi_{\mathcal{I}_k}, \pi_{k+}) d_{\psi_k,t}(\pi_{k-,0}, \widetilde{\pi}_{\mathcal{I}_k}, \pi_{k+})', \quad (3.14)$$

where the subscript 0 denotes the limit of the true value as $n \rightarrow \infty$.

Assumption 4. $\lambda_{\min}(H_k(\pi_{\mathcal{I}_k}, \pi_{\mathcal{I}_k} | \pi_{k+})) \geq \varepsilon$ for some $\varepsilon > 0 \forall \pi_{\mathcal{I}_k} \in \Pi_{\mathcal{I}_k}, \pi_{k+} \in \Pi_{\mathcal{I}_{k+}}$ and $k = 1, \dots, K$.

The following Lemma establishes consistency for the semi-strong identification groups using the limit of $Q_n^c(\pi_{\mathcal{I}_k}, \pi_{k+})$. This Lemma is proved by induction. In step k , part (a) of the Lemma is used to show the consistency in part (b) and the rate of convergence in part (c). The latter two in turn are used to obtain part (a) for step $k + 1$. Let d_β, d_ζ , and d_{k-} denote the dimensions of β, ζ , and β_{k-} .

Lemma 2 (consistency for semi-strong identification groups by induction)

Suppose Assumptions 1-4 hold. Then, for $k = 2, \dots, K - 1$,

(a) the concentrated sample criterion function satisfies

$$\begin{aligned} & \|\beta_{\mathcal{I}_k, n}\|^{-2} \left(Q_n^c(\pi_{\mathcal{I}_k}, \pi_{k+}) - Q_n(\psi_{k-, n}^0) \right) \\ \rightarrow_p & -\frac{1}{2} \Delta_k' H_k(\pi_{\mathcal{I}_k}, \pi_{\mathcal{I}_{k,0}} | \pi_{k+})' [H_k(\pi_{\mathcal{I}_k}, \pi_{\mathcal{I}_k} | \pi_{k+})]^{-1} H_k(\pi_{\mathcal{I}_k}, \pi_{\mathcal{I}_{k,0}} | \pi_{k+}) \Delta_k, \end{aligned}$$

where $\Delta_k = (0_{1 \times d_{k-}}, \omega_{k,0}', 0_{1 \times (d_\zeta + d_{k-})})'$ and $\omega_{k,0} = \lim_{n \rightarrow \infty} \beta_{\mathcal{I}_k, n} / \|\beta_{\mathcal{I}_k, n}\|$ is the angel parameter;

(b) the estimator of $\pi_{\mathcal{I}_k}$ satisfies

$$\sup_{\pi_{k+} \in \Pi_{k+}} \|\widehat{\pi}_{\mathcal{I}_k}(\pi_{k+}) - \pi_{\mathcal{I}_k, n}\| \rightarrow_p 0;$$

(c) the estimator of $\psi_{k-} = (\beta', \zeta', \pi_{\mathcal{I}_1}', \dots, \pi_{\mathcal{I}_{k-1}}')$ satisfies

$$\|\beta_{\mathcal{I}_k, n}\|^{-1} \begin{pmatrix} \widehat{\beta}_{k-}(\pi_{k+}) - \beta_{k-, n} \\ \widehat{\beta}_{\mathcal{I}_k}(\pi_{k+}) - \beta_{\mathcal{I}_k, n} \\ \widehat{\beta}_{k+}(\pi_{k+}) \\ \widehat{\zeta} - \zeta_n \\ \text{diag}\{\beta_{k-, n}\}(\widehat{\pi}_{k-}(\pi_{k+}) - \pi_{k-, n}) \end{pmatrix} \rightarrow_p 0.$$

Comments. 1. Part (a) is obtained by a quadratic expansion of $Q_n(\widehat{\psi}_{k-}(\pi_k, \pi_{k+}), \pi_{\mathcal{I}_k}, \pi_{k+})$ around the centering term $Q_n(\psi_{k-, n}^0)$. This expansion relies on the consistency of $\widehat{\psi}_{k-}(\pi_k, \pi_{k+})$, which follows from Lemma 1 and part (b) up to step $k - 1$.

2. This quadratic expansion has some non-standard features. First, the expansion is around $\psi_{k-, n}^0$ instead of the true value of ψ_{k-} . The choice of $\psi_{k-, n}^0$ ensures that the left hand side of part (a)

is minimized by $\widehat{\pi}_{\mathcal{I}_k}(\pi_{k+})$. The right hand side of part (a) is uniquely minimized at $\pi_{\mathcal{I}_k} = \pi_{\mathcal{I}_k,0}$ by a matrix Cauchy-Schwarz inequality. Therefore, the argmax continuous mapping theorem (Theorem 3.2.2 in van der Vaart and Wellner (1996)) gives consistency in part (b). Second, in this quadratic expansion, both the first and second order derivatives have mixed rate of convergence.

3. Part (c) provides the rate of convergence of $\widehat{\psi}_{k-}(\widehat{\pi}_{\mathcal{I}_k}(\pi_{k+}), \pi_{k+})$, which is crucial for deriving the asymptotic distribution in part (a) for step $k + 1$. As k gets larger, the rate of convergence $\|\beta_{\mathcal{I}_k,n}\|^{-1}$ also gets faster and this rate is improved in a sequential manner.

To sum up, Lemma 2 shows that all parameters in the semi-strong identification groups can be consistently estimated, uniformly over $\pi_K \in \Pi_K$, i.e.,

$$\sup_{\pi_K \in \Pi_K} \|\widehat{\pi}_{K-}(\pi_K) - \pi_{K,n}\| \rightarrow_p 0. \quad (3.15)$$

3.3 Asymptotic Distribution in the Reparameterized Model

Next we show the asymptotic distribution of the least squares estimator under mixed identification strength. There are two cases: (a) The last group involves weak identification, i.e., $n^{1/2}\beta_{\mathcal{I}_K} \rightarrow b_{\mathcal{I}_K} \in R^{d_K}$. (b) There are no weakly-identified parameters and the last group only involves strong or semi-strong identification. In case (a), $\pi_{\mathcal{I}_K}$ cannot be consistently estimated because its signal does not dominate the noise from the error. In case (b), we apply the arguments in Lemma 2 to $k = K$ and obtain consistency of $\widehat{\pi}_{\mathcal{I}_K}$.

To characterize the non-standard distribution under weak identification, let $G(\pi_{\mathcal{I}_K})$ be a mean-zero Gaussian process with covariance kernel

$$\Omega(\pi_{\mathcal{I}_K}, \widetilde{\pi}_{\mathcal{I}_K}) = \mathbb{E}_{\gamma_0} U_t^2 d_{\psi_K,t}(\pi_{\mathcal{I}_K-,0}, \pi_{\mathcal{I}_K}) d_{\psi_K,t}(\pi_{\mathcal{I}_K-,0}, \widetilde{\pi}_{\mathcal{I}_K})'. \quad (3.16)$$

Building on this Gaussian process, define

$$\begin{aligned} \tau(\pi_{\mathcal{I}_K}) &= [H_K(\pi_{\mathcal{I}_K}, \pi_{\mathcal{I}_K})]^{-1} [H_K(\pi_{\mathcal{I}_K}, \pi_{\mathcal{I}_K,0}) S_{\mathcal{I}_K} b_{\mathcal{I}_K} + G(\pi_{\mathcal{I}_K})], \\ \xi(\pi_{\mathcal{I}_K}) &= -\frac{1}{2} \tau(\pi_{\mathcal{I}_K})' H_K(\pi_{\mathcal{I}_K}, \pi_{\mathcal{I}_K}) \tau(\pi_{\mathcal{I}_K}), \\ \pi_{\mathcal{I}_K}^* &= \arg \min_{\pi_K \in \Pi_K} \xi(\pi_{\mathcal{I}_K}). \end{aligned} \quad (3.17)$$

We assume that each sample path of the non-central chi-square process $\xi(\pi_{\mathcal{I}_K})$ has a unique minimizer with probability one and call this minimizer $\pi_{\mathcal{I}_K}^*$. In the presence of weak identification, Theorem 1 below shows that $\xi(\pi_{\mathcal{I}_K})$ appears in the limit of the concentrated criterion function

$Q_n^c(\pi_{\mathcal{I}_K})$. In contrast to the right hand of part (a) in Lemma 2, $\xi(\pi_{\mathcal{I}_K})$ cannot identify the true value of $\pi_{\mathcal{I}_K}$. The localization parameter $b_{\mathcal{I}_K}$ represents the signal to noise ratio.

To define the joint distribution in case (b), define covariance matrices

$$\begin{aligned}\Sigma(\pi) &= H^{-1}(\pi)\Omega_\theta(\pi)H^{-1}(\pi), \text{ where} \\ H(\pi) &= \mathbb{E}_{\gamma_0} d_{\theta,t}(\pi)d_{\theta,t}(\pi)' \text{ and } \Omega_\theta(\pi) = \mathbb{E}_{\gamma_0} U_t^2 d_{\theta,t}(\pi)d_{\theta,t}(\pi)' \text{ with} \\ d_{\theta,t}(\pi) &= (g(X_t, \pi)', Z_t', g_\pi(X_t, \pi_k)')'.\end{aligned}\tag{3.18}$$

Assumption 5. (i) $\lambda_{\min}(H(\pi)) \geq \varepsilon$, $\lambda_{\min}(\Omega_\theta(\pi)) \geq \varepsilon$, for some $\varepsilon > 0 \forall \pi \in \Pi$.

(ii) Each sample path of the stochastic process $\{\xi(\pi_{\mathcal{I}_K}) : \pi_{\mathcal{I}_K} \in \Pi_{\mathcal{I}_K}\}$ is minimized at a unique point with probability one.

Theorem 1 (asymptotic distribution of estimators)

Suppose Assumptions 1-5 hold. Then,

(a) with weakly identified parameters: If $n^{1/2}\beta_{\mathcal{I}_K} \rightarrow b_{\mathcal{I}_K} \in R^{d_K}$,

$$n(Q_n^c(\pi_{\mathcal{I}_K}) - Q_n(\psi_{K,n}^0)) \Rightarrow \xi(\pi_{\mathcal{I}_K}),$$

and

$$\begin{pmatrix} n^{1/2}\mathbf{B}(\beta_{K^-,n}) \left(\widehat{\psi}_{K^-} - \psi_{K^-,n} \right) \\ \widehat{\pi}_{\mathcal{I}_K} \end{pmatrix} \Rightarrow \begin{pmatrix} \tau(\pi_{\mathcal{I}_K}^*) - S_{\mathcal{I}_K} b_{\mathcal{I}_K} \\ \pi_{\mathcal{I}_K}^* \end{pmatrix},$$

where $\psi_{K^-} = (\beta', \zeta', \pi_{\mathcal{I}_1}', \dots, \pi_{\mathcal{I}_{K-1}}')'$, $S_{\mathcal{I}_k} = [0_{d_k \times d_{k^-}}, I_{d_k}, 0_{d_k \times (d_\zeta + d_{k^-})}]'$, and $\mathbf{B}(\beta_{K^-,n}) = \text{diag}\{(1_{d_\beta + d_\zeta}, \beta_{K^-,n}'')'\}$;

(b) without weakly identified parameters: If $\|n^{1/2}\beta_{\mathcal{I}_K}\| \rightarrow \infty$, Lemma 2 applies to $k = K$ and

$$n^{1/2}\mathbf{B}(\beta_n) \left(\widehat{\theta} - \theta_n \right) \rightarrow_d N(0, \Sigma(\pi_0)),$$

where $\mathbf{B}(\beta_n) = \text{diag}\{(1_{d_\beta + d_\zeta}, \beta_n')'\}$.

Comments. 1. In case (a), $\widehat{\psi}_{K^-} = (\widehat{\beta}', \widehat{\zeta}', \widehat{\pi}_{\mathcal{I}_1}', \dots, \widehat{\pi}_{\mathcal{I}_{K-1}}')'$ is consistent but it has a non-standard asymptotic distribution. The distribution involves the Gaussian process $\tau(\pi_K)$ and the inconsistent estimator $\pi_{\mathcal{I}_K}^*$. In addition, the rate of convergence of $\widehat{\pi}_{\mathcal{I}_2}, \dots, \widehat{\pi}_{\mathcal{I}_{K-1}}$ are all slower than $n^{-1/2}$.

2. Without weakly identified parameters, the distribution in part (b) is analogous to standard results except for the rescaling matrix $\mathbf{B}(\beta_n)$.

Example (Cont.) In the example $y_t = \beta_1 g_1(X_t, \pi_1) + \beta_2 g_2(X_t, \pi_2) + U_t$, consider the distribution of the least squares estimator when $\beta_{1,n} \rightarrow 0$, $|n^{1/2}\beta_{1,n}| \rightarrow \infty$, and $n^{1/2}\beta_{2,n} \rightarrow b_2 \in R$. Following the grouping rule, the group indices are $\mathcal{I}_1 = \emptyset$, $\mathcal{I}_2 = \{1\}$, $\mathcal{I}_3 = \{2\}$ and the number of groups is $K = 3$. In this case, $\beta = (\beta_1, \beta_2)'$ is strongly identified, π_1 is semi-strongly identified, and π_2 is weakly identified.

The asymptotic results apply to this example as follows. First, Lemma 1 implies that $\widehat{\beta}(\pi)$ is consistent uniformly over $\pi = (\pi_1, \pi_2)'$. Second, applying Lemma 2 with $k = 2$ and $\psi_{2-} = (\beta, \pi_1)'$ yields that $\widehat{\beta}(\pi_2) = \widehat{\beta}(\widehat{\pi}_1(\pi_2), \pi_2)$ and $\widehat{\pi}_1(\pi_2)$ are both consistent uniformly over π_2 . Third, apply Theorem 1(a) with $K = 3$ and $\mathcal{I}_K = \{2\}$, we obtain

$$\begin{pmatrix} n^{1/2}(\widehat{\beta} - \beta_n) \\ n^{1/2}\beta_{1n}(\widehat{\pi}_1 - \pi_{1,n}) \\ \widehat{\pi}_2 \end{pmatrix} \Rightarrow \begin{pmatrix} \tau(\pi_2^*) - S_2 b_2 \\ \pi_2^* \end{pmatrix}, \quad (3.19)$$

where $S_2 b_2 = (0, b_2, 0)'$, $G(\pi_2)$, $\tau(\pi_2)$, and π_2^* are as defined in (3.16) and (3.17) with

$$\begin{aligned} H_K(\pi_2, \pi_{2,0}) &= \mathbb{E}_{\gamma_0} d_{\psi_K, t}(\pi_{1,0}, \pi_2) d_{\psi_K, t}(\pi_{1,0}, \pi_{2,0})', \\ \Omega(\pi_2, \widetilde{\pi}_2) &= \mathbb{E}_{\gamma_0} U_t^2 d_{\psi_K, t}(\pi_{1,0}, \pi_2) d_{\psi_K, t}(\pi_{1,0}, \widetilde{\pi}_2)', \text{ where} \\ d_{\psi_K, t}(\pi_{1,0}, \pi_2) &= (g_1(X_t, \pi_{1,0}), g_2(X_t, \pi_2), g_{\pi_1}(X_t, \pi_{1,0}))'. \quad \square \end{aligned} \quad (3.20)$$

4 Wald Test and t Test with Mixed Identification Strength

Under drifting true parameters, we consider tests of the null hypothesis $H_0 : R\theta_n = v_n$ for some $d_r \times d_\theta$ matrix R of rank d_r . We establish the asymptotic distributions of the Wald statistic and the t statistic, allowing $R\theta$ to involve parameters with different identification strength. Both θ_n and v_n may change with n . This is particularly useful for confidence set construction. For the test $H_0 : \beta_p = 0$, $v_n = 0$.

Under strong identification, Theorem 1(b) implies that $\mathbf{B}^{-1}(\beta_0)\Sigma(\pi_0)\mathbf{B}^{-1}(\beta_0)$ is the asymptotic covariance matrix of the least squares estimator $\widehat{\theta}$. Following the definition of $\Sigma(\pi)$ in (3.18), we estimate $\Sigma(\pi)$ by

$$\begin{aligned} \widehat{\Sigma} &= [\widehat{H}]^{-1} \widehat{\Omega}_\theta [\widehat{H}]^{-1}, \text{ where} \\ \widehat{H} &= n^{-1} \sum_{t=1}^n d_{\theta, t}(\widehat{\pi}) d_{\theta, t}(\widehat{\pi})', \quad \widehat{\Omega}_\theta = n^{-1} \sum_{t=1}^n \widehat{U}_t^2 d_{\theta, t}(\widehat{\pi}) d_{\theta, t}(\widehat{\pi})', \end{aligned} \quad (4.1)$$

and \widehat{U}_i is the regression residual. The standard definition of the Wald statistic for the null hypothesis $H_0 : R\theta_n = v_n$ is

$$W_n(R) = n \left(R\widehat{\theta} - v_n \right)' \left(R\mathbf{B}^{-1}(\widehat{\beta})\widehat{\Sigma}\mathbf{B}^{-1}(\widehat{\beta})R' \right)^{-1} \left(R\widehat{\theta} - v_n \right). \quad (4.2)$$

This is the standard Wald statistic typically used in empirical work. Obviously a standard critical value from the chi-square distribution is justified under strong identification. Below we show that the Wald statistic has a different asymptotic distribution under weak identification. Therefore, a different critical value should be employed. We use the Wald statistic for presentation of the main results. Analogous results hold for the t statistic.

Section 4.1 introduces an orthogonal rotation on the restriction matrix R that separates restrictions on parameters of different identification strength. Section 4.2 uses a rescaling matrix to deal with the asymptotic singularity of the covariance matrix. This section disassemble the Wald statistic into a sandwich form where each part has a non-degenerate limit. The non-standard asymptotic distribution of the test statistics are presented in Section 4.3.

4.1 Rotation

Under mixed identification strength, the estimator $\widehat{\theta}$ involves both inconsistent estimators and consistent estimators of different rates of convergence. It is essential to separate the restrictions on different groups. This is achieved by an orthogonal rotation of the restriction matrix R .

