Journal of Econometrics 189 (2015) 207–228

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Robust inference in nonlinear models with mixed identification strength[☆]

Xu Cheng[*]

Department of Economics, University of Pennsylvania, United States

ARTICLE INFO

Article history:
Received 10 May 2014
Received in revised form
1 July 2015
Accepted 13 July 2015
Available online 3 August 2015

JEL classification:
C12
C15

Keywords:
Mixed rates
Nonlinear regression
Robust inference
Uniformity
Weak identification

ABSTRACT

The paper studies inference in regression models composed of nonlinear functions with unknown transformation parameters and loading coefficients that measure the importance of each component. In these models, non-identification and weak identification present in multiple parts of the parameter space, resulting in mixed identification strength for different unknown parameters. This paper proposes robust tests and confidence intervals for sub-vectors and linear functions of the unknown parameters. In particular, the results cover 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. To construct this robust inference procedure, we develop a local limit theory that models mixed 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.

1. Introduction

Economic theory and empirical studies often suggest nonlinear relationships among economic variables. These relationships are commonly specified in a parametric form involving nonlinear component functions with unknown transformation parameters and loading coefficients that measure the importance of each nonlinear component. Generalizing the linear regression model, these nonlinear regression models take the form

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

where $\pi_j \in \mathbb{R}^{d_{\pi_j}}$ is the unknown coefficient in the smooth nonlinear function $g_j(\cdot, \pi_j)$, and $\beta_j \in \mathbb{R}^{d_{\beta_j}}$ and $\zeta \in \mathbb{R}^{d_\zeta}$ are coefficients of the nonlinear and linear regressors, respectively. In this model, $\|\beta_j\|$ determines the identification strength of π_j. If $d_{\beta_j} > 1$, each element of $g_j(X_t, \pi_j)$ depends on the whole vector of π_j such that the identification of π_j is lost only if $\beta_j = 0$. For $j = 1, \ldots, p$, $\beta_j = 0$ yields p different sources of identification failures. In finite-sample estimation, small $\|\beta_j\|$ results in the weak identification of π_j. Inference is non-standard because non/weak identification occurs in multiple areas of the parameter space and the unknown parameters may have mixed identification strength.

Several classes of nonlinear functions are popular in empirical applications. One is the smooth transition autoregressive model (STAR, see Granger and Terasvirta (1993) and Terasvirta (1994)), where $g_j(x, \pi_j) = \phi(x, \pi_j)x$ and $\phi(x, \pi_j)$ is the logistic function or exponential function with unknown location parameter π_j. Each nonlinear function links two regimes. Multiple regime STAR model and its applications to business cycles and real exchange rate dynamics are studied by van Dijk and Franses (1999), McAleer and Medeiros (2008), Bec et al. (2010) and Shintani et al. (2013), among others. Another popular nonlinear function is the Box–Cox transformation (Box and Cox (1964)), where $g_j(x, \pi_j) = (x^{\pi_j} - 1)/\pi_j$. Its application to the estimation of production function and cost function are considered by Caves et al. (1980), Clark (1984), and Giannakas et al. (2000), etc. In the neural network (see White (1989) and Kuan and White (1994)), $g_j(x, \pi_j) = \phi(\pi_j'x)$, where $\phi(\cdot)$ is the logistic function. Additional nonlinear transformations are discussed in Hansen (1996).

[☆] This paper supersedes a working paper previously circulated under the title "Robust Confidence Intervals in Nonlinear Regression under Weak Identification". The author thanks the co-editor Jianqing Fan, an associate editor, three referees, Donald Andrews, Patrik Guggenberger, Adam McCloskey, and Eric Renault for helpful comments.

[*] Correspondence to: 3718 Locust Walk, Philadelphia, PA, 19104, United States. *E-mail address:* xucheng@econ.upenn.edu.

http://dx.doi.org/10.1016/j.jeconom.2015.07.003
0304-4076/© 2015 Elsevier B.V. All rights reserved.

Mixed identification strength brings new challenges to hypothesis testing and the construction of confidence sets. 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 $j = 1, \ldots, p-1$ 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), Luukkonen et al. (1988), Andrews and Ploberger (1994), and Hansen (1996), where some nuisance parameters are not identified under the null. These classical results apply to testing the null hypothesis $H_0 : \beta = (\beta_1', \ldots, \beta_p')' = 0$, where the nuisance parameter $\pi = (\pi_1', \ldots, \pi_p')'$ 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 or linear functions of $\theta = (\beta', \zeta', \pi')'$ that is robust to weak identification. There is a large literature on inference robust to weak identification following Staiger and Stock (1997) and Stock and Wright (2000). While many important results are developed for the full vector of θ, sub-vector inference typically depends on projection or concentration out of strongly-identified nuisance parameters. In the nonlinear regression model considered in this paper, the direction of weak identification is known. Making use of this structure, we propose robust and non-conservative tests and confidence sets for sub-vectors of θ, allowing the nuisance parameters to be strongly identified or weakly identified.