We first introduce the rotation matrix for the general case. Partition the restriction matrix R into

$$R = [R_1 : R_2 : \cdots : R_K], \quad (4.3)$$

where R_1 is the submatrix of R associated with $(\beta', \zeta', \pi'_{\mathcal{I}_1})$, the strongly identified parameters, and R_k is the submatrix of R associated with $\pi_{\mathcal{I}_k}$ for $k = 2, \dots, K$. Thus, R_1 is a $d_r \times (d_\beta + d_\zeta + d_1)$ matrix and R_k is a $d_r \times d_k$ matrix for $k = 2, \dots, K$, where d_k is the number of elements in \mathcal{I}_k . Let

$$A = [A_1 : A_2 : \cdots : A_K] \in \mathcal{O}(d_r) \quad (4.4)$$

be an orthogonal matrix that satisfies two conditions below:

$$(i) A'R = \begin{bmatrix} A'_1 R_1 & 0 & 0 & 0 & 0 \\ A'_2 R_1 & A'_2 R_2 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & 0 & 0 \\ A'_{K-1} R_1 & A'_{K-1} R_2 & \cdots & A'_{K-1} R_{K-1} & 0 \\ A'_K R_1 & A'_K R_2 & \cdots & A'_K R_{K-1} & A'_K R_K \end{bmatrix} \text{ is block upper diagonal} \quad (4.5)$$

and

$$(ii) R^* = \begin{bmatrix} A'_1 R_1 & 0 & 0 & 0 & 0 \\ 0 & A'_2 R_2 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & A'_{K-1} R_{K-1} & 0 \\ 0 & 0 & 0 & 0 & A'_K R_K \end{bmatrix} \text{ has full rank.} \quad (4.6)$$

This rotation matrix A can be obtained as follows. For $k = K$, let $d_K^* = \text{rank}(R_K)$ and A_K be the $d_r \times d_K^*$ matrix whose columns span the column space of R_K . For $k = K - 1$, let $d_{K-1}^* = \text{rank}([R_{K-1} : R_K]) - \text{rank}(R_K)$ and A_{K-1} be a $d_r \times d_{K-1}^*$ matrix such that the rows of $[A_{K-1} : A_K]$ span the columns space of $[R_{K-1} : R_K]$. Continue this step sequentially to $k = K - 2, \dots, 2$. In each step, let

$$d_k^* = \text{rank}[R_k : \cdots : R_K] - \text{rank}[R_{k+1} : \cdots : R_K] \quad (4.7)$$

and A_k be a $d_r \times d_k^*$ matrix such that the columns of $[A_k : \cdots : A_K]$ span the column space of $[R_k : \cdots : R_K]$. Finally, the columns of A_1 ensures that A is an orthogonal matrix. When $d_k = 0$, A_k disappears from the construction of A . In the special case where $d_1, \dots, d_{k-1} = 0$, A_k is chosen to ensure that A is an orthogonal matrix. The rotation is similar to that used by Antoine and Renault (2012) for mixed-rate distribution in different directions.

Following the rotation by A , the linear restrictions in R are separated for parameters with different rates of convergence, including possible inconsistent estimators in group K . In the asymptotic distribution derived below, we show that the block diagonal matrix R^* appears in place of R asymptotically. Under the null, the Wald statistic defined in (4.2) satisfies

$$W_n(R) = W_n(A'R) = W_n(R^*) + \varepsilon_n, \quad (4.8)$$

where ε_n is explicitly defined as the difference between $W_n(A'R)$ and $W_n(R^*)$. In the proof of Theorem 2 below, we show that ε_n is asymptotically negligible. The analysis roughly goes as

follows. Under the null, consider $A'R(\widehat{\theta} - \theta_n)$ in $W_n(A'R)$. For $k = 2, \dots, K$,

$$\begin{aligned} A'_k R(\widehat{\theta} - \theta_n) &= \sum_{\ell < k} A'_k R_\ell (\widehat{\psi}_{k^-} - \psi_{k^-,n}) + A'_k R_k (\widehat{\pi}_{\mathcal{I}_k} - \pi_{\mathcal{I}_k,n}) \text{ and} \\ \widehat{\psi}_{k^-} - \psi_{k^-,n} &= o_p(\|\widehat{\pi}_{\mathcal{I}_k} - \pi_{\mathcal{I}_k,n}\|) \end{aligned} \quad (4.9)$$

where the first equality holds because $A'R$ is upper block-diagonal and the second equality follows from Theorem 1. Therefore, only the block-diagonal elements remain asymptotically and the asymptotic distribution of $W_n(R)$ is determined by that of $W_n(R^*)$.

Example (Cont.) Here we use examples to illustrate the restriction matrix R^* in the simple model $y_t = \beta_1 g(X_t, \pi_1) + \beta_2 g(X_t, \pi_2) + U_t$.

(1) $H_0 : \beta_2 = 0$. In this case, $R = (0, 1, 0, 0)$ and $R^* = R$.

(2) $H_0 : \pi_1 - \pi_2 = 0$. In this case, $R = (0, 0, 1, -1)$. The real restriction vector R^* depends on the identification strength of π_1 and π_2 . (i) If both π_1 and π_2 are strongly identified, $R^* = R$. (ii) If the identification strength of π_1 is stronger such that π_1 is estimated with a faster rate, $R^* = (0, 0, 0, -1)$. (iii) If both π_1 and π_2 are weakly identified, π_1 and π_2 again belong to the same group and $R^* = R$.

(3). $H_0 : \beta_1 + \pi_1 = 0$ and $\pi_1 - \pi_2 = 0$. (i) If β_1 is strongly identified, π_1 is semi-strongly identified (estimated at a rate slower than $n^{-1/2}$), and π_2 is weakly identified,

$$R = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \text{ and } R^* = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \quad (4.10)$$

(ii) If β_1 is strongly identified but π_1 and π_2 are both weakly identified,

$$R = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \text{ and } R^* = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}. \quad \square \quad (4.11)$$

4.2 Rescaling Matrix for Asymptotic Singularity of Covariance Matrix

Under the null, $W_n(R^*)$ can be written as

$$W_n(R^*) = n (\widehat{\theta} - \theta_n)' R^{*'} \left(R^* \mathbf{B}^{-1}(\widehat{\beta}) \widehat{\Sigma} \mathbf{B}^{-1}(\widehat{\beta}) R^{*'} \right)^{-1} R^* (\widehat{\theta} - \theta_n). \quad (4.12)$$

To deal with the asymptotic singularity of the covariance matrix, we start with the diagonal matrix $\mathbf{B}(\widehat{\beta})$. This matrix can be decomposed into two matrices associated with the norm and the angle

of β_k for $k = 1, \dots, K$, i.e.,

$$\begin{aligned}\mathbf{B}(\beta) &= \mathbf{D}(\beta)\mathbf{B}(\omega), \text{ where} \\ \mathbf{D}(\beta) &= \text{diag}\{(1_{d_\beta+d_\zeta}, \|\beta_1\| 1_{d_1}, \dots, \|\beta_K\| 1_{d_K})'\}, \\ \mathbf{B}(\omega) &= \text{diag}\{(1_{d_\beta+d_\zeta}, \omega'_1, \dots, \omega'_K)'\}.\end{aligned}\tag{4.13}$$

Write $\mathbf{B}(\hat{\beta}) = \mathbf{D}(\hat{\beta})\mathbf{B}(\hat{\omega})$. The angel parameters in $\mathbf{B}(\hat{\omega})$ do not converge to zero, following the grouping rule. To deal with the asymptotic singularity of $\mathbf{D}(\hat{\beta})$, define a new diagonal matrix $\mathbf{D}^*(\hat{\beta})$ as

$$\mathbf{D}^*(\beta) = \text{diag}\{(1_{d_1^*}, \|\beta_2\| 1_{d_2^*}, \dots, \|\beta_K\| 1_{d_K^*})'\} \in R^{d_r \times d_r}.\tag{4.14}$$

Because R^* is a block-diagonal matrix, it follows that

$$R^*\mathbf{D}(\beta) = \mathbf{D}^*(\beta)R^*.\tag{4.15}$$

With probability approaching one,

$$W_n(R^*) = W_n(\mathbf{D}^*(\hat{\beta})R^*) = \rho'_n V_n^{-1} \rho_n,\tag{4.16}$$

where

$$\begin{aligned}\rho_n &= n^{1/2}\mathbf{D}^*(\hat{\beta})R^*(\hat{\theta} - \theta_n) = n^{1/2}R^*\mathbf{D}(\hat{\beta})(\hat{\theta} - \theta_n) \\ &= R^*\mathbf{B}^{-1}(\hat{\omega})\xi_n \text{ with } \xi_n = n^{1/2}\mathbf{B}(\hat{\beta})(\hat{\theta} - \theta_n),\end{aligned}\tag{4.17}$$

and

$$\begin{aligned}V_n &= \mathbf{D}^*(\hat{\beta})R^*\mathbf{B}^{-1}(\hat{\beta})\hat{\Sigma}\mathbf{B}^{-1}(\hat{\beta})R^{*\prime}\mathbf{D}^*(\hat{\beta}) \\ &= R^*\mathbf{B}^{-1}(\hat{\omega})\hat{\Sigma}\mathbf{B}^{-1}(\hat{\omega})R^{*\prime}.\end{aligned}\tag{4.18}$$

The equality in (4.15) and $\mathbf{B}(\hat{\beta}) = \mathbf{D}(\hat{\beta})\mathbf{B}(\hat{\omega})$ are used in both (4.17) and (4.18). An important implication of the calculation in (4.18) is that V_n is non-singular asymptotically and V_n^{-1} appears as the rescaling covariance matrix in (4.16). Below we derive the asymptotic distribution of ξ_n , $\mathbf{B}^{-1}(\hat{\omega})$, and $\hat{\Sigma}$ under all identification scenarios, which in turn yields the asymptotic distribution of the Wald statistic following (4.16)-(4.18).

4.3 Non-standard Distribution of the Test Statistic

First consider the re-centered and re-scaled parameter ξ_n defined in (4.17). Following the asymptotic distribution in Theorem 1(a), define a function of the Gaussian process $\tau(\pi_K)$:

$$\xi(\pi_{\mathcal{I}_K}) = \begin{pmatrix} \tau(\pi_{\mathcal{I}_K}) - S_{\mathcal{I}_K} b_{\mathcal{I}_K} \\ \text{diag}(\tau_{\beta_K}(\pi_{\mathcal{I}_K})) (\pi_{\mathcal{I}_K} - \pi_{\mathcal{I}_K,0}) \end{pmatrix}, \text{ where } \tau_{\beta_K}(\pi_{\mathcal{I}_K}) = S'_{\mathcal{I}_K} \tau(\pi_K). \quad (4.19)$$

Under weak identification, we show $\xi_n \Rightarrow \xi(\pi_{\mathcal{I}_K}^*)$ in the proof of Theorem 2 below. To handle the matrix $\mathbf{B}^{-1}(\hat{\omega})$ in ρ_n and V_n , define

$$\begin{aligned} \omega(\pi_{\mathcal{I}_K}) &= \left(\omega'_{1,0}, \omega'_{2,0}, \dots, \omega'_{K-1,0}, \frac{\tau_{\beta_K}(\pi_{\mathcal{I}_K})'}{\|\tau_{\beta_K}(\pi_{\mathcal{I}_K})\|} \right)' \text{ and} \\ \mathbf{B}_{\omega}(\pi_{\mathcal{I}_K}) &= \mathbf{B}(\omega(\pi_{\mathcal{I}_K})) = \text{diag} \left\{ \left(1_{d_{\beta}+d_{\zeta}}, \omega'_{1,0}, \dots, \omega'_{K-1,0}, \frac{\tau_{\beta_K}(\pi_{\mathcal{I}_K})'}{\|\tau_{\beta_K}(\pi_{\mathcal{I}_K})\|} \right)' \right\}. \end{aligned} \quad (4.20)$$

For the strong and semi-strong identification groups, the angel parameters are estimated consistently. This is the reason that $\omega_{k,0}$ shows up in (4.20) for $k = 1, \dots, K-1$. For group K , $\tau_{\beta_K}(\pi_{\mathcal{I}_K}^*)/\|\tau_{\beta_K}(\pi_{\mathcal{I}_K}^*)\|$ characterizes the limit of the angel parameter.

In the proof of Theorem 2, we show that

(a) under weak identification, i.e., $n^{1/2}\beta_K \rightarrow b_K \in R^{d_K}$,

$$\xi_n \Rightarrow \xi(\pi_{\mathcal{I}_K}^*), \hat{\omega} \Rightarrow \omega(\pi_{\mathcal{I}_K}^*), \hat{\Sigma} \Rightarrow \Sigma(\pi_{\mathcal{I}_K}^*), \quad (4.21)$$

(b) without weak identification, i.e., $\|n^{1/2}\beta_K\| \rightarrow \infty$,

$$\xi_n \rightarrow_d \xi \sim N(0, \Sigma(\pi_0)), \hat{\omega} \rightarrow_p \omega_0, \hat{\Sigma}(\hat{\pi}) \rightarrow_p \Sigma(\pi_0). \quad (4.22)$$

All convergence holds jointly. Put the distributions in (4.21) and (4.22) together with the decomposition in (4.16)-(4.18), the asymptotic distribution of the Wald statistic is given below.

Theorem 2 (Wald statistic with mixed identification strength)

Suppose Assumptions 1-5 hold. Then,

(a) with weakly identified parameters: If $n^{1/2}\beta_{\mathcal{I}_K} \rightarrow b_{\mathcal{I}_K} \in R^{d_K}$,

$W_n(R) \Rightarrow \mathcal{W}(\pi_{\mathcal{I}_K}^*)$, where

$$\mathcal{W}(\pi_{\mathcal{I}_K}) = [R^* B_{\omega}^{-1}(\pi_{\mathcal{I}_K}) \xi(\pi_{\mathcal{I}_K})]' [R^* (B_{\omega}^{-1}(\pi_{\mathcal{I}_K}) \Sigma(\pi_{\mathcal{I}_K}) B_{\omega}^{-1}(\pi_{\mathcal{I}_K})) R^{*'}]^{-1} [R^* B_{\omega}^{-1}(\pi_{\mathcal{I}_K}) \xi(\pi_{\mathcal{I}_K})];$$

(b) without weakly identified parameters: If $\|n^{1/2}\beta_{\mathcal{I}_K}\| \rightarrow \infty$, $W_n(R) \rightarrow_d \chi_{d_r}^2$.

Comments: 1. The asymptotic distribution of the Wald statistic not only depends on the weak identification group through $b_{\mathcal{I}_K}$, but also depends on the rest of the group specification through R^* .

2. Theorem 2 shows that the Wald statistic has a non-standard asymptotic distribution if some parameters are weakly identified. Quantiles of this non-standard distribution can be obtained by simulation. The Wald statistic has a chi-square distribution asymptotically as long as all parameters are at least semi-strongly identified. Semi-strong identification affects the rate of convergence of the estimators but not the asymptotic distribution of the Wald statistic. The Wald statistic for tests with linear restrictions is self-corrected when all parameters are consistently estimated.

For single hypothesis $H_0 : R\theta_n = v_n$ where $d_r = 1$, we can also use the t statistic:

$$t_n(R) = \frac{n^{1/2} (R\hat{\theta} - v_n)}{\sqrt{R\mathbf{B}^{-1}(\hat{\beta})\hat{\Sigma}\mathbf{B}^{-1}(\hat{\beta})R'}}. \quad (4.23)$$

This is the standard definition of the t statistic.

Corollary 1 (t statistic with mixed identification strength)

Suppose Assumptions 1-5 hold and $d_r = 1$. Then,

(a) with weakly identified parameters: If $n^{1/2}\beta_{\mathcal{I}_K} \rightarrow b_{\mathcal{I}_K} \in R^{d_K}$,

$$t_n(R) \Rightarrow \mathcal{T}(\pi_{\mathcal{I}_K}^*), \text{ where } \mathcal{T}(\pi_{\mathcal{I}_K}) = \frac{R^* B_\omega^{-1}(\pi_{\mathcal{I}_K}) \xi(\pi_{\mathcal{I}_K})}{\sqrt{R^* (B_\omega^{-1}(\pi_{\mathcal{I}_K}) \Sigma(\pi_{\mathcal{I}_K}) B_\omega^{-1}(\pi_{\mathcal{I}_K})) R^{*'}}};$$

(b) without weakly identified parameters: If $\|n^{1/2}\beta_{\mathcal{I}_K}\| \rightarrow \infty$, $t_n(R) \rightarrow_d N(0, 1)$.

Example (Cont.) Now we get back to the example $y_t = \beta_1 g_1(X_t, \pi_1) + \beta_2 g_2(X_t, \pi_2) + U_t$ and consider the null hypothesis $H_0 : \beta_2 = 0$. The restriction matrix is $R = R^* = (0, 0, 0, 1)$. Under the null, $n^{1/2}\beta_{2,n} = b_2 = 0$. The distribution of the Wald statistic depends on the identification strength of π_1 .

(1) If $|n^{1/2}\beta_{1,n}| \rightarrow \infty$, which includes both strong and semi-strong identification of π_1 , $\mathcal{I}_K = \{2\}$ and $b_2 = 0$. In this case, $\pi_{\mathcal{I}_K} = \pi_2$. The elements in $\mathcal{T}(\pi_2)$ are specified as follows: $\xi(\pi_2)$ is as specified in (4.19) with $\tau(\pi_2)$ given in (3.20), $S_2 = (0, 1, 0)'$, $b_2 = 0$, and $B_\omega(\pi_2) = I_4$.

(2) If $n^{1/2}\beta_{1,n} \rightarrow b_1 \in R$, $\mathcal{I}_K = \{1, 2\}$ and $b = (b_1, b_2)' = (b_1, 0)'$. In this case, $\pi_{\mathcal{I}_K} = \pi$. The elements in $\mathcal{T}(\pi)$ are specified as follows: $G(\pi)$, $\tau(\pi)$, and π^* are as defined in (3.16) and (3.17) with

$$\begin{aligned} H_K(\pi, \pi_0) &= \mathbb{E}_{\gamma_0} d_{\psi_K, t}(\pi) d_{\psi_K, t}(\pi_0)', \\ \Omega(\pi, \tilde{\pi}) &= \mathbb{E}_{\gamma_0} U_t^2 d_{\psi_K, t}(\pi) d_{\psi_K, t}(\tilde{\pi})', \text{ where} \\ d_{\psi_K, t}(\pi) &= (g_1(X_t, \pi_1), g_2(X_t, \pi_2))', \end{aligned} \quad (4.24)$$

the selector matrix is $S_{\mathcal{I}_K} = I_2$, and

$$\begin{aligned} S_{\mathcal{I}_K} b_{\mathcal{I}_K} &= b = (b_1, 0)', \quad \tau_{\beta_K}(\pi_{\mathcal{I}_K}) = \tau(\pi), \\ \mathbf{B}_\omega(\pi_{\mathcal{I}_K}) &= \text{diag} \left\{ \left(1, 1, \frac{\tau(\pi)'}{\|\tau(\pi)\|} \right)' \right\}. \quad \square \end{aligned} \quad (4.25)$$

5 Robust Inference

Next, we link the asymptotic distributions under all group specifications to the asymptotic size of tests and confidence sets, which approximates the finite-sample size of tests and confidence sets, respectively. To this end, we first formally define the asymptotic size. For fixed v , the asymptotic size of a test for the null hypothesis: $H_0 : R\theta_n = v$ is

$$AsySz = \limsup_{n \rightarrow \infty} \left[\sup_{\gamma \in \Gamma: R\theta = v} \mathbb{P}_\gamma (T_n(R) > c_{n, 1-\alpha}(v)) \right], \quad (5.1)$$

which is the limsup of the finite-sample size of the test. A nominal $1 - \alpha$ confidence set for $R\theta$ is obtained by inverting the tests for $H_0 : R\theta_n = v_n$, i.e., $CS_n = \{v_n : T_n(R) \leq c_{n, 1-\alpha}(v_n)\}$. The asymptotic size of this confidence set is

$$AsySz = \liminf_{n \rightarrow \infty} \inf_{\gamma \in \Gamma} \mathbb{P}_\gamma \left(T_n(R) \leq c_{n, 1-\alpha}(v_n) \right), \quad (5.2)$$

which is the liminf of the finite-sample size of the confidence set.

5.1 Potential Size Distortion

Theorem 2 and Corollary 1 show that the asymptotic distributions of the Wald statistic and t statistic depend on

$$h = (\mathcal{I}, b_{\mathcal{I}_K}, \omega_0, \gamma_0), \quad (5.3)$$

where \mathcal{I} is the group specification, $n^{1/2}\beta_{\mathcal{I}_K,n} \rightarrow b_{\mathcal{I}_K}$ measures the identification strength of group K , $\omega_{\mathcal{I}_k,n} \rightarrow \omega_{\mathcal{I}_k,0}$ is the angle parameter in group k , $\gamma_n \rightarrow \gamma_0 \in \Gamma$. Let \mathcal{H}_I denote the collection of all group specifications. Then the parameter space of h is

$$H = \{h = (\mathcal{I}, b_{\mathcal{I}_K}, \omega, \gamma) : \mathcal{I} \in \mathcal{H}_I, b_{\mathcal{I}_K} \in (R \cup \{\pm\infty\})^{d_K}, \|\omega_{\mathcal{I}_k}\| = 1, \gamma \in \Gamma\}. \quad (5.4)$$

When the null hypothesis is $H_0 : R\theta = v$ for fixed v , the value of parameter h that is consistent with the null hypothesis is collected in

$$H(v) = \{h \in H : R\theta_0 = v\}. \quad (5.5)$$

Along a sequence of true parameters $\{\gamma_n \in \Gamma : n \geq 1\}$ associated with h , define

$$\mathcal{W}(h) = \begin{cases} \mathcal{W}(\pi_K^*), & \text{if Theorem 2 (a) holds,} \\ \chi_{d_r}^2, & \text{if Theorem 2 (b) holds.} \end{cases} \quad (5.6)$$

For the t test, define $\mathcal{T}(h)$ similarly to $\mathcal{W}(h)$, with $\mathcal{W}(\pi_K^*)$ and $\chi_{d_r}^2$ replaced by $\mathcal{T}(\pi_K^*)$ and $N(0, 1)$, respectively.

For a standard Wald test, the $1 - \alpha$ quantile of $\chi_{d_r}^2$, denoted by $\chi_{d_r, 1-\alpha}^2$, is used as the critical value. For a standard symmetric two sided t test, the $1 - \alpha/2$ quantile of $N(0, 1)$, denoted by $z_{1-\alpha/2}$, is used as the critical value.

Assumption CV1. (i) $\mathcal{W}(h)$ is continuous at $\chi_{d_r, 1-\alpha}^2 \forall h \in H$.

(ii) $|\mathcal{T}(h)|$ is continuous at $z_{1-\alpha/2} \forall h \in H$.

Theorem 3 (size distortion of standard test and confidence set)

Suppose Assumptions 1-5 and CV1 hold. Then,

- (a) the asymptotic size of a standard Wald test is $\sup_{h \in H(v)} \Pr(\mathcal{W}(h) > \chi_{d_r, 1-\alpha}^2)$;
- (b) the asymptotic size of a standard Wald confidence set is $\inf_{h \in H} \Pr(\mathcal{W}(h) \leq \chi_{d_r, 1-\alpha}^2)$;
- (c) parts (a) and (b) apply to the symmetric two-sided t test and confidence set by replacing $\mathcal{W}(h)$ with $\mathcal{T}(h)$ and replacing $\chi_{d_r, 1-\alpha}^2$ with $z_{1-\alpha/2}$.

Comments. 1. The degree of size distortion for a standard test and confidence set can be simulated using the formula in Theorem 3 and the distributions derived in Theorem 2 and Corollary 1.

2. Theorem 3 is proved by applying the generic results in Andrews, Cheng, and Guggenberger (2011). A reparameterization of h is introduced to fit this problem in the generic set up.

5.2 Data-Dependent Non-Standard Critical Values

To avoid size distortion, the ideal critical value to use is the $1 - \alpha$ quantile of $\mathcal{W}(h)$ or $\mathcal{T}(h)$ in the presence of weak identification. However, these distributions depend on the unknown parameter h specified in (5.3). When constructing a robust critical value, the general strategy is to plug in elements of h that can be consistently estimated and take a supreme of the quantiles over the elements of h that cannot be consistently estimated.

A special element of h is the group specification \mathcal{I} . The group specification \mathcal{I} cannot be consistently estimated, however, an identification-category-selection (ICS) method can significantly reduce the number of group specifications relevant for robust inference. This ICS procedure uses data to determine the weak identification group \mathcal{I}_K , leaving the semi-strong identification groups $\mathcal{I}_2, \dots, \mathcal{I}_{K-1}$ and the strong identification group \mathcal{I}_1 unspecified. This method is closely related to the generalized moment selection method in Andrews and Soares (2010) and the type 1 robust critical value in Andrews and Cheng (2012). Different from these papers, the group specification \mathcal{I} cannot be fully determined by the ICS procedure. Nevertheless, this selection yields a less conservative choice of the critical value.

For $j = 1, \dots, p$, let

$$ICS_{j,n} = n^{1/2}(\widehat{\Sigma}_j)^{-1/2} \left| \widehat{\beta}_j \right|, \quad (5.7)$$

where $\widehat{\Sigma}_j$ is the j -th diagonal element of $\widehat{\Sigma}$. Roughly speaking, $ICS_{j,n} = O_p(1)$ only if $\beta_{j,n} = O(n^{-1/2})$. We select the weak identification group by

$$\widehat{\mathcal{I}}_W = \{j : ICS_{j,n} \leq \kappa_{j,n}\}, \quad (5.8)$$

where $\{\kappa_{j,n} : n \geq 1\}$ is a sequence of constants such that $\kappa_{j,n} \rightarrow \infty$ and $\kappa_{j,n}/n^{1/2} \rightarrow 0$ for $j = 1, \dots, p$. For the null hypothesis $H_0 : \beta_k = 0$, we put k in $\widehat{\mathcal{I}}_W$ without selection. The regressors are selected one by one in $\widehat{\mathcal{I}}_W$. If prior information is available for a group structure, the selection statistic $ICS_{j,n}$ can be modified to take the form of a Wald statistic. Define

$$\widehat{H} = \{h \in H : \mathcal{I}_K = \widehat{\mathcal{I}}_W, \omega_{\mathcal{I}_k} = \widehat{\beta}_{\mathcal{I}_k} / \|\widehat{\beta}_{\mathcal{I}_k}\| \text{ and } \pi_{\mathcal{I}_k} = \widehat{\pi}_k \text{ for } k < K\}^5 \quad (5.9)$$

⁵The asymptotic distribution $\mathcal{W}(\pi_K^*)$ does not depend on the true values of β and ζ although both of them can be consistently estimated. Hence, we do not plug in the estimators of β and ζ .

Let $\mathcal{W}_{1-\alpha}(h)$ denote the $1 - \alpha$ quantile of $\mathcal{W}(h)$ defined in (5.6). To obtain a confidence set by inverting tests for $H_0 : R\theta_n = v_n$ with the Wald statistic, we suggest the plug-in critical value

$$\hat{c}_{n,1-\alpha} = \sup_{h \in \hat{H}} \mathcal{W}_{1-\alpha}(h). \quad (5.10)$$

Because \hat{H} is a subset of H , $\hat{c}_{n,1-\alpha}$ is smaller than $\sup_{h \in H} \mathcal{W}_{1-\alpha}(h)$, which is the least favorable critical value. To test the null hypothesis $H_0 : R\theta_n = v$ for fixed v , the plug-in critical value $\hat{c}_{n,1-\alpha}(v)$ is obtained by replacing \hat{H} with $\hat{H}(v) = \hat{H} \cap H(v)$. When the t statistic is used for a symmetric two-sided test, the plug-in critical values is constructed with $\mathcal{W}_{1-\alpha}(h)$ replaced by the $1 - \alpha$ quantile of $|\mathcal{T}(h)|$. We call test and confidence set based on this plug-in critical value robust test and robust confidence set.

In empirical implementation, the first step is to specify \hat{H} by the ICS method. Second, simulate $\mathcal{W}_{1-\alpha}(h)$ for each h using the asymptotic distribution in Theorem 4. Simulation methods for a Gaussian processes is given in Hansen (1996). Finally, obtain the plug-in critical value following (5.10). The computation depends on the number of nonlinear regressors in the model as well the parameter of interest. In many cases, $\mathcal{W}_{1-\alpha}(h)$ does not depend on \mathcal{I} except for the weak identification group \mathcal{I}_K . The procedure becomes computation intensive as the number of nonlinear regressors increases. For this reason, the current paper suggests a simple data-dependent rule in (5.8). More sophisticated and computation intensive data-dependent choices are considered for other models or for general set-ups in Andrews and Barwick (2012), Andrews and Cheng (2012), McCloskey (2012), among others. These methods can be adapted to the present model using the asymptotic distributions developed in Sections 3 and 4.

Assumption CV2. (i) $\mathcal{W}(h)$ is uniformly continuous in $\omega_{\mathcal{I}_k}$ and $\pi_{\mathcal{I}_k}$ for $k = 1, \dots, K - 1$ on $h \in H$.

(ii) $\mathcal{W}(h)$ is continuous at its $1 - \alpha$ quantile for all $h \in H$ and $\alpha \in (0, 1/2)$.

(iii) Parts (a) and (b) hold with $\mathcal{W}(h)$ replaced by $|\mathcal{T}(h)|$.

The following result holds for the robust test and confidence set based on the Wald statistic and the t statistic.

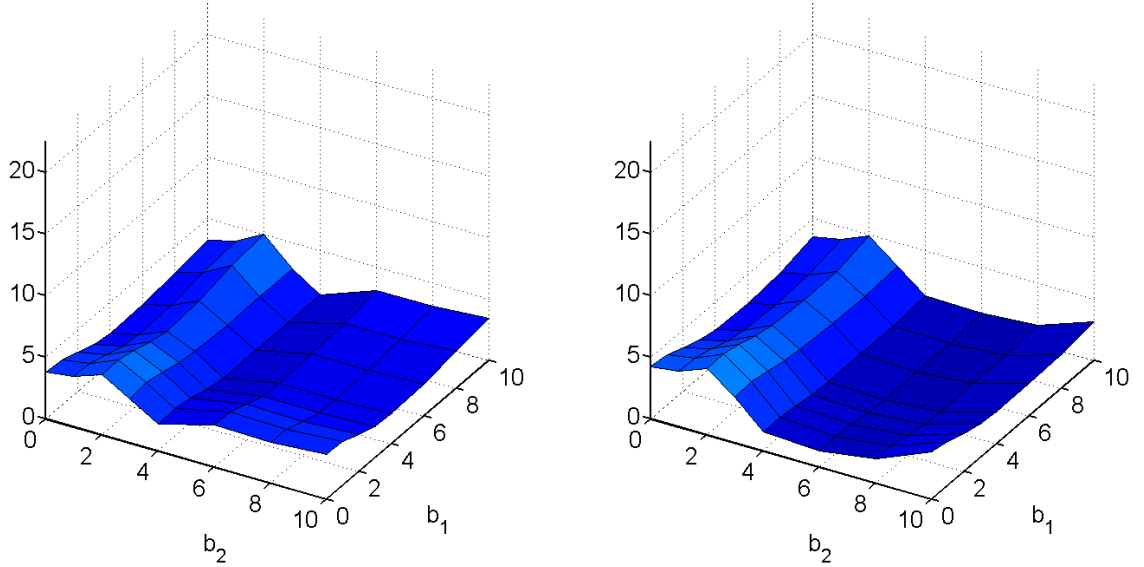
Theorem 4 (robust test and confidence set)

Suppose Assumptions 1-5 and Assumption CV2 hold. Then,

(a) *the asymptotic size of the robust test of $H_0 : R\theta = v$ is α ;*

(b) *the asymptotic size of the robust confidence set of $R\theta$ is $1 - \alpha$.*

Figure 2: Robust Test: Asymptotic (left) and Finite-Sample (right, $n = 500$) Rejection Probability ($\times 100$) for $H_0 : \beta_2 = \beta_{2,0}$.



Notes: DGP is the same as that for Figure 1, nominal size $\alpha = 5\%$, the true values of β_1 and β_2 are $\beta_{1,0} = b_1/\sqrt{500}$ and $\beta_{2,0} = b_2/\sqrt{500}$ in the right panel.

Example (Cont.) Figure 2 presents numerical results for robust tests in $y_t = \beta_1 g_1(X_t, \pi_1) + \beta_2 g_2(X_t, \pi_2) + U_t$. The DGP is the same as that for Figure 1 so that the performance of the standard test and the robust test can be compared. The test statistic is the symmetric two-sided t statistic, coupled with the standard critical value in Figure 1 and the robust critical value in Figure 2. The left panel of Figure 2 is obtained by drawing the t statistic and the ICS statistic from their asymptotic distributions.⁶ Both figures demonstrate how the null rejection probability of the test changes with the true values of β_1 and β_2 .

Table 1 below focuses on the test $H_0 : \beta_2 = 0$ and shows the null rejection probability as a function of b_1 and the true value of π_1 , denoted by $\pi_{1,0}$. Under the null, the true value of π_2 is irrelevant.

In this example, the nonlinear functions are the exponential smooth transition function. Specifically, $x = (x_1, x_2, x_3)'$, $g_1(X, \pi_1) = x_1(1 - \exp(-c(x_3 - \pi_1)^2))$, $g_2(x, \pi_2) = x_2(1 - \exp(-c(x_3 - \pi_2)^2))$. The marginal effect of x_1 and x_2 are both nonlinear, depending on the transition variable x_3 . The

⁶The asymptotic distribution of the t statistic and the ICS statistic are given in Corollary 1 and (C.11) in the appendix. The ICS statistics are non-centered t statistics. Thus, their asymptotic distributions follow the same arguments for the t statistic.

marginal distribution of X_{1t} , X_{2t} , X_{3t} , U_t are all standard normal and independent across observations. The correlation coefficient between X_{1t} and X_{2t} is 0.5, both are uncorrelated with X_{3t} . The error U_t is independent of all other variables. The true values of β_1 and β_2 are b_1/\sqrt{n} and b_2/\sqrt{n} , respectively, for finite-sample results with sample size n .⁷ The true values of π_1 and π_2 are both 0 for Figures 1 and 2. The optimization parameter space for π_1 and π_2 are both $[-1, 1]$. The constant c is 10. In all cases, 50,000 simulation repetitions are conducted.

The right panel of Figure 2 is comparable to the right panel of Figure 1 with the standard test replaced by the robust test. The left panel of Figure 2 is an asymptotic version of the right panel obtained by drawing the t statistic and the ICS statistic from their asymptotic distributions. To demonstrate the effect of the ICS procedure for different values of b_1 and b_2 , we consider $\pi_{1,0} = 0$ and $\pi_{2,0} = 0$ when constructing the robust critical value in Figure 2.

In Figure 2, the ICS procedure is based on a data-dependent choice of the tuning parameter. First, the ICS statistic $ICS_{1,n}$ and $ICS_{2,n}$ are constructed following (5.7). They are compared with tuning parameters $\kappa_{1,n} = c_1 \log(\log(n))$ and $\kappa_{2,n} = c_2 \log(\log(n))$ to determine the weak identification set $\widehat{\mathcal{I}}_W$. The constants c_1 and c_2 are tuned by the asymptotic null rejection probabilities through simulation. Replacing the t statistic and the ICS statistic by draws from their asymptotic distributions, we simulate the null rejection probability of the robust test for any values of c_1 and c_2 . For large values of c_1 and c_2 , the ICS procedure favors the least favorable critical value, which controls the maximum rejection probability but tends to under reject for some values of b_1 and b_2 . In the simulation for Figure 2, we choose c_1 and c_2 that minimize the average probability of under rejection, provided that the maximum rejection probability is no larger than $\alpha + \varepsilon$, where ε is a tolerance level close to 0. We set $\alpha = 5\%$ and $\varepsilon = 0.1\%$ in the simulation. The same constants c_1 and c_2 are used in the two panels of Figure 2. These choices minimize the non-similarity of the test over b_1 and b_2 while controlling the maximum rejection probability.