The paper derives a local limit theory for the least squares estimator and the Wald statistic when β_j for $j = 1, \ldots, p$ converges to 0 at various rates or is bounded away from 0. Because the identification strength is unknown, all convergence rates and all combinations across $j = 1, \ldots, p$ are considered for uniform inference. 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, 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 $H_0 : \beta_p = 0$. When the number of such crucial points increases from one to multiple, this new asymptotic theory is required for uniform inference with mixed identification strength.

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 et al. (1990) and 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.

This paper contributes to the growing literature on 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 et al. (2012), Andrews and Mikusheva (2012, 2015) and Chen et al. (2014), among others. Kleibergen (2014) considers efficient subset inference in linear instrumental variable models. In a general nonlinear model, the geometric approach in Andrews and Mikusheva (2015) provides an informative robust test.

Mixed identification strength also is considered by Andrews and Guggenberger (2014a,b) in moment condition models. They show that it is important to consider cases where the singular values of the Jacobian drift to zero at different rates in order to establish the uniform validity of an identification-robust test. Andrews and Guggenberger (2014a,b) investigate the uniform validity of some existing tests and proceed to propose three new tests that are robust to both weak identification of a general form and singular variance matrix of the moments. These papers focus on full vector inference in a general moment condition model, whereas the present paper studies sub-vector inference in a nonlinear regression model. Thus, different types of robust tests are used.

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 et al. (2006), Montiel Olea (2013) and Andrews (2013), and other papers referenced in Andrews and Stock (2007). Guerron-Quintana et al. (2013), Andrews and Mikusheva (2012, 2015) and Qu (2014) consider weak identification in DSGE models, an important issue discussed in Schorfheide (2013) and 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

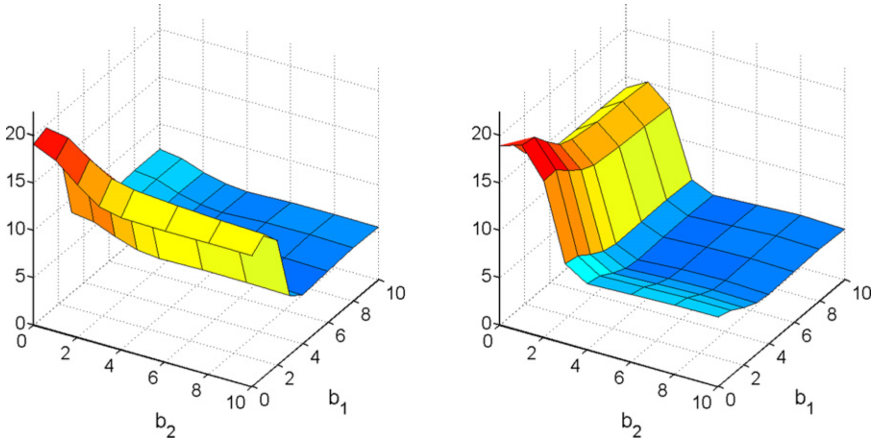

Fig. 1. Standard Two-Sided t Test: Finite-Sample Rejection Probability ($\times 100$) for $H_0 : \beta_1 = \beta_{1,0}$ (left) and $H_0 : \beta_2 = \beta_{2,0}$ (right).

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. For a robust test, the critical value $c_{n,1-\alpha}(v)$ may depend on both the sample size and the null value. 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, Fig. 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 $\beta_1 \in \mathbb{R}$ and $\beta_2 \in \mathbb{R}$. 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 finite-sample distribution also depends on the identification strength of π_1, measured by $\|\beta_1\|$. In consequence, a non-standard distribution that works well when β_1 is far from 0 may work poorly when β_1 is close to 0. Fig. 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 $j = 1, \ldots, p$. Due to the nonlinear structure of the model, π_j is strongly identified only if $\beta_{j,n} \to \beta_{j,0} \neq 0$. 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 consider sequences of $\beta_{j,n}$ that satisfy one of the following conditions:

$$\text{(i) } \beta_{j,n} \to 0, \qquad n^{1/2}\beta_{j,n} \to b_j \in \mathbb{R}^{d_{\beta_j}},$$

(weak identification) or

$$\text{(ii) } \beta_{j,n} \to 0, \qquad n^{1/2}\|\beta_{j,n}\| \to \infty,$$

(semi-strong identification) or

$$\text{(iii) } \beta_{j,n} \to \beta_{j,0} \neq 0 \quad \text{(strong identification)}. \tag{2.1}$$

For $j = 1, \ldots, p$, (i), (ii), or (iii) could be the case. In addition, $\lim_{n\to\infty} \|\beta_{j,n}\| / \|\beta_{j',n}\| \in R \cup \{\pm\infty\}$ for sequences in (ii) and (iii).[1] 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. 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 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 $j = 1, \ldots, p$ is a closed set in $\mathbb{R}^{d_{\beta_j}}$ that includes both zero and non-zero values. Thus, the area where non/weak identification occurs is part of the parameter space. The parameter space Π^* is compact. 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. The space Φ^* is a compact metric space with a metric that induces weak convergence of bivariate distribution (W_i, W_{i+m}) for all $i, m \geq 1$.[2] In parametric models, the finite-dimensional parameter θ fully specifies the distribution of the data and ϕ does not exist. Let $\mathbb{P}_\gamma$ and $\mathbb{E}_\gamma$ denote the probability and expectation under the distribution indexed by γ.