Table 1 focuses on the test $H_0 : \beta_2 = 0$ under different values of b_1 and $\pi_{1,0}$. Under the null, the data does not depend on π_2 . Because $b_2 = 0$, the ICS procedure only compares $ICS_{1,n}$ with $\kappa_{1,n} = c_1 \log(\log(n))$. Similar to Figure 2, we choose c_1 to minimize the average rate of under rejection over b_1 and $\pi_{1,0}$, provided that the maximum null rejection probability is controlled. When the sample size is 500, the maximum rejection probability of robust test is 5.7% and the minimum rejection probability of the robust test is 4.5%. \square

Tests proposed in this paper are robust to identification loss in multiple areas of the parameters space. It is particularly useful for sub-vector inference when the nuisance parameters have mixed

⁷In simulations, the grids for b_1 and b_2 are $\{1, 2, 3, 4, 5, 6, 8, 10, 20, 30\}$. Only results for b_1 and b_2 up to 10 are reported because they are stable for larger values of b_1 and b_2 .

Table 1: Rejection Probability (x100) of Tests for $H_0 : \beta_2 = 0$ versus $H_0 : \beta_2 \neq 0$

| $\pi_{1,0}$ | b_1 | Robust | | | Standard | | | |
|-------------|-------|-----------|-----------|------------|-----------|-----------|------------|------|
| | | $n = 200$ | $n = 500$ | $n = 1000$ | $n = 200$ | $n = 500$ | $n = 1000$ | |
| 0 | 0 | 6.2 | 5.5 | 5.3 | 21.0 | 19.9 | 19.7 | |
| | 1 | 6.1 | 5.2 | 5.0 | 20.0 | 19.1 | 18.8 | |
| | 2 | 5.7 | 4.7 | 4.5 | 18.0 | 16.9 | 16.7 | |
| | 3 | 5.4 | 4.5 | 4.1 | 16.5 | 15.4 | 15.2 | |
| | 4 | 5.5 | 4.6 | 4.2 | 15.9 | 14.8 | 14.6 | |
| | 6 | 6.0 | 5.0 | 4.6 | 15.8 | 14.8 | 14.5 | |
| | 8 | 6.3 | 5.4 | 5.2 | 15.8 | 14.7 | 14.5 | |
| | 10 | 6.3 | 5.4 | 5.3 | 15.8 | 14.7 | 14.4 | |
| | 0.3 | 0 | 6.4 | 5.5 | 5.4 | 21.3 | 19.9 | 19.4 |
| | | 1 | 6.1 | 5.3 | 5.1 | 20.3 | 19.1 | 18.5 |
| 2 | | 5.8 | 4.8 | 4.5 | 18.1 | 16.9 | 16.4 | |
| 3 | | 5.5 | 4.5 | 4.2 | 16.8 | 15.3 | 15.1 | |
| 4 | | 5.6 | 4.6 | 4.3 | 16.3 | 14.8 | 14.5 | |
| 6 | | 6.3 | 5.1 | 4.9 | 16.2 | 14.8 | 14.4 | |
| 8 | | 6.5 | 5.5 | 5.3 | 16.2 | 14.8 | 14.5 | |
| 10 | | 6.4 | 5.5 | 5.4 | 16.2 | 14.7 | 14.4 | |
| 0.5 | | 0 | 6.2 | 5.6 | 5.2 | 20.9 | 20.4 | 19.5 |
| | | 1 | 6.0 | 5.3 | 4.9 | 19.8 | 19.3 | 18.6 |
| | 2 | 5.6 | 4.8 | 4.4 | 17.9 | 17.2 | 16.5 | |
| | 3 | 5.4 | 4.6 | 4.1 | 16.5 | 15.7 | 14.9 | |
| | 4 | 5.5 | 4.7 | 4.3 | 16.0 | 15.1 | 14.4 | |
| | 6 | 6.0 | 5.2 | 4.9 | 16.0 | 15.0 | 14.3 | |
| | 8 | 6.3 | 5.6 | 5.3 | 16.0 | 15.1 | 14.4 | |
| | 10 | 6.2 | 5.6 | 5.4 | 15.9 | 15.0 | 14.3 | |
| | 0.8 | 0 | 6.1 | 5.5 | 5.1 | 21.1 | 20.0 | 19.5 |
| | | 1 | 5.8 | 5.2 | 4.9 | 20.0 | 18.9 | 18.5 |
| 2 | | 5.5 | 4.8 | 4.3 | 17.9 | 16.8 | 16.4 | |
| 3 | | 5.3 | 4.6 | 4.1 | 16.5 | 15.4 | 14.9 | |
| 4 | | 5.4 | 4.7 | 4.3 | 16.0 | 14.9 | 14.5 | |
| 6 | | 5.9 | 5.3 | 5.0 | 16.0 | 15.0 | 14.4 | |
| 8 | | 6.2 | 5.7 | 5.4 | 16.1 | 15.0 | 14.4 | |
| 10 | | 6.1 | 5.7 | 5.5 | 16.1 | 14.9 | 14.3 | |
| max | | | 6.5 | 5.7 | 5.5 | 21.3 | 20.4 | 19.7 |

identification strength. The ICS procedure and the plug-in method improve the efficiency of the robust test, however, the test does not have optimality properties, such as those discussed in Elliott, Müller, and Watson (2012). Besides the Wald statistic and the t statistic, one can derive the asymptotic distributions of the QLR and LM statistics along drifting parameters and simulate their robust critical values in a similar fashion. Andrews and Cheng (2012) study the QLR statistic when identification loss occurs at one point. With multiple points of non-identification in this paper, the sequential peeling method developed in Section 3.2 is useful to analyze the constrained sample criterion function. We leave these alternative robust tests and their comparison for future work.

Appendix

This version: May 8, 2014

For notational simplicity, in this appendix we use subscript k rather than \mathcal{I}_k to denote parameters in group k for notational simplicity. For example, π_k and β_k are used in place of $\pi_{\mathcal{I}_k}$ and $\beta_{\mathcal{I}_k}$. The continuous mapping theorem is abbreviated to CMT. Left hand side and right hand are abbreviated to lhs and rhs. With probability approaching one is written as w.p.a.1.

A Auxiliary Lemmas

Let $s(W, \theta)$ denote a function of θ that is differentiable on the support of W . Its derivative is denoted by $s_\theta(W, \theta)$. The following lemmas apply to strictly stationary strong mixing time series under Assumption 2 or i.i.d. data under Assumption 2*.

Lemma A.1 (uniform law of large numbers)

Suppose (i) Assumption 2(i) or 2*(i) holds, (ii) $E_\gamma(\sup_{\theta \in \Theta} \|s(W_t, \theta)\|^{1+\delta} + \sup_{\theta \in \Theta} \|s_\theta(W_t, \theta)\|^{1+\delta}) \leq C \forall \gamma \in \Gamma$ for some $C < \infty$ and $\delta > 0$, and (iii) Θ is compact. Then, (i) $\sup_{\theta \in \Theta} \|n^{-1} \sum_{t=1}^n s(W_t, \theta) - E_{\gamma_n} s(W_t, \theta)\| \rightarrow_p 0$ under any sequence of true parameters $\{\gamma_n \in \Gamma : n \geq 1\}$ and $\gamma_n \rightarrow \gamma_0 \in \Gamma$. (ii) $E_{\gamma_0} s(W_t, \theta)$ is uniformly continuous on $\Theta \forall \gamma_0 \in \Gamma$.

Lemma A.2 (stochastic equicontinuity)

(a) Suppose (i) Assumption 2(i) holds, (ii) $E_\gamma(\sup_{\theta \in \Theta} \|s(W_t, \theta)\|^q + \sup_{\theta \in \Theta} \|s_\theta(W_t, \theta)\|^q) \leq C \forall \gamma \in \Gamma$ for some $C < \infty$ and q as in Assumption 2(i). Then, $\nu_n s(\theta) = n^{-1/2} \sum_{t=1}^n (s(W_t, \theta) - E_{\gamma_n} s(W_t, \theta))$ 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 $\limsup_{n \rightarrow \infty} P[\sup_{\theta_1, \theta_2 \in \Theta: \|\theta_1 - \theta_2\| < \delta} \|\nu_n s(\theta_1) - \nu_n s(\theta_2)\| > \eta] < \varepsilon \forall \gamma_0 \in \Gamma$.

(b) Part (a) holds if Assumption 2(i) is replaced by Assumption 2*(i) and q is replaced by $2 + \delta$ for some $\delta > 0$.

Lemma A.3 (central limit theorem)

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

(b) Part (a) holds if Assumption 2(i) is replaced by Assumption 2*(i) and q is replaced by $2 + \delta$ for some $\delta > 0$.

Lemmas A.1-A.3 are proved in Lemmas 11.3-11.5 in the supplemental appendix of Andrews and Cheng (2013) for the strong mixing arrays. Lemma A.1 automatically extends to the i.i.d. data. Lemma A.2 holds for the i.i.d. data with q replaced by $2 + \delta$ by applying stochastic equicontinuity results for the type II class (Lipschitz functions) in Andrews (1994). Lemma A.3 extends to i.i.d. data with q replaced by $2 + \delta$ following the Lyapunov central limit theorem for row-wise i.i.d. triangular arrays.

B Proofs for Asymptotic Distributions of Estimators and Test Statistics

Proof of Lemma 1. The sample least squares criterion function is

$$\begin{aligned} Q_n(\theta) &= n^{-1} \sum_{t=1}^n U_t^2(\theta)/2, \text{ where} \\ U_t(\theta) &= Y_t - g(X_t, \pi)' \beta - Z_t' \zeta \\ &= U_t + g(X_t, \pi_n)' \beta_n + Z_t' \zeta_n - g(X_t, \pi)' \beta - Z_t' \zeta. \end{aligned} \quad (\text{B.1})$$

Applying Lemma A.1, $Q_n(\theta)$ converges to a non-random function $Q(\theta)$ uniformly over $\theta \in \Theta$. The population criterion function is

$$Q(\theta) = \mathbb{E}_{\gamma_0} U_t^2/2 + \mathbb{E}_{\gamma_0} [g(X_t, \pi_0)' \beta_0 + Z_t' \zeta_0 - g(X_t, \pi)' \beta - Z_t' \zeta]^2/2 \quad (\text{B.2})$$

and $Q(\theta)$ is continuous in θ on Θ . Note that $\beta_{\mathcal{I}_1,0} \neq 0$ and $\beta_{\mathcal{I}_k,0} = 0$ for $k > 1$ by the group specification.

Define

$$\psi = (\beta', \zeta')'. \quad (\text{B.3})$$

Let ψ_n denote the true value of ψ for sample size n and $\psi_n \rightarrow \psi_0$. We write the criterion function $Q(\theta)$ as $Q(\psi, \pi_1|\pi_{1+})$ and analyze $Q(\psi, \pi_1|\pi_{1+})$ as a function of (ψ, π_1) for a fixed value of π_{1+} .

Now we show that for any π_{1+} , $Q(\psi, \pi_1|\pi_{1+})$ is uniquely minimized by $(\psi_0, \pi_{1,0})$. Note that $\beta_{k,0} = 0$ for $k > 1$ by the grouping rule. Therefore, $Q(\psi_0, \pi_{1,0}|\pi_{1+}) = 0$. For fixed π_{1+} ,

$$\begin{aligned} &Q(\psi, \pi_1|\pi_{1+}) - Q(\psi_0, \pi_{1,0}|\pi_{1+}) \\ &= \mathbb{E}_{\gamma_0} \left[g_1(X_t, \pi_{1,0})' \beta_{1,0} - g_1(X_t, \pi_1)' \beta_1 - \sum_{k>1}^K g_k(X_t, \pi_k)' \beta_k + Z_t' (\zeta_0 - \zeta) \right]^2 / 2. \end{aligned} \quad (\text{B.4})$$

By Assumption 3,

$$\mathbb{P}_{\gamma_0} (a' [g(X_t, \pi)', g(X_t, \pi_0)', Z_t] = 0) < 1 \quad (\text{B.5})$$

for any $a \neq 0$ and $\pi \neq \pi_0$. Because $\beta_{1,0} \neq 0$, the rhs of (B.4) is greater than 0 for any $\pi_1 \neq \pi_{1,0}$. When $\pi_1 = \pi_{1,0}$, (B.5) implies that the rhs of (B.4) is greater than 0 unless $\beta = \beta_0$ and $\zeta = \zeta_0$.

Given that (i) the population criterion function $Q(\psi, \pi_1 | \pi_{1+})$ is uniquely minimized by $(\psi_0, \pi_{1,0})$ for any π_{1+} , (ii) $Q(\psi, \pi_1 | \pi_{1+})$ is continuous, and (iii) the parameter spaces are all compact, we have the identification uniqueness condition

$$\inf_{\pi_{1+} \in \Pi_{1+}} \inf_{\psi \in \Psi, \pi_1 \in \Pi_1} \{Q(\psi, \pi_1 | \pi_{1+}) - Q(\psi_0, \pi_{1,0} | \pi_{1+})\} > 0 \quad (\text{B.6})$$

uniformly over Π_{1+} , following Lemma 8.1 in the supplemental appendix of Andrews and Cheng (2012). Finally, (B.6) implies the uniform consistency of $\widehat{\psi}(\pi_{1+})$ and $\widehat{\pi}_1(\pi_{1+})$ by Lemma 3.1 of Andrews and Cheng (2012). This Lemma extends the consistency proof for extremum estimators to uniform consistency. \square

Proof of Lemma 2. The proof is by induction. Step 1 shows that Lemma 2(b) and 2(c) hold for $k = 1$. Step 2 shows that, if Lemma 2(b) and 2(c) hold for $k - 1$, Lemma 2(a)-2(c) hold for k .

Step 1. For $k = 1$, Lemma 2(b) is

$$\sup_{\pi_{1+} \in \Pi_{1+}} \|\widehat{\pi}_1(\pi_{1+}) - \pi_{1,n}\| \rightarrow_p 0, \quad (\text{B.7})$$

which follows from Lemma 1. For $k = 1$, Lemma 2(c) becomes

$$\|\beta_{1,n}\|^{-1} \begin{pmatrix} \widehat{\beta}_1(\pi_{1+}) - \beta_{1,n} \\ \widehat{\beta}_{1+}(\pi_{1+}) \\ \widehat{\zeta} - \zeta_n \end{pmatrix} \rightarrow_p 0 \quad (\text{B.8})$$

uniformly over π_{1+} , which follows from Lemma 1, $\beta_{1,n} \rightarrow \beta_{1,0} \neq 0$, and $\beta_{k,n} \rightarrow \beta_{k,0} = 0$ for $k > 1$.

Step 2. Suppose Lemma 2 holds for $k - 1$. For $\psi_{k-} = (\beta', \zeta', \pi_1', \dots, \pi_{k-1}')'$, the result for $k - 1$ yields uniform consistency of $\widehat{\psi}_{k-}(\pi_k, \pi_{k+})$ over (π_k, π_{k+}) . Now we show Lemma 2 holds for k .

Let $D_{\psi_k}^1(\theta)$ and $D_{\psi_k}^2(\theta)$ denote the first and second order partial derivatives of $Q_n(\theta)$ wrt ψ_{k-} ,

where $\theta = (\psi'_{k-}, \pi'_k, \pi'_{k+})'$. In this model, the first order derivative is

$$\begin{aligned} D_{\psi_k}^1(\theta) &= -n^{-1} \sum_{t=1}^n \mathbf{B}(\beta_{k-}) d_{\psi_k,t}(\pi) U_t(\theta), \text{ where} \\ \mathbf{B}(\beta_{k-}) &= \text{diag}\{(\mathbf{1}_{d_\beta+d_\zeta}, \beta'_{k-})'\}, \\ d_{\psi_k,t}(\pi) &= (g(X_t, \pi)', Z'_t, g_{\pi_{k-}}(X_t, \pi_{k-})')'. \end{aligned} \quad (\text{B.9})$$

The second order derivative is

$$\begin{aligned} D_{\psi_k}^2(\theta) &= \mathbf{B}(\beta_{k-}) \left(n^{-1} \sum_{t=1}^n d_{\psi_k,t}(\pi) d_{\psi_k,t}(\pi)' - n^{-1} \sum_{t=1}^n d_{\psi_k,t}^*(\theta) U_t(\theta) \right) \mathbf{B}(\beta_{k-}), \text{ where} \quad (\text{B.10}) \\ d_{\psi_k,t}^*(\theta) &= \begin{pmatrix} 0_{d_\beta+d_\zeta} & \text{diag}\{g_{\pi_{k-}}(X_t, \pi_{k-})\} [\text{diag}\{\beta_{k-}\}]^{-1} \\ \text{diag}\{g_{\pi_{k-}}(X_t, \pi_{k-})\} [\text{diag}\{\beta_{k-}\}]^{-1} & \text{diag}\{g_{\pi_{k-}}(X_t, \pi_{k-})\} [\text{diag}\{\beta_{k-}\}]^{-1} \end{pmatrix}. \end{aligned}$$

Recall that

$$\psi_{k-,n}^0 = (\beta_{k-,n}, \beta_k^0, \beta_{k+}^0, \zeta_n, \pi_{k-,n}), \text{ where } \beta_k^0 = 0 \text{ and } \beta_{k+}^0 = 0. \quad (\text{B.11})$$

We set $\beta_k^0 = 0$ and $\beta_{k+}^0 = 0$ in $\psi_{k-,n}^0$ so that the criterion function $Q_n(\theta)$ does not depend on (π_k, π_{k+}) when evaluated at $\psi_{k-,n}^0$. Hence, we write $Q_n(\psi_{k-,n}^0) = Q_n(\psi_{k-,n}^0, \pi_k, \pi_{k+})$.

Part (a). Because $\widehat{\psi}_{k-}(\pi_k, \pi_{k+})$ minimizes $Q_n(\psi_{k-}, \pi_k, \pi_{k+})$ for any (π_k, π_{k+}) , a mean-value expansion of the first order condition (FOC) around $\psi_{k-} = \psi_{k-,n}^0$ implies that

$$\begin{aligned} 0 &= D_{\psi_k}^1(\widehat{\psi}_{k-}(\pi_k, \pi_{k+}), \pi_k, \pi_{k+}) \\ &= D_{\psi_k}^1(\psi_{k-,n}^0, \pi_k, \pi_{k+}) + D_{\psi_k}^2(\psi_{k-,n}^*, \pi_k, \pi_{k+}) \left(\widehat{\psi}_{k-}(\pi_k, \pi_{k+}) - \psi_{k-,n}^0 \right), \end{aligned} \quad (\text{B.12})$$

for some $\psi_{k-,n}^*$ between $\widehat{\psi}_{k-}(\pi_k, \pi_{k+})$ and $\psi_{k-,n}^0$ ($\psi_{k-,n}^*$ may depend on π_k and π_{k+}). This expansion implies that

$$\widehat{\psi}_{k-}(\pi_k, \pi_{k+}) - \psi_{k-,n}^0 = - \left[D_{\psi_k}^2(\psi_{k-,n}^*, \pi_k, \pi_{k+}) \right]^{-1} D_{\psi_k}^1(\psi_{k-,n}^0, \pi_k, \pi_{k+}). \quad (\text{B.13})$$

We first study the first-order partial derivative in (B.13). Normalize it by $[\mathbf{B}(\beta_{k-,n})]^{-1}$,

$$\begin{aligned} &[\mathbf{B}(\beta_{k-,n})]^{-1} D_{\psi_k}^1(\psi_{k-,n}^0, \pi_k, \pi_{k+}) \\ &= -n^{-1} \sum_{t=1}^n d_{\psi_k,t}(\pi_{k-,n}, \pi_k, \pi_{k+}) [g_k(X_t, \pi_{k,n})' \beta_{k,n} + g_{k+}(X_t, \pi_{k+,n})' \beta_{k+,n} + U_t]. \end{aligned} \quad (\text{B.14})$$

We normalize both sides of (B.14) by $\|\beta_{k,n}\|^{-1}$ and obtain

$$\begin{aligned} \|\beta_{k,n}\|^{-1} \left([\mathbf{B}(\beta_{k-,n})]^{-1} D_{\psi_k}^1(\psi_{k-,n}^0, \pi_k, \pi_{k+}) \right) &\rightarrow_p -\Phi_k(\pi_k, \pi_{k,0}|\pi_{k+})\omega_{k,0}, \text{ where} \quad (\text{B.15}) \\ \Phi_k(\pi_k, \pi_{k,0}|\pi_{k+}) &= \mathbb{E}_{\gamma_0} d_{\psi_k,t}(\pi_{k-,0}, \pi_k, \pi_{k+})g_k(X_t, \pi_{k,0})'. \end{aligned}$$

The convergence follows from (i) applying Lemma A.1 to $n^{-1} \sum_{t=1}^n d_{\psi_k,t}(\pi_{k-,n}, \pi_k, \pi_{k+})g_k(X_t, \pi_{k,n})'$ and $n^{-1} \sum_{t=1}^n d_{\psi_k,t}(\pi_{k-,n}, \pi_k, \pi_{k+})g_{k+}(X_t, \pi_{k+,n})'$, (ii) applying Lemmas A.2 and A.3 to the empirical process $n^{-1/2} \sum_{t=1}^n d_{\psi_k,t}(\pi_{k-,n}, \pi_k, \pi_{k+})U_t$, (iii) $\beta_{k+,n} = o(\|\beta_{k,n}\|)$, and (iv) $\|n^{1/2}\beta_{k,n}\| \rightarrow \infty$. Note that $\Phi_k(\pi_k, \pi_{k,0}|\pi_{k+}) = H_k(\pi_k, \pi_{k,0}|\pi_{k+})S_k$, where $H_k(\pi_k, \pi_{k,0}|\pi_{k+})$ is defined in (3.14) and S_k is a selector matrix such that $g_k(X_t, \pi_{k,0}) = S_k' d_{\psi_k,t}(\pi_{k-,0}, \pi_k, \pi_{k+})$.

Next we study the second-order partial derivative in (B.13). Pre- and post-multiply $D_{\psi_k}^2(\theta)$ by $[\mathbf{B}(\beta_{k-})]^{-1}$,

$$[\mathbf{B}(\beta_{k-})]^{-1} D_{\psi_k}^2(\theta) [\mathbf{B}(\beta_{k-})]^{-1} = n^{-1} \sum_{t=1}^n d_{\psi_k,t}(\pi) d_{\psi_k,t}(\pi)' - n^{-1} \sum_{t=1}^n d_{\psi_k,t}^*(\theta) U_t(\theta). \quad (\text{B.16})$$

Lemma A.1 implies uniform convergence of the first term on the rhs. Now we show the second term on the rhs is negligible, i.e.,

$$n^{-1} \sum_{t=1}^n U_t(\theta) d_{\psi_k,t}^*(\theta) = o_p(1) \text{ at } \theta = (\psi_{k-,n}^{*'}, \pi_k', \pi_{k+}')', \quad (\text{B.17})$$

uniformly over (π_k, π_{k+}) , where $\psi_{k-,n}^{*}$ is between $\widehat{\psi}_{k-}(\pi_k, \pi_{k+})$ and $\psi_{k-,n}^0$. Let $\mathcal{I}_{k-} = \mathcal{I}_1 \cup \dots \cup \mathcal{I}_{k-1}$ denote the indices of regressors in groups 1 to $k-1$. Given the definition of $d_{\psi_k,t}^*(\theta)$, it is sufficient to show that for $j \in \mathcal{I}_{k-}$,

$$n^{-1} \sum_{t=1}^n [g_{\pi_j}(X_t, \pi_j) + g_{\pi\pi_j}(X_t, \pi_j)] U_t(\theta) / \beta_j = o_p(1) \quad (\text{B.18})$$

uniformly over (π_k, π_{k+}) when evaluated at $\theta = (\psi_{k-,n}^{*'}, \pi_k', \pi_{k+}')'$. Note that this analysis is element by element for each $j = 1, \dots, p$ rather than by groups.

Next we show (B.18) holds for $j \in \mathcal{I}_{k-}$. To differentiate a single element β_j from a group, we use β_ℓ to denote group ℓ . For $j \in \mathcal{I}_{k-}$ and $\ell = k-1$, we have the following results:

$$\begin{aligned} \frac{\|\beta_{\ell,n}\|}{\beta_{j,n}} &= O_p(1) \text{ and} \\ \frac{\|\beta_{\ell,n}\|}{\widehat{\beta}_{j,n}(\pi_k, \pi_{k+})} &= \left(\frac{\widehat{\beta}_{j,n}(\pi_k, \pi_{k+}) - \beta_{j,n}}{\|\beta_{\ell,n}\|} + \frac{\beta_{j,n}}{\|\beta_{\ell,n}\|} \right)^{-1} = O_p(1), \quad (\text{B.19}) \end{aligned}$$

because (i) the coefficients $\beta_{j,n}$ are grouped in a decreasing order and (ii) Lemma 2(c) applies to $\ell = k - 1$. Given (B.19), we have

$$\frac{\|\beta_{\ell,n}\|}{\beta_j} = O_p(1) \quad (\text{B.20})$$

for any β_j between $\beta_{j,n}$ and $\widehat{\beta}_{j,n}(\pi_k, \pi_{k+})$. For $\ell = k - 1$, the error $U_t(\theta)$ can be written as

$$\begin{aligned} U_t(\theta) &= [U_t + g_{\ell^-}(X_t, \pi_{\ell^-,n})' \beta_{\ell^-,n} + g_{\ell}(X_t, \pi_{\ell,n})' \beta_{\ell,n} + g_{\ell^+}(X_t, \pi_{\ell^+,n}) \beta_{\ell^+,n}] \\ &\quad - [g_{\ell^-}(X_t, \pi_{\ell^-})' \beta_{\ell^-} + g_{\ell}(X_t, \pi_{\ell})' \beta_{\ell} + g_{\ell^+}(X_t, \pi_{\ell^+}) \beta_{\ell^+}]. \end{aligned} \quad (\text{B.21})$$

Using this expansion, write

$$n^{-1} \sum_{t=1}^n g_{\pi_j}(X_t, \pi_j) U_t(\theta) / \beta_j = (A_j + B_j + C_j) \frac{\|\beta_{\ell,n}\|}{\beta_j}, \quad (\text{B.22})$$

where $\|\beta_{\ell,n}\|/\beta_j = O_p(1)$ following (B.20) and A_j, B_j, C_j are specified as follows. The first term is

$$A_j = \frac{n^{-1/2} \sum_{t=1}^n g_{\pi_j}(X_t, \pi_j) U_t}{n^{1/2} \|\beta_{\ell,n}\|}. \quad (\text{B.23})$$

The second term is

$$\begin{aligned} B_j &= n^{-1} \sum_{t=1}^n [g_{\pi_j}(X_t, \pi_j) g_{\ell^-}(X_t, \pi_{\ell^-,n})]' \frac{\beta_{\ell^-,n}}{\|\beta_{\ell,n}\|} - n^{-1} \sum_{t=1}^n [g_{\pi_j}(X_t, \pi_j) g_{\ell^-}(X_t, \pi_{\ell^-})]' \frac{\beta_{\ell^-}}{\|\beta_{\ell,n}\|} \\ &= n^{-1} \sum_{t=1}^n [g_{\pi_j}(X_t, \pi_j) (g_{\ell^-}(X_t, \pi_{\ell^-,n}) - g_{\ell^-}(X_t, \pi_{\ell^-}))]' \frac{\beta_{\ell^-,n}}{\|\beta_{\ell,n}\|} \\ &\quad - n^{-1} \sum_{t=1}^n [g_{\pi_j}(X_t, \pi_j) g_{\ell^-}(X_t, \pi_{\ell^-})]' \frac{\beta_{\ell^-} - \beta_{\ell^-,n}}{\|\beta_{\ell,n}\|} \\ &= (\pi_{\ell^-,n} - \pi_{\ell^-})' \left[n^{-1} \sum_{t=1}^n g_{\pi_j}(X_t, \pi_j) \text{diag}\{g_{\pi_{\ell^-}}(X_t, \pi_{\ell^{**}})\} \right] \frac{\beta_{\ell^-,n}}{\|\beta_{\ell,n}\|} \\ &\quad - n^{-1} \sum_{t=1}^n [g_{\pi_j}(X_t, \pi_j) g_{\ell^-}(X_t, \pi_{\ell^-})]' \frac{\beta_{\ell^-} - \beta_{\ell^-,n}}{\|\beta_{\ell,n}\|}, \end{aligned} \quad (\text{B.24})$$

for some $\pi_{\ell^-}^{**}$ between $\pi_{\ell^-,n}$ and π_{ℓ^-} by a mean-value expansion. The third term is

$$\begin{aligned} C_j &= \left[n^{-1} \sum_{t=1}^n g_{\pi_j}(X_t, \pi_j) g_{\ell}(X_t, \pi_{\ell,n}) \right]' \frac{\beta_{\ell,n}}{\|\beta_{\ell,n}\|} - \left[n^{-1} \sum_{t=1}^n g_{\pi_j}(X_t, \pi_j) g_{\ell}(X_t, \pi_{\ell}) \right]' \frac{\beta_{\ell}}{\|\beta_{\ell,n}\|} \\ &\quad + \left[n^{-1} \sum_{t=1}^n g_{\pi_j}(X_t, \pi_j) g_{\ell^+}(X_t, \pi_{\ell^+,n}) \right]' \frac{\beta_{\ell^+,n}}{\|\beta_{\ell,n}\|} - \left[n^{-1} \sum_{t=1}^n g_{\pi_j}(X_t, \pi_j) g_{\ell^+}(X_t, \pi_{\ell^+}) \right]' \frac{\beta_{\ell^+}}{\|\beta_{\ell,n}\|}. \end{aligned} \quad (\text{B.25})$$

Now we show $A_j, B_j, C_j = o_p(1)$. Note that the rate of convergence in Lemma 2(c) holds when $\widehat{\psi}_{k^-}(\pi_k, \pi_{k+})$ is replaced by ψ_n . Hence, it also holds for any ψ_{k^-} between $\widehat{\psi}_{k^-}(\pi_k, \pi_{k+})$ and ψ_n . First, $A_j = o_p(1)$ because (i) $n^{-1/2} \sum_{t=1}^n g_{\pi_j}(X_t, \pi_j) U_t = O_p(1)$ uniformly over π_j by Lemma A.2 and Lemma A.3 and (ii) $n^{1/2} \|\beta_{\ell, n}\| \rightarrow \infty$. Second, $B_j = o_p(1)$ because, (i) for ψ_{k^-} between $\widehat{\psi}_{k^-}(\pi_k, \pi_{k+})$ and ψ_n , $(\pi_{\ell^-, n} - \pi_{\ell^-})' \beta_{\ell^-, n} / \|\beta_{\ell, n}\| = o_p(1)$, $(\beta_{\ell^-} - \beta_{\ell^-, n}) / \|\beta_{\ell, n}\| = o_p(1)$ by Lemma 2(c), and (ii) the sample means are $O_p(1)$ by the ULLN in Lemma A.1. Third, $C_j = o_p(1)$ holds because (i) for ψ_{k^-} between $\widehat{\psi}_{k^-}(\pi_k, \pi_{k+})$ and ψ_n , $\beta_{\ell, n} / \|\beta_{\ell, n}\| \rightarrow \omega_{k, 0}$, $\beta_{\ell} / \|\beta_{\ell, n}\| \rightarrow \omega_{k, 0}$, $\beta_{\ell^+, n} / \|\beta_{\ell, n}\| \rightarrow 0$, $\beta_{\ell^+} / \|\beta_{\ell, n}\| \rightarrow 0$ and (ii) the sample means are $O_p(1)$ by the ULLN in Lemma A.1. Similarly, (B.22) holds when $g_{\pi_j}(X_t, \pi_j)$ is replaced by $g_{\pi\pi, j}(X_t, \pi_j)$. This proves (B.18), which in turn implies (B.17).

It follows from (B.16) and (B.17) that, for $\theta = (\psi'_{k^-}, \pi'_k, \pi'_{k+})'$, where ψ_{k^-} is between $\widehat{\psi}_{k^-}(\pi_k, \pi_{k+})$ and $\psi_{k^-, n}^0$, the normalized second order partial derivative satisfies

$$\begin{aligned} [\mathbf{B}(\beta_{k^-})]^{-1} D_{\psi_k}^2(\theta) [\mathbf{B}(\beta_{k^-})]^{-1} &\rightarrow_p H_k(\pi_k, \pi_k | \pi_{k+}), \text{ where} \\ H_k(\pi_k, \pi_k | \pi_{k+}) &= \mathbb{E}_{\gamma_0} d_{\psi_k, t}(\pi_{k^-, 0}, \pi_k | \pi_{k+}) d_{\psi_k, t}(\pi_{k^-, 0}, \pi_k | \pi_{k+})'. \end{aligned} \quad (\text{B.26})$$

Next we show

$$[\mathbf{B}(\beta_{k^-, n})]^{-1} [\mathbf{B}(\beta_{k^-})] \rightarrow_p I_{d_{\beta} + d_{\zeta} + d_{k^-}}, \quad (\text{B.27})$$

where d_{k^-} is the number of elements in β_{k^-} , so that rescaling by $\mathbf{B}(\beta_{k^-})$ and by $\mathbf{B}(\beta_{k^-, n})$ is asymptotically equivalent. For $j \in \mathcal{I}_{k^-}$,

$$\frac{\widehat{\beta}_j(\pi_k, \pi_{k+})}{\beta_{j, n}} - 1 = \frac{\widehat{\beta}_j(\pi_k, \pi_{k+}) - \beta_{j, n} \|\beta_{k-1, n}\|}{\|\beta_{k-1, n}\| \beta_{j, n}} \rightarrow 0 \quad (\text{B.28})$$

by applying Lemma 2(c) to $k-1$. This implies that for $j \in \mathcal{I}_{k^-}$, $\beta_j / \beta_{j, n} \rightarrow 1$ for any β_j between $\widehat{\beta}_j(\pi_k, \pi_{k+})$ and $\beta_{j, n}$, which further implies the desired result in (B.27).

Normalizing the equality in (B.13), we obtain

$$\begin{aligned} &\mathbf{B}(\beta_{k^-, n}) \left(\widehat{\psi}_{k^-}(\pi_k, \pi_{k+}) - \psi_{k^-, n}^0 \right) \\ &= - \left\{ [\mathbf{B}(\beta_{k^-, n})]^{-1} D_{\psi_k}^2(\psi_{k^-, n}^*, \pi_k, \pi_{k+}) [\mathbf{B}(\beta_{k^-, n})]^{-1} \right\}^{-1} \left\{ [\mathbf{B}(\beta_{k^-, n})]^{-1} D_{\psi_k}^1(\psi_{k^-, n}^0, \pi_k, \pi_{k+}) \right\}. \end{aligned} \quad (\text{B.29})$$

Applying (B.15), (B.26), and (B.27) to (B.29) yields

$$\begin{aligned}
& \|\beta_{k,n}\|^{-1} \left(\mathbf{B}(\beta_{k^-,n}) \left(\widehat{\psi}_{k^-}(\pi_k, \pi_{k^+}) - \psi_{k^-,n}^0 \right) \right) \\
& \rightarrow_p [H_k(\pi_k, \pi_k | \pi_{k^+})]^{-1} \Phi_k(\pi_k, \pi_{k,0} | \pi_{k^+}) \omega_{k,0} \\
& = [H_k(\pi_k, \pi_k | \pi_{k^+})]^{-1} H_k(\pi_k, \pi_{k,0} | \pi_{k^+}) \Delta_k
\end{aligned} \tag{B.30}$$

uniformly over (π_k, π_{k^+}) , where $\Delta_k = S_k \omega_{k,0}$ by definition.

We expand the criterion function $Q_n^c(\pi_k, \pi_{k^+}) = Q_n(\widehat{\psi}_{k^-}(\pi_k, \pi_{k^+}), \pi_k, \pi_{k^+})$ around $(\psi_{k^-,n}^0, \pi_k, \pi_{k^+})$ for fixed (π_k, π_{k^+}) . Note that $Q_n(\psi_{k^-,n}^0) = Q_n(\psi_{k^-,n}^0, \pi_k, \pi_{k^+})$ does not depend on (π_k, π_{k^+}) and we have shown the consistency of $\widehat{\psi}_{k^-}(\pi_k, \pi_{k^+})$. By a second order Taylor expansion,

$$\begin{aligned}
& Q_n^c(\pi_k, \pi_{k^+}) - Q_n(\psi_{k^-,n}^0) \\
& = D_{\psi_{k^-}}^1(\psi_{k^-,n}^0, \pi_k, \pi_{k^+})' \left(\widehat{\psi}_{k^-}(\pi_k, \pi_{k^+}) - \psi_{k^-,n}^0 \right) \\
& \quad + \frac{1}{2} \left(\widehat{\psi}_{k^-}(\pi_k, \pi_{k^+}) - \psi_{k^-,n}^0 \right)' D_{\psi_{k^-}}^2(\psi_{k^-,n}^{**}, \pi_k, \pi_{k^+}) \left(\widehat{\psi}_{k^-}(\pi_k, \pi_{k^+}) - \psi_{k^-,n}^0 \right) \\
& = \left(D_{\psi_{k^-}}^1(\psi_{k^-,n}^0, \pi_k, \pi_{k^+})' [\mathbf{B}(\beta_{k^-,n})]^{-1} \right) \left(\mathbf{B}(\beta_{k^-,n}) \left(\widehat{\psi}_{k^-}(\pi_k, \pi_{k^+}) - \psi_{k^-,n}^0 \right) \right) \\
& \quad + \frac{1}{2} \left(\mathbf{B}(\beta_{k^-,n}) \left(\widehat{\psi}_{k^-}(\pi_k, \pi_{k^+}) - \psi_{k^-,n}^0 \right) \right)' \left([\mathbf{B}(\beta_{k^-,n})]^{-1} D_{\psi_{k^-}}^2(\psi_{k^-,n}^{**}, \pi_2) [\mathbf{B}(\beta_{k^-,n})]^{-1} \right) \\
& \quad \times \left(\mathbf{B}(\beta_{k^-,n}) \left(\widehat{\psi}_{k^-}(\pi_k, \pi_{k^+}) - \psi_{k^-,n}^0 \right) \right)
\end{aligned} \tag{B.31}$$

for some $\psi_{k^-,n}^{**}$ between $\widehat{\psi}_{k^-}(\pi_k, \pi_{k^+})$ and $\psi_{k^-,n}^0$. Applying the results for the first and second order derivatives in (B.15) and (B.26) and the results for $\mathbf{B}(\beta_{k^-,n})(\widehat{\psi}_{k^-}(\pi_k, \pi_{k^+}) - \psi_{k^-,n}^0)$ in (B.29) and (B.30), we obtain the desired result in part (a).

Part (b). Following the definitions of $H_k(\pi_k, \pi_k | \pi_{k^+})$ and $\Delta_k = [0_{1 \times d_{k^-}}, \omega'_{k,0}, 0_{1 \times (d_c + d_{k^-})}]'$, the matrix Cauchy-Schwarz inequality (see Tripathi (1999)) implies that $\Delta_k' H_k(\pi_k, \pi_{k,0} | \pi_{k^+})' [H_k(\pi_k, \pi_k | \pi_{k^+})]^{-1} H_k(\pi_k, \pi_{k,0} | \pi_{k^+}) \Delta_k$ is uniquely maximized at $\pi_k = \pi_{k,0}$ provided that for $a \neq 0$ and some $\varepsilon > 0$,

$$\mathbb{P}_\gamma \left(a [\omega'_{k,0} g_k(X_t, \pi_{k,0})] + b' [g_{k^-}(X_t, \pi_{k^-,0}), g_k(X_t, \pi_k), g_{k^+}(X_t, \pi_{k^+}), Z_t', g_{\pi_{k^-}}(X_t, \pi_{k^-,0})] = 0 \right) \leq 1 - \varepsilon \tag{B.32}$$

for $\pi_k \neq \pi_{k,0}$. Because each element in $\omega_{k,0}$ is different from 0 following the grouping rule, the desired result in (B.32) is implied by Assumption 3. Thus, part (b) follows from part(a), the argmax CMT (Theorem 3.2.2 in van der Vaart and Wellner (1996)(1996, p. 286)), and $\pi_{k,n} \rightarrow \pi_{k,0}$ as $n \rightarrow \infty$.

Part (c). Part (c) follows from (B.30), the consistency in part (b), and replacing $\beta_{k,n}^0$, which

is a vector of zeros, with $\beta_{k,n}$ in the centering term.

This completes the proof of step 2 in the induction arguments, and therefore completes the proof of Lemma 2. \square

Proof of Theorem 1. Part (a). For $k = K$, normalizing (B.14) by $n^{1/2}$ yields

$$\begin{aligned} & n^{1/2} [\mathbf{B}(\beta_{K^-,n})]^{-1} D_{\psi_K}^1(\psi_{K^-,n}^0, \pi_K) \\ &= -n^{-1} \sum_{t=1}^n d_{\psi_K,t}(\pi_{K^-,n}, \pi_K) g_K(X_t, \pi_{K,n})' \left(n^{1/2} \beta_{K,n} \right) - n^{-1/2} \sum_{t=1}^n U_t d_{\psi_K,t}(\pi_{K^-,n}, \pi_K) \\ &\Rightarrow -[H_K(\pi_K, \pi_{K,0}) S_K b_K + G(\pi_K)] \end{aligned} \quad (\text{B.33})$$

following Lemmas A.1-A.3 and $n^{1/2} \beta_{K,n} \rightarrow b_K$. For $k = K$, (B.26) yields

$$[\mathbf{B}(\beta_{K^-})]^{-1} D_{\psi_K}^2(\theta) [\mathbf{B}(\beta_{K^-})]^{-1} \rightarrow_p H_K(\pi_K, \pi_K) \quad (\text{B.34})$$

for any $\theta = (\psi'_{K^-}, \pi'_K)'$ where ψ_{K^-} is between $\widehat{\psi}_{K^-}(\pi_K)$ and $\psi_{K^-,n}^0$. In addition, (B.27) gives

$$[\mathbf{B}(\beta_{K^-,n})]^{-1} [\mathbf{B}(\beta_{K^-})] \rightarrow_p I_{d_\beta + d_\zeta + d_{K^-}}. \quad (\text{B.35})$$

For $k = K$, normalizing (B.29) by $n^{1/2}$, we obtain

$$\begin{aligned} & n^{1/2} \mathbf{B}(\beta_{K^-,n}) \left(\widehat{\psi}_{K^-}(\pi_K) - \psi_{K^-,n}^0 \right) \\ &= - \left([\mathbf{B}(\beta_{K^-,n})]^{-1} D_{\psi_K}^2(\psi_{K^-,n}^*, \pi_K) [\mathbf{B}(\beta_{K^-,n})]^{-1} \right)^{-1} n^{1/2} [\mathbf{B}(\beta_{K^-,n})]^{-1} D_{\psi_K}^1(\psi_{K^-,n}^0, \pi_K) \end{aligned} \quad (\text{B.36})$$

for $\psi_{K^-,n}^*$ between $\widehat{\psi}_{K^-}(\pi_K)$ and $\psi_{K^-,n}^0$. Combining (B.33)-(B.36) yields

$$\begin{aligned} & n^{1/2} \mathbf{B}(\beta_{K^-,n}) \left(\widehat{\psi}_{K^-}(\pi_K) - \psi_{K^-,n}^0 \right) \Rightarrow \tau(\pi_K), \text{ where} \\ & \tau(\pi_K) = [H_K(\pi_K, \pi_K)]^{-1} [H_K(\pi_K, \pi_{K,0}) S_K b_K + G(\pi_K)]. \end{aligned} \quad (\text{B.37})$$

Applying (B.31) to $k = K$ and normalizing the criterion function by n , we obtain

$$\begin{aligned}
& n \left(Q_n^c(\pi_K) - Q_n(\psi_{K^-,n}^0) \right) \\
&= \left(n^{1/2} D_{\psi_K}^1(\psi_{K^-,n}^0, \pi_K)' [\mathbf{B}(\beta_{K^-,n})]^{-1} \right) \left(n^{1/2} \mathbf{B}(\beta_{K^-,n}) \left(\widehat{\psi}_{K^-(\pi_K)} - \psi_{K^-,n}^0 \right) \right) \\
&+ \frac{1}{2} \left(n^{1/2} \mathbf{B}(\beta_{K^-,n}) \left(\widehat{\psi}_{K^-(\pi_K)} - \psi_{K^-,n}^0 \right) \right)' \left([\mathbf{B}(\beta_{K^-,n})]^{-1} D_{\psi_K}^2(\psi_{K^-,n}^{**}, \pi_K) [\mathbf{B}(\beta_{K^-,n})]^{-1} \right) \\
&\times \left(n^{1/2} \mathbf{B}(\beta_{K^-,n}) \left(\widehat{\psi}_{K^-(\pi_K)} - \psi_{K^-,n}^0 \right) \right) \tag{B.38} \\
&\Rightarrow -\frac{1}{2} [H_K(\pi_K, \pi_{K,0}) S_K b_K + G_K(\pi_K)]' [H_K(\pi_K, \pi_K)]^{-1} [H_K(\pi_K, \pi_{K,0}) S_K b_K + G_K(\pi_K)]
\end{aligned}$$

following (B.33), (B.34), and (B.37). Because $\widehat{\pi}_K$ minimizes $Q_n^c(\pi_K)$, applying the argmax CMT, we obtain

$$\widehat{\pi}_K \Rightarrow \pi_K^*. \tag{B.39}$$

Because $\widehat{\psi}_{K^-(\widehat{\pi}_K)} = \widehat{\psi}_{K^-}$, the CMT and (B.37) yield

$$\begin{aligned}
& n^{1/2} \mathbf{B}(\beta_{K^-,n}) \left(\widehat{\psi}_{K^-} - \psi_{K^-,n} \right) \\
&= n^{1/2} \mathbf{B}(\beta_{K^-,n}) \left(\widehat{\psi}_{K^-(\widehat{\pi}_K)} - \psi_{K^-,n}^0 \right) - n^{1/2} \mathbf{B}(\beta_{K^-,n}) \left(\psi_{K^-,n} - \psi_{K^-,n}^0 \right) \\
&\Rightarrow \tau_K(\pi_K^*) - S_K b_K, \tag{B.40}
\end{aligned}$$

where $S_K b_K$ is a vector of the same size as ψ_{K^-} but with the sub-vector of β_K replaced by b_K and the rest replaced by zeros. The convergence in (B.39) and (B.40) hold jointly because there are both functionals of the same underlying stochastic processes. This completes the proof. \square

Part (b). When $\|n^{1/2} \beta_{K,n}\| \rightarrow \infty$, Lemma 2 applies to $k = K$ with π_{k^+} omitted in the expression. This provides (i) consistency of $\widehat{\theta}$ and (ii) the rate of convergence in Lemma 2 (c) with $k = K$.

Define the first and second order derivatives of $Q_n(\theta)$ wrt θ by

$$\begin{aligned}
D_{\theta}^1(\theta) &= -n^{-1} \sum_{t=1}^n U_t(\theta) \mathbf{B}(\beta) d_{\theta,t}(\pi), \text{ with} \\
\mathbf{B}(\beta_k) &= \text{diag}\{(1_{d_{\beta}+d_{\zeta}}, \beta')'\}, \\
d_{\theta,t}(\pi) &= (g(X_t, \pi)', Z_t', g_{\pi}(X_t, \pi)')'. \tag{B.41}
\end{aligned}$$

and

$$D_{\theta}^2(\theta) = \mathbf{B}(\beta) \left(n^{-1} \sum_{t=1}^n d_{\theta,t}(\pi) d_{\theta,t}(\pi)' - n^{-1} \sum_{t=1}^n U_t(\theta) d_{\theta,t}^*(\theta) \right) \mathbf{B}(\beta), \text{ where} \quad (\text{B.42})$$

$$d_{\theta,t}^*(\theta) = \begin{pmatrix} 0_{d_{\beta}+d_{\zeta}} & \text{diag}\{g_{\pi}(X_t, \pi)\} [\text{diag}\{\beta\}]^{-1} \\ \text{diag}\{g_{\pi}(X_t, \pi)\} [\text{diag}\{\beta\}]^{-1} & \text{diag}\{g_{\pi\pi}(X_t, \pi)\} [\text{diag}\{\beta\}]^{-1} \end{pmatrix}.$$

Because $\hat{\theta}$ minimizes $Q_n(\theta)$, a mean-value expansion of the FOC around θ_n implies that

$$\hat{\theta} - \theta_n = - [D_{\theta}^2(\theta^*)]^{-1} D_{\theta}^1(\theta_n). \quad (\text{B.43})$$

for some θ^* between $\hat{\theta}$ and θ_n .

Evaluate $D_{\theta}^1(\theta)$ at θ_n and normalize it by $n^{1/2} [\mathbf{B}(\beta_n)]^{-1}$,

$$n^{1/2} [\mathbf{B}(\beta_n)]^{-1} D_{\theta}^1(\theta_n) \rightarrow_d N(0, \Omega_{\theta}(\pi_0)). \quad (\text{B.44})$$

Pre- and post-multiply $D_{\theta}^2(\theta)$ by $[\mathbf{B}(\beta)]^{-1}$,

$$[\mathbf{B}(\beta)]^{-1} D_{\theta}^2(\theta) [\mathbf{B}(\beta)]^{-1} = n^{-1} \sum_{t=1}^n d_{\theta,t}(\pi) d_{\theta,t}(\pi)' - n^{-1} \sum_{t=1}^n U_t(\theta) d_{\theta,t}^*(\theta), \quad (\text{B.45})$$

where we have

$$n^{-1} \sum_{t=1}^n U_t(\theta) d_{\theta,t}^*(\theta) = o_p(1) \text{ at } \theta = \theta^*, \quad (\text{B.46})$$

for any θ^* between $\hat{\theta}$ and θ_n following the arguments used to show (B.17). It follows that

$$[\mathbf{B}(\beta)]^{-1} D_{\theta}^2(\theta) [\mathbf{B}(\beta)]^{-1} \rightarrow_p H(\pi_0) \quad (\text{B.47})$$

for any θ between $\hat{\theta}$ and θ_n . In addition, Lemma 2(c) for $k = K$ implies that $[\mathbf{B}(\beta_n)]^{-1} \mathbf{B}(\beta^*) \rightarrow_p I_{2d_{\beta}+d_{\zeta}}$ for β^* between $\hat{\beta}$ and β_n .

Putting together results for the first and second order derivatives, we obtain

$$\begin{aligned} n^{1/2} \mathbf{B}(\beta_n) \left(\hat{\theta} - \theta_n \right) &= - ([\mathbf{B}(\beta_n)]^{-1} D_{\theta}^2(\theta^*) [\mathbf{B}(\beta_n)]^{-1})^{-1} n^{1/2} [\mathbf{B}(\beta_n)]^{-1} D_{\theta}^1(\theta_n) \\ &\rightarrow_d N(0, H(\pi_0)^{-1} \Omega_{\theta}(\pi_0) H(\pi_0)^{-1}). \end{aligned} \quad (\text{B.48})$$

Proof of Theorem 2. Under the null hypothesis $H_0 : R\theta_n = v_n$, the Wald statistic $W_n(R)$ is

$$W_n(R) = n \left[R \left(\hat{\theta} - \theta_n \right) \right]' \left[R \mathbf{B}^{-1}(\hat{\beta}) \hat{\Sigma}_n \mathbf{B}^{-1}(\hat{\beta}) R' \right]^{-1} \left[R \left(\hat{\theta} - \theta_n \right) \right]. \quad (\text{B.49})$$

We first show

$$\varepsilon_n = W_n(R) - W_n(R^*) = o_p(1). \quad (\text{B.50})$$

Because $\mathbf{D}^*(\hat{\beta})$ is non-singular with w.p.a.1, $W_n(R) = W_n(\mathbf{D}^*(\hat{\beta})A'R)$ w.p.a.1. Decompose the the rotated matrix $A'R$ as

$$A'R = R^* + \varepsilon_R^*, \quad (\text{B.51})$$

where R^* is the block-diagonal matrix and $\varepsilon_R^* = A'R - R^*$ is composed of the rest. Using this decomposition, we have

$$\begin{aligned} W_n(\mathbf{D}^*(\hat{\beta})A'R) &= \bar{\varrho}' \left[\bar{R} \left(\mathbf{B}^{-1}(\hat{\omega}) \hat{\Sigma}_n \mathbf{B}^{-1}(\hat{\omega}) \right) \bar{R}' \right]^{-1} \bar{\varrho}, \text{ where} \\ \bar{\rho} &= n^{1/2} \mathbf{D}^*(\hat{\beta}) (R^* + \varepsilon_R^*) (\hat{\theta}_n - \theta_n) \\ \bar{R} &= \mathbf{D}^*(\hat{\beta}) (R^* + \varepsilon_R^*) \mathbf{D}^{-1}(\hat{\beta}). \end{aligned} \quad (\text{B.52})$$

Because R^* is block-diagonal,

$$\bar{R} = R^* + \mathbf{D}^*(\hat{\beta}) \varepsilon_R^* \mathbf{D}^{-1}(\hat{\beta}), \quad (\text{B.53})$$

where $\mathbf{D}^*(\hat{\beta}) \varepsilon_R^* \mathbf{D}^{-1}(\hat{\beta}) = o_p(1)$ because (i) the matrix $A'_k R_j$ in ε_R^* is multiplied by $\|\hat{\beta}_k\| \cdot \|\hat{\beta}_j\|^{-1}$, which is $o_p(1)$ for $j < k$ and (ii) $A'R$ is upper block diagonal by construction. To study $\bar{\rho}$, write it as

$$\begin{aligned} \bar{\rho} &= \rho_n + n^{1/2} \mathbf{D}^*(\hat{\beta}) \varepsilon_R^* (\hat{\theta}_n - \theta_n), \text{ where} \\ \rho_n &= n^{1/2} \mathbf{D}^*(\hat{\beta}) R^* (\hat{\theta}_n - \theta_n). \end{aligned} \quad (\text{B.54})$$

The second term $n^{1/2} \mathbf{D}^*(\hat{\beta}) \varepsilon_R^* (\hat{\theta}_n - \theta_n) = o_p(1)$ because its components are $n^{1/2} \|\hat{\beta}_k\| (A'_k R_j) (\hat{\pi}_j - \pi_{j,n})$ for $j < k$. By Theorem 1, the convergence rate of $\hat{\pi}_j$ is $n^{1/2} \|\beta_{j,n}\|$, which is an order of magnitude larger than $n^{1/2} \|\hat{\beta}_k\|$ for $j < k$. Putting together (B.52)-(B.54), we have

$$\begin{aligned} &W_n(R^*) + \varepsilon_n \\ &= W_n(\mathbf{D}^*(\hat{\beta})A'R) \\ &= (\rho_n + o_p(1))' \left[(R^* + o_p(1)) \left(\mathbf{B}^{-1}(\hat{\omega}) \hat{\Sigma}_n \mathbf{B}^{-1}(\hat{\omega}) \right) (R^* + o_p(1))' \right]^{-1} (\rho_n + o_p(1)). \end{aligned} \quad (\text{B.55})$$

Applying $W_n(R^*) = \rho_n' V_n^{-1} \rho_n$ in (4.16) and comparing the first and third line of (B.55) shows that $\varepsilon_n = o_p(1)$ provided that (i) $\rho_n = O_p(1)$, (ii) $\mathbf{B}^{-1}(\widehat{\omega}) \widehat{\Sigma}_n \mathbf{B}^{-1}(\widehat{\omega}) = O_p(1)$, and (iii) $\lambda_{\min}(\mathbf{B}^{-1}(\widehat{\omega}) \widehat{\Sigma}_n^{-1} \mathbf{B}^{-1}(\widehat{\omega})) > 0$ w.p.a.1., given that R^* has full rank by construction. We investigate these terms below.

We first consider weak identification in part (a). Following (4.17), $\rho_n = R^* \mathbf{B}^{-1}(\widehat{\omega}) \xi_n$, where $\xi_n = n^{1/2} \mathbf{B}(\widehat{\beta})(\widehat{\theta} - \theta_n)$. To derive the asymptotic distribution of ξ_n , define a stochastic process indexed by π_K :

$$\xi_n(\pi_K) = \begin{pmatrix} n^{1/2} \mathbf{B}(\widehat{\beta}_{K^-,n}) (\widehat{\psi}_{K^-}(\pi_K) - \psi_{K^-,n}) \\ n^{1/2} \text{diag}\{\widehat{\beta}_K(\pi_K)\} (\pi_K - \pi_{K,n}) \end{pmatrix}. \quad (\text{B.56})$$

Applying (B.27) with $k = K$, we have $\mathbf{B}(\widehat{\beta}_{K^-,n}) [\mathbf{B}(\beta_{K^-,n})]^{-1} = I_{d_{K^-}} + o_p(1)$. Applying it together with Theorem 1(a) and the CMT yields

$$\xi_n = \xi_n(\widehat{\pi}_K) \Rightarrow \xi(\pi_K^*), \quad (\text{B.57})$$

where

$$\xi(\pi_K) = \begin{pmatrix} \tau_K(\pi_K) - S_K b_K \\ \text{diag}\{\tau_{\beta_K}(\pi_K)\} (\pi_K - \pi_{K,0}) \end{pmatrix}. \quad (\text{B.58})$$

To study $\mathbf{B}(\widehat{\omega})$ with $\widehat{\omega} = (\widehat{\omega}'_1, \dots, \widehat{\omega}'_K)'$, note that for $k = 1, \dots, K-1$, $\|\beta_{k,n}\|^{-1} (\widehat{\beta}_k - \beta_{k,n}) = o_p(1)$ following Lemma 2(c). This implies $\widehat{\beta}_k = \beta_{k,n} + \|\beta_{k,n}\| o_p(1)$ and $\|\widehat{\beta}_k\| / \|\beta_{k,n}\| = 1 + o_p(1)$. Hence,

$$\widehat{\omega}_k = \frac{\widehat{\beta}_k}{\|\widehat{\beta}_k\|} = \frac{\widehat{\beta}_k - \beta_{k,n}}{\|\beta_{k,n}\|} \frac{\|\beta_{k,n}\|}{\|\widehat{\beta}_k\|} + \frac{\beta_{k,n}}{\|\beta_{k,n}\|} \frac{\|\beta_{k,n}\|}{\|\widehat{\beta}_k\|} \rightarrow_p \omega_{k,0}. \quad (\text{B.59})$$

For the last group,

$$\widehat{\omega}_K = n^{1/2} \widehat{\beta}_K / \|n^{1/2} \widehat{\beta}_K\| \Rightarrow \frac{\tau_{\beta_K}(\pi_K^*)}{\|\tau_{\beta_K}(\pi_K^*)\|} \quad (\text{B.60})$$

by Theorem 1(a) and the CMT. Therefore,

$$\mathbf{B}(\widehat{\omega}) \Rightarrow \mathbf{B}_\omega(\pi_K^*), \quad (\text{B.61})$$

which is non-singular w.p.a.1.

The covariance matrix is $\widehat{\Sigma} = \widehat{\Sigma}(\widehat{\theta})$, where $\widehat{\Sigma}(\theta) = [\widehat{H}(\pi)]^{-1} \widehat{\Omega}_\theta(\theta) [\widehat{H}(\pi)]^{-1}$. Lemma A.1 implies that

$$\widehat{H}(\pi) \rightarrow_p H(\pi) \quad (\text{B.62})$$

uniformly over $\pi \in \Pi$. For the other term, we have

$$\begin{aligned}
\widehat{\Omega}_\theta(\theta) &= n^{-1} \sum_{t=1}^n \widehat{U}_t^2(\theta) d_{\theta,t}(\pi) d_{\theta,t}(\pi)' \\
&= n^{-1} \sum_{t=1}^n U_t^2 d_{\theta,t}(\pi) d_{\theta,t}(\pi)' \\
&\quad + 2n^{-1} \sum_{t=1}^n U_t \left(\sum_{k=1}^K (g(X_t, \pi_{k,n})' \beta_{k,n} - g(X_t, \pi_k)' \beta_k) \right) d_{\theta,t}(\pi) d_{\theta,t}(\pi)' \\
&\quad + n^{-1} \sum_{t=1}^n \left(\sum_{k=1}^K (g(X_t, \pi_{k,n})' \beta_{k,n} - g(X_t, \pi_k)' \beta_k) \right)^2 d_{\theta,t}(\pi) d_{\theta,t}(\pi)' \tag{B.63} \\
&\rightarrow_p \mathbb{E}_{\gamma_0} [U_t^2 d_{\theta,t}(\pi) d_{\theta,t}(\pi)'] + \mathbb{E}_{\gamma_0} \left[\left(\sum_{k=1}^K (g(X_t, \pi_{k,0})' \beta_{k,0} - g(X_t, \pi_k)' \beta_k) \right)^2 d_{\theta,t}(\pi) d_{\theta,t}(\pi)' \right]
\end{aligned}$$

uniformly over $\theta \in \Theta$ following the ULLN in Lemma A.1. Given the uniform consistency of $\widehat{\psi}_{K^-}(\pi_K)$ over π_K , we have

$$\widehat{\Omega}(\widehat{\psi}_{K^-}(\pi_K), \pi_K) \rightarrow_p \mathbb{E}_{\gamma_0} [U_t^2 d_{\theta,t}(\pi_{K^-,0}, \pi_K) d_{\theta,t}(\pi_{K^-,0}, \pi_K)'] = \Omega_\theta(\pi_{K^-,0}, \pi_K), \tag{B.64}$$

where the convergence holds uniformly over $\pi_K \in \Pi_K$. Putting together (B.62) and (B.64), we have

$$\begin{aligned}
\widehat{\Sigma}(\widehat{\psi}_{K^-}(\pi_K), \pi_K) &\rightarrow_p \Sigma(\pi_K) = [H(\pi_K)]^{-1'} \Omega_\theta(\pi_K) [H(\pi_K)]^{-1}, \text{ where} \\
H(\pi_K) &= H(\pi_{K^-,0}, \pi_K) \text{ and } \Omega_\theta(\pi_K) = \Omega_\theta(\pi_{K^-,0}, \pi_K). \tag{B.65}
\end{aligned}$$

By the CMT and Theorem 1(a),

$$\widehat{\Sigma} = \widehat{\Sigma}(\widehat{\psi}_{K^-}(\widehat{\pi}_K), \widehat{\pi}_K) \Rightarrow \Sigma(\pi_{K^*}). \tag{B.66}$$

By Assumptions 4 and 5, $\lambda_{\min}(\Sigma(\pi_{K^*}))$ is bounded away from 0.

Putting together (B.57), (B.61), (B.66), we obtain $\varepsilon_R = o_p(1)$ by (B.55). Furthermore, these results hold jointly. Therefore,

$$\begin{aligned}
W_n(R) &= \rho'_n V_n^{-1} \rho_n + o_p(1) \tag{B.67} \\
&= (R^* \mathbf{B}^{-1}(\widehat{\omega}) \xi_n)' \left[R^* \left(\mathbf{B}^{-1}(\widehat{\omega}) \widehat{\Sigma} \mathbf{B}^{-1}(\widehat{\omega}) \right) R^{*'} \right]^{-1} (R^* \mathbf{B}^{-1}(\widehat{\omega}) \xi_n) + o_p(1) \\
&\Rightarrow (R^* \mathbf{B}_\omega^{-1}(\pi_K^*) \xi(\pi_K^*))' \left[R^* \left(\mathbf{B}_\omega^{-1}(\pi_K^*) \widehat{\Sigma}(\pi_K^*) \mathbf{B}_\omega^{-1}(\pi_K^*) \right) R^{*'} \right]^{-1} (R^* \mathbf{B}_\omega^{-1}(\pi_K^*) \xi_n),
\end{aligned}$$

where the first equality follows from (B.55) and $\varepsilon_n = o_p(1)$, the second equality follows from the definition of ρ_n and V_n , and the convergence follows from the joint convergence of those in (B.57), (B.61), and (B.66).

Next, we prove part (b). Theorem 1(b) implies that

$$\xi_n(\widehat{\pi}_K) \rightarrow_d \xi \sim N(0, \Sigma(\pi_0)) \quad (\text{B.68})$$

because $\mathbf{B}^{-1}(\widehat{\beta}_K(\pi_K))\mathbf{B}(\beta_{K,n}) = 1_{d_K} + o_p(1)$ when group K involves semi-strong or strong identification. In addition, the angel parameters and the covariance matrix satisfy

$$\widehat{\omega} \rightarrow_p \omega_0 = (\omega'_{1,0}, \dots, \omega'_{K,0})' \text{ and } \widehat{\Sigma} \rightarrow_p \Sigma(\pi_0) \quad (\text{B.69})$$

following the arguments in (B.59) for $k = K$ and those in (B.65) with the consistency of $\widehat{\pi}_K$. Therefore, $\varepsilon_n = o_p(1)$ following the calculation in (B.55). Furthermore, the Wald statistic satisfies

$$\begin{aligned} W_n(R) &\rightarrow_d [R^*\mathbf{B}^{-1}(\omega_0)\xi]' [R^* (\mathbf{B}^{-1}(\omega_0)\Sigma_0\mathbf{B}^{-1}(\omega_0)) R^{*'}]^{-1} [R^*\mathbf{B}^{-1}(\omega_0)\xi] \\ &\sim \chi_{d_r}^2 \end{aligned} \quad (\text{B.70})$$

because R^* , $\mathbf{B}^{-1}(\omega_0)$, and Σ_0 all have full rank. This completes the proof. \square

Corollary 1 follows directly from Theorem 2.

C Proofs for the Asymptotic Size

Proof of Theorem 3. We prove this theorem using the generic results in Andrews, Cheng, and Guggenberger (2011) (hereafter ACG) after reparameterizing the model to fit the set-up in ACG. We invoke Corollary 2.1(b) of ACG, which requires the verification of Assumptions B1, B2*, and C1 in ACG. This gives the desired asymptotic size result for a confidence set. The asymptotic size of a test follows from the same arguments. The verification of these high-level assumptions in ACG reply on the reparameterization proposed below and the asymptotic distribution derived in Theorem 2.

To employ the notation in ACG, reparameterize β as $(\|\beta\|, g(\beta))$, where

$$g(\beta) = \left(\frac{\beta_j}{\beta_\ell} \right)_{j \neq \ell} = \left(\frac{\beta_1}{\beta_2}, \dots, \frac{\beta_1}{\beta_p}, \dots, \frac{\beta_p}{\beta_1}, \dots, \frac{\beta_p}{\beta_{p-1}} \right) \quad (\text{C.1})$$

and define $\beta_j/\beta_\ell = \infty$ if $\beta_\ell = 0$. Note that it is a one-to-one transformation between β and $(\|\beta\|, g(\beta))$ because

$$\beta_j = \frac{\|\beta\|}{\sqrt{\sum_{\ell=1}^p \left(\frac{\beta_\ell}{\beta_j}\right)^2}}. \quad (\text{C.2})$$

Hence, $\gamma = (\theta, \phi)$ can be reparameterized as

$$\lambda = (\|\beta\|, g(\beta), \zeta', \pi', \phi)'. \quad (\text{C.3})$$

Define a function

$$h_n(\lambda_n) = (n^{1/2}\beta_n, \|\beta_n\|, g(\beta_n), \zeta_n, \pi_n, \phi_n). \quad (\text{C.4})$$

Recall

$$h = (\mathcal{I}, b_{\mathcal{I}K}, \omega_0, \gamma_0) = (\mathcal{I}, b_{\mathcal{I}K}, \omega_0, \beta_0, \zeta_0, \pi_0, \phi_0). \quad (\text{C.5})$$

It is a one-to-one transformation between h and the limit of $h_n(\lambda_n)$.⁸ In particular, the grouping rule \mathcal{I} is determined by the limit of $h_n(\lambda_n)$ because (i) the strong identification group are for $\beta_{j,n} \rightarrow \beta_{j,0} \neq 0$, (ii) the weak identification groups are for $n^{1/2}\beta_{j,n} \rightarrow b_j \in R$, and (iii) the group structure for the semi-strong identification groups are determined by the relative convergence rates represented by the limit of $g(\beta_n)$. Given this grouping rule \mathcal{I} , the limit of $h_n(\lambda_n)$ determines $b_{\mathcal{I}K}$ and the group angel parameter ω_0 . The other direction from h to the limit of $h_n(\lambda_n)$ is obvious.

For any sequences of true parameters $\{\lambda_n : n \geq 1\}$ for which $h_n(\lambda_n)$ converges to a limit that can be reparameterized as $h \in H$, Theorem 2 shows that $W_n(R) \rightarrow_d \mathcal{W}(h)$. Under Assumption CV1, $\mathcal{W}(h)$ is continuous at $\chi_{dr,1-\alpha}^2 \forall h \in H$. Therefore, the coverage probability satisfies

$$\Pr(W_n(R) \leq \chi_{dr,1-\alpha}^2) \rightarrow \Pr(\mathcal{W}(h) \leq \chi_{dr,1-\alpha}^2). \quad (\text{C.6})$$

This verifies both Assumption B1 and C1 of ACG with the limit of the coverage probability $CP(h) = \Pr(\mathcal{W}(h) \leq \chi_{dr,1-\alpha}^2)$. (ACG allows for a lower and an upper bound for the limit of the finite-sample coverage probability, denoted by $CP^-(h)$ and $CP^+(h)$, respectively. In (C.6), $CP^-(h) = CP^+(h) = CP(h)$.)

Assumption B2*(i) of ACG holds with $(\lambda_1, \dots, \lambda_q)' = (\|\beta\|, g(\beta), \zeta', \pi')'$ and $\lambda_{q+1} = \phi$. Assumption B2*(ii) of ACG holds with $h_n(\lambda_n) = (n^{1/2}\beta_n, \lambda_n)$ and, as a result, Assumption B2*(iii) holds automatically. Assumption B2*(iv) of ACG holds because the parameter space is a product space and the parameter space of β_j includes a neighborhood around 0.

⁸We do not differentiate ∞ and $-\infty$ for this proof.

Invoking Corollary 2.1(b) of ACG, we obtain Theorem 3(b). Theorem 3(a) follows from the same arguments with H replaced by $H(v)$ for fixed v under the null and the coverage probability replaced by the rejection probability. Results for test and confidence set based on the t statistic follow from the same arguments. \square

Proof of Theorem 4. As in the proof of Theorem 3, we invoke Corollary 2.1(b) of ACG for this proof. The same reparameterization for λ and $h_n(\lambda_n)$ is necessary. Assumption B2* of ACG is the same for the standard test and the robust test, thus it remains to verify Assumptions B1 and C1 of ACG for the robust test and confidence interval based on the plug-in critical value.

We first introduce some notations. For a sequence of constants $\{c_n : n \geq 1\}$, let $c_n \rightarrow [c_1, c_2]$ denote $c_1 \leq \liminf_{n \rightarrow \infty} c_n \leq \limsup_{n \rightarrow \infty} c_n \leq c_2$. To verify Assumption B1 of ACG, we first show that for any sequence of true parameters $\{\lambda_n : n \geq 1\}$ for which $h_n(\lambda_n)$ converges to a limit that can be reparameterized as $h_0 \in H$, the coverage probability satisfies

$$\Pr(W_n(R) \leq \widehat{c}_{n,1-\alpha}) \rightarrow [CP^-(h_0), CP^+(h_0)] \quad (\text{C.7})$$

for some $CP^-(h_0), CP^+(h_0) \in [0, 1]$. Here we use $h_0 \in H$ rather than $h \in H$ to denote the sequence under consideration, whereas h is a generic notation in the definition of the plug-in critical value. To verify Assumption C1 of ACG, we show $CP^-(h_L) = CP^+(h_L)$ for some $h_L \in H$ such that $CP^-(h_L) = \inf_{h \in H} CP^-(h) = 1 - \alpha$. Then, Corollary 2.1(b) of ACG implies that the asymptotic size is $1 - \alpha$.

For a given $h_0 \in H$, its corresponding elements are $\mathcal{I}_{K,0}, \omega_{\mathcal{I}_k,0}, \pi_{\mathcal{I}_k,0}$. We define an infeasible critical value under h_0 as

$$\begin{aligned} \bar{c}_{1-\alpha}(h_0) &= \sup_{h \in H_0} \mathcal{W}_{1-\alpha}(h), \text{ where} \\ H_0 &= \{h \in H : \mathcal{I}_K = \mathcal{I}_{K,0}, \omega_{\mathcal{I}_k} = \omega_{\mathcal{I}_k,0}, \pi_{\mathcal{I}_k} = \pi_{\mathcal{I}_k,0} \text{ for } k < K\}. \end{aligned} \quad (\text{C.8})$$

This infeasible critical value $\bar{c}_{1-\alpha}(h_0)$ does not depend on the data. Because $h_0 \in H_0$,

$$\bar{c}_{1-\alpha}(h_0) \geq \mathcal{W}_{1-\alpha}(h_0). \quad (\text{C.9})$$

Recall the plug-in critical value defined as

$$\begin{aligned}\widehat{c}_{n,1-\alpha} &= \sup_{h \in \widehat{H}} \mathcal{W}_{1-\alpha}(h), \text{ where} \\ \widehat{H} &= \{h \in H : \mathcal{I}_K = \widehat{\mathcal{I}}_W, \omega_{\mathcal{I}_k} = \widehat{\beta}_{\mathcal{I}_k} / \|\widehat{\beta}_{\mathcal{I}_k}\| \text{ and } \pi_{\mathcal{I}_k} = \widehat{\pi}_{\mathcal{I}_k} \text{ for } k < K\}.\end{aligned}\quad (\text{C.10})$$

In the definition of \widehat{H} , \mathcal{I}_K , $\omega_{\mathcal{I}_k}$, $\pi_{\mathcal{I}_k}$ for $k < K$ are estimated. The grouping rule \mathcal{I} is not specified except for the last group \mathcal{I}_K .

Along a sequence of true parameters $\{\lambda_n : n \geq 1\}$ for which $h_n(\lambda_n)$ converges to a limit that can be reparameterized as $h_0 \in H$, we first show that the the estimated weak identified set $\widehat{\mathcal{I}}_W$ is no smaller than the true weak identification set $\mathcal{I}_{K,0}$ w.p.a.1, i.e., $\Pr(\mathcal{I}_{K,0} \subseteq \widehat{\mathcal{I}}_W) \rightarrow 1$. Therefore, imposing \mathcal{I}_K to be $\widehat{\mathcal{I}}_W$ in \widehat{H} is less restrictive than imposing \mathcal{I}_K to be $\mathcal{I}_{K,0}$ in H_0 . Here we assume there exist weakly identified regressors and they are collected in $\mathcal{I}_{K,0}$ following the grouping rule. When no regressors are weakly identified, the Wald statistic has a chi-square distribution and the limit of the coverage probability is greater than or equal to $1 - \alpha$ because $\widehat{c}_{n,1-\alpha} \geq \chi_{d_r,1-\alpha}^2$ by construction.

Consider $j \in \mathcal{I}_{K,0}$, Theorem 1 and (B.65) imply

$$ICS_{j,n} = n^{1/2}(\widehat{\Sigma}_j)^{-1/2} \left| \widehat{\beta}_j \right| \rightarrow_d (\Sigma_j(\pi_K^*))^{-1/2} \left| \tau_{\beta_j}(\pi_K^*) \right|, \quad (\text{C.11})$$

where $\tau_{\beta_j}(\pi_K)$ is an element of $\tau(\pi)$ associated with β_j and $\Sigma_j(\pi_K)$ is an element of $\Sigma(\pi)$ associated with β_j , for both of which π_1, \dots, π_{K-1} are evaluated at the limit of the true values. By Assumption 5, $\inf_{\pi_K \in \Pi_K} \Sigma_j(\pi_K) > 0$. Hence, $ICS_{j,n} = O_p(1)$ and $ICS_{j,n} < \kappa_n$ w.p.a.1. because $\kappa_n \rightarrow \infty$. This proves

$$\Pr(\mathcal{I}_{K,0} \subseteq \widehat{\mathcal{I}}_W) \rightarrow 1. \quad (\text{C.12})$$

It follows that any element that does not belong to $\widehat{\mathcal{I}}_W$ must be in the semi-strong or strong identification group. Therefore, $\widehat{\beta}_{\mathcal{I}_k} / \|\widehat{\beta}_{\mathcal{I}_k}\| \rightarrow_p \omega_{\mathcal{I}_k,0}$ and $\widehat{\pi}_k \rightarrow_p \pi_{\mathcal{I}_k,0}$ for $k < K$ for any group specification \mathcal{I} where $\mathcal{I}_K = \widehat{\mathcal{I}}_W$.

For a given group specification \mathcal{I} , the quantile $\mathcal{W}_{1-\alpha}(h)$ with $\omega_{\mathcal{I}_k} = \widehat{\beta}_{\mathcal{I}_k} / \|\widehat{\beta}_{\mathcal{I}_k}\|$ and $\pi_{\mathcal{I}_k} = \widehat{\pi}_{\mathcal{I}_k}$ converge in probability to the quantile of $\mathcal{W}_{1-\alpha}(h)$ with $\omega_{\mathcal{I}_k} = \omega_{\mathcal{I}_k,0}$, $\pi_{\mathcal{I}_k} = \pi_{\mathcal{I}_k,0}$ under Assumption CV2. This follows the same line of arguments for Theorem 3 of Andrews and Guggenberger (2009b). Because $\Pr(\mathcal{I}_{K,0} \subseteq \widehat{\mathcal{I}}_W) \rightarrow 1$, w.p.a.1,

$$\bar{c}_{1-\alpha}(h_0) \leq \widehat{c}_{n,1-\alpha} + o_p(1). \quad (\text{C.13})$$

Combing it together with (C.9), w.p.a.1, we have

$$\mathcal{W}_{1-\alpha}(h_0) \leq \widehat{c}_{n,1-\alpha} + o_p(1). \quad (\text{C.14})$$

Under the sequence of true parameters associated with $h_0 \in H$, Theorem 2 shows that $W_n(R) \rightarrow_d \mathcal{W}(h_0)$. Therefore,

$$\begin{aligned} & \Pr(W_n(R) \leq \widehat{c}_{n,1-\alpha}) \\ & \geq \Pr(W_n(R) + o_p(1) \leq \mathcal{W}_{1-\alpha}(h_0) \ \& \ \mathcal{W}_{1-\alpha}(h_0) \leq \widehat{c}_{n,1-\alpha} + o_p(1)) \\ & = \Pr(W_n(R) + o_p(1) \leq \mathcal{W}_{1-\alpha}(h_0)) \\ & \quad - \Pr(W_n(R) + o_p(1) \leq \mathcal{W}_{1-\alpha}(h_0) \ \& \ \mathcal{W}_{1-\alpha}(h_0) > \widehat{c}_{n,1-\alpha} + o_p(1)) \\ & \geq \Pr(W_n(R) + o_p(1) \leq \mathcal{W}_{1-\alpha}(h_0)) - \Pr(\mathcal{W}_{1-\alpha}(h_0) > \widehat{c}_{n,1-\alpha} + o_p(1)) \\ & \rightarrow 1 - \alpha, \end{aligned} \quad (\text{C.15})$$

where the convergence follows from $W_n(R) \rightarrow_d \mathcal{W}(h_0)$, the Slutsky's theorem, and (C.14). Therefore, for any $h_0 \in H$, (C.7) holds with $CP^-(h_0) = 1 - \alpha$. The value of $CP^+(h)$ does not matter for asymptotic size. We simply take $CP^+(h_0) = 1$.

To show $\inf_{h \in H} CP^-(h) = 1 - \alpha$, we consider the case where all parameters are strongly identified, e.g., $\beta_{j,n} \rightarrow \beta_{j,0} \neq 0$ for all $j = 1, \dots, p$. In this case,

$$\kappa_n^{-1} |ICS_{j,n}| = \left(\kappa_n^{-1} n^{1/2} \right) (\widehat{\Sigma}_j)^{-1/2} \left| \widehat{\beta}_j \right| \rightarrow \infty \quad (\text{C.16})$$

because κ_n diverges to ∞ slower than $n^{1/2}$. Therefore, when all parameters are strongly identified, $\widehat{\mathcal{I}}_W = \mathcal{O}$ w.p.a.1, which implies that $\widehat{c}_{n,1-\alpha} = \chi_{d_r,1-\alpha}^2$ w.p.a.1 in this case. In addition, Theorem 2 shows that $\mathcal{W}(h_0) \sim \chi_{d_r}^2$ in this case. Therefore, when all parameters are strongly identified,

$$\Pr(W_n(R) \leq \widehat{c}_{n,1-\alpha}) \rightarrow \Pr(\mathcal{W}(h_0) \leq \chi_{d_r,1-\alpha}^2) = 1 - \alpha. \quad (\text{C.17})$$

Let h_L denote the limit of $h_n(\lambda_n)$ when all parameters are strongly identified, i.e., $\beta_{0,j} \neq 0$ for all j in h_L . (C.17) shows $CP^-(h_L) = CP^+(h_L) = 1 - \alpha$. This completes the verification of Assumption C1 of ACG and concludes that the asymptotic size of the robust confidence set is $1 - \alpha$. The proof for the test is the same except that $H, H(v), \widehat{c}_{n,1-\alpha}$ are replaced by $H(v), \widehat{H}(v), \widehat{c}_{n,1-\alpha}(v)$, respectively, and the coverage probability is replaced by the rejection probability. The same arguments apply to robust tests and confidence sets based on the the t statistic. \square

References

- ANDREWS, D. W. K. (1994): “Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity,” *Econometrica*, 62(1), 43–72.
- ANDREWS, D. W. K., AND P. J. BARWICK (2012): “Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure,” *Econometrica*, 80(6), 2805–2826.
- ANDREWS, D. W. K., AND X. CHENG (2012): “Estimation and Inference With Weak, Semi-Strong, and Strong Identification,” *Econometrica*, 80(5), 2153–2211.
- (2013): “Maximum likelihood estimation and uniform inference with sporadic identification failure,” *Journal of Econometrics*, 173(1), 36–56.
- (2014): “GMM Estimation and Uniform Subvector Inference with Possible Identification Failure,” *Econometric Theory*, 30, 287–333.
- ANDREWS, D. W. K., X. CHENG, AND P. GUGGENBERGER (2011): “Generic Results for Establishing the Asymptotic Size of Confidence Sets and Tests,” Cowles Foundation Discussion Papers 1813, Cowles Foundation for Research in Economics, Yale University.
- ANDREWS, D. W. K., AND P. GUGGENBERGER (2009a): “Hybrid and Size-Corrected Subsampling Methods,” *Econometrica*, 77(3), 721–762.
- (2009b): “Validity Of Subsampling And Plug-In Asymptotic Inference For Parameters Defined By Moment Inequalities,” *Econometric Theory*, 25(03), 669–709.
- (2010): “Asymptotic size and a problem with subsampling and with the m out of n Bootstrap,” *Econometric Theory*, 26, 426–468.
- ANDREWS, D. W. K., M. J. MOREIRA, AND J. H. STOCK (2006): “Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression,” *Econometrica*, 74(3), 715–752.
- ANDREWS, D. W. K., AND W. PLOBERGER (1994): “Optimal Tests When a Nuisance Parameter Is Present Only under the Alternative,” *Econometrica*, 62(6), 1383–1414.
- ANDREWS, D. W. K., AND G. SOARES (2010): “Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection,” *Econometrica*, 78(1), 119–157.
- ANDREWS, D. W. K., AND J. H. STOCK (2007): “Testing with many weak instruments,” *Journal of Econometrics*, 138(1), 24–46.

- ANDREWS, I. (2013): “Conditional Linear Combination Tests for Weakly Identified Models,” Discussion paper, MIT.
- ANDREWS, I., AND A. MIKUSHEVA (2012): “A geometric approach to weakly identified econometric models,” Discussion paper, MIT.
- (2013): “Maximum Likelihood Inference in Weakly Identified DSGE Models,” *Quantitative Economics*, p. forthcoming.
- ANTOINE, B., AND E. RENAULT (2009): “Efficient GMM with nearly-weak instruments,” *Econometrics Journal*, 12(s1), S135–S171.
- (2012): “Efficient minimum distance estimation with multiple rates of convergence,” *Journal of Econometrics*, 170(2), 350–367.
- CANER, M. (2010): “Testing, Estimation in GMM and CUE with Nearly-Weak Identification,” *Econometric Reviews*, 29(3), 330–363.
- CHAUDHURI, S., AND E. ZIVOT (2011): “A new method of projection-based inference in GMM with weakly identified nuisance parameters,” *Journal of Econometrics*, 164(2), 239–251.
- CHEN, X., M. PONOMAREVA, AND E. TAMER (2013): “Likelihood Inference in Some Finite Mixture Models,” *Journal of Econometrics*, p. forthcoming.
- CHOI, I., AND P. C. B. PHILLIPS (1992): “Asymptotic and finite sample distribution theory for IV estimators and tests in partially identified structural equations,” *Journal of Econometrics*, 51(1-2), 113–150.
- DAVIES, R. B. (1977): “Hypothesis testing when a nuisance parameter is present only under the alternative,” *Biometrika*, 64(2), 247–254.
- DAVIES, R. B. (1987): “Hypothesis testing when a nuisance parameter is present only under the alternative,” *Biometrika*, 74(1), 33–43.
- DUFOUR, J.-M. (1997): “Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models,” *Econometrica*, 65(6), 1365–1388.
- DUFOUR, J.-M., AND M. TAAMOUTI (2005): “Projection-Based Statistical Inference in Linear Structural Models with Possibly Weak Instruments,” *Econometrica*, 73(4), 1351–1365.
- (2007): “Further results on projection-based inference in IV regressions with weak, collinear or missing instruments,” *Journal of Econometrics*, 139(1), 133–153.

- ELLIOTT, G., U. K. MÜLLER, AND M. W. WATSON (2012): “Nearly Optimal Tests when a Nuisance Parameter is Present Under the Null Hypothesis,” Discussion paper, UCSD and Princeton University.
- GUERRON-QUINTANA, P., A. INOUE, AND L. KILIAN (2013): “Frequentist inference in weakly identified dynamic stochastic general equilibrium models,” *Quantitative Economics*, 4(2), 197–229.
- GUGGENBERGER, P., F. KLEIBERGEN, S. MAVROEIDIS, AND L. CHEN (2012): “On the Asymptotic Sizes of Subset AndersonCRubin and Lagrange Multiplier Tests in Linear Instrumental Variables Regression,” *Econometrica*, 80(6), 2649–2666.
- GUGGENBERGER, P., AND R. J. SMITH (2005): “Generalized Empirical Likelihood Estimators And Tests Under Partial, Weak, And Strong Identification,” *Econometric Theory*, 21(04), 667–709.
- HAHN, J., AND G. KUERSTEINER (2002): “Discontinuities of weak instrument limiting distributions,” *Economics Letters*, 75(3), 325–331.
- HANSEN, B. E. (1996): “Inference When a Nuisance Parameter Is Not Identified under the Null Hypothesis,” *Econometrica*, 64(2), 413–30.
- KITAMURA, Y., AND P. C. B. PHILLIPS (1997): “Fully modified IV, GIVE and GMM estimation with possibly non-stationary regressors and instruments,” *Journal of Econometrics*, 80(1), 85–123.
- KLEIBERGEN, F. (2002): “Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression,” *Econometrica*, 70(5), 1781–1803.
- (2005): “Testing Parameters in GMM Without Assuming that They Are Identified,” *Econometrica*, 73(4), 1103–1123.
- LEE, L.-F. (2005): “Classical inference with ML and GMM estimates with various rates of convergence,” Discussion paper, Ohio State University.
- (2010): “Pooling Estimates With Different Rates Of Convergence: A Minimum 2 Approach With Emphasis On A Social Interactions Model,” *Econometric Theory*, 26(01), 260–299.
- MA, J., AND C. R. NELSON (2010): “Valid Inference for a Class of Models Where Standard Inference Performs Poorly: Including Nonlinear Regression, ARMA, GARCH, and Unobserved Components,” Economics Series 256, Institute for Advanced Studies.

- MCCLOSKEY, A. (2012): “Bonferroni-Based Size-Correction for Nonstandard Testing Problems,” Working Papers 2012-16, Brown University, Department of Economics.
- MOREIRA, M. J. (2003): “A Conditional Likelihood Ratio Test for Structural Models,” *Econometrica*, 71(4), 1027–1048.
- NELSON, C. R., AND R. STARTZ (1990): “Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator,” *Econometrica*, 58(4), 967–76.
- NELSON, C. R., AND R. STARTZ (2007): “The zero-information-limit condition and spurious inference in weakly identified models,” *Journal of Econometrics*, 138(1), 47–62.
- OLEA, J. L. M. (2013): “Efficient Conditionally Similar Tests: Finite-Sample Theory and Large-Sample Applications,” Discussion paper, New York University.
- PHILLIPS, P. (1989): “Partially Identified Econometric Models,” *Econometric Theory*, 5(02), 181–240.
- PHILLIPS, P. C. B., AND J. Y. PARK (1988): “On the Formulation of Wald Tests of Nonlinear Restrictions,” *Econometrica*, 56(5), 1065–83.
- QU, Z. (2013): “Inference in DSGE Models with Possible Weak Identification,” *Quantitative Economics*, p. forthcoming.
- RADCHENKO, P. (2008): “Mixed-Rates Asymptotics,” *The Annals of Statistics*, 36(1), pp. 287–309.
- SARGAN, J. D. (1983): “Identification and Lack of Identification,” *Econometrica*, 51(6), 1605–33.
- SCHORFHEIDE, F. (2013): *Advances in Economics and Econometrics: Theory and Applications, Tenth World Congress* vol. 3, chap. Estimation and Evaluation of DSGE Models: Progress and Challenges, pp. 184–230. Cambridge University Press.
- SHI, X., AND P. C. PHILLIPS (2012): “Nonlinear Cointegrating Regression Under Weak Identification,” *Econometric Theory*, 28(03), 509–547.
- SIMS, C. A., J. H. STOCK, AND M. W. WATSON (1990): “Inference in Linear Time Series Models with Some Unit Roots,” *Econometrica*, 58(1), 113–44.
- STAIGER, D., AND J. H. STOCK (1997): “Instrumental Variables Regression with Weak Instruments,” *Econometrica*, 65(3), 557–586.

STOCK, J. H., AND J. WRIGHT (2000): “GMM with Weak Identification,” *Econometrica*, 68(5), 1055–1096.

TRIPATHI, G. (1999): “A matrix extension of the Cauchy-Schwarz inequality,” *Economics Letters*, 63(1), 1–3.

VAN DER VAART, A., AND J. WELLNER (1996): *Weak Convergence and Empirical Processes*, Springer series in statistics. Springer.

VAN DIJK, D., T. TERASVIRTA, AND P. H. FRANSES (2002): “Smooth Transition Autoregressive Models A Survey Of Recent Developments,” *Econometric Reviews*, 21(1), 1–47.