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

[2] For example, the Prokhorov metric on probability measures induces weak convergence. The compactness assumption is not restrictive following the Prokhorov's Theorem (Theorem 6.1 of Billingsley (1968)). If a set of probability measures is tight, its closure is sequentially compact, which gives a convergent subsequence and is equivalent to compact on a metric space.

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

$$\begin{aligned} &\theta_n = (\beta_n', \zeta_n', \pi_n')' \in \mathbb{R}^{d_\theta}, \qquad \beta_n = (\beta_{1,n}', \ldots, \beta_{p,n}')' \in \mathbb{R}^{d_\beta}, \\ &\pi_n = (\pi_{1,n}', \ldots, \pi_{p,n}')' \in \mathbb{R}^{d_\pi}, \quad \text{and} \quad \gamma_n = (\theta_n, \phi_n) \end{aligned} \tag{3.1}$$

where $\theta_n \to \theta_0 = (\beta_0', \zeta_0', \pi_0')'$, $\gamma_n \to \gamma_0 \in \Gamma$, and the subscript 0 denotes the limit of true values.[3] 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 function[4] is

$$Q_n(\theta) = \frac{1}{2n} \sum_{t=1}^{n} \left(Y_t - \sum_{j=1}^{p} g_j(X_t, \pi_j)'\beta_j - Z_t'\zeta \right)^2. \tag{3.2}$$

The least squares estimator $\widehat{\theta}_n$ 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 $j = 1, \ldots, p$ 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 derive asymptotic distributions along sequences of true parameters $\{\gamma_n \in \Gamma : n \geq 1\}$, assuming that the following assumptions hold for any $\gamma \in \Gamma$.

Let $g_{j\ell}(x, \pi_j) \in \mathbb{R}$ denote the ℓ-th element of $g_j(x, \pi_j) \in \mathbb{R}^{d_{\beta_j}}$. Assumptions 1, 2, and 2* holds for all j and ℓ.

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

For time series data, the following assumption holds. Let d_θ denote the dimensional of θ. Let C denote a generic finite constant.

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) $\mathbb{E}_\gamma(\sup_{\pi_j \in \Pi_j}[g_{j\ell}\left(X_t, \pi_j\right)^{2q} + \|g_{j\ell}^{\pi}\left(X_t, \pi_j\right)\|^{2q} + \|g_{j\ell}^{\pi\pi}\left(X_t, \pi_j\right)\|^{2q}] + M_{j\ell}(X_t)^{2q}) \leq C$.

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) $\mathbb{E}_\gamma(\sup_{\pi_j \in \Pi_j}[g_{j\ell}(X_t, \pi_j)^{4+\delta} + \|g_{j\ell}^{\pi}(X_t, \pi_j)\|^{4+\delta} + \|g_{j\ell}^{\pi\pi}(X_t, \pi_j)\|^{4+\delta}] + M_{j\ell}(X_t)^{4+\delta}) \leq C$.

Let $g(X_t, \pi) = (g_1(X_t, \pi_1)', \ldots, g_p(X_t, \pi_p)')'$ denote the collection of all nonlinear regressors.

Assumption 3. $\forall \pi, \pi_0 \in \Pi$ and some $\varepsilon > 0$, $\mathbb{P}_\gamma([g(X_t, \pi)', g(X_t, \pi_0)', Z_t']a = 0) \leq 1 - \varepsilon$ 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. Assumption 3 requires no multi-collinearity between $g(X_t, \pi)$, $g(X_t, \pi_0)$, and Z_t for any $\pi \neq \pi_0$, which rules out the case where $g(X_t, \pi)$ is a linear in π. These are standard assumptions in nonlinear regression analysis.

3.1. Grouping rules and reparameterization

To derive asymptotic results with mixed identification strength, we first group $g_1(X_t, \pi_1), \ldots, g_p(X_t, \pi_p)$ based on the order of magnitude of $\|\beta_{1,n}\|, \ldots, \|\beta_{p,n}\|$. The grouping rule is specified based on $\|\beta_{j,n}\|$, but the grouping result applies to the jth regressor and it categorizes the identification strength of π_j. 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 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.

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

$$\mathcal{I}_k = \{k_1, \ldots, k_{p_k}\} \tag{3.3}$$

denote the indices for group k. For example, suppose $p = 7$, $\beta_{1,n} = 3$, $\beta_{2,n} = 1$, $\beta_{3,n} = n^{-1/4}$, $\beta_{4,n} = n^{-1/3}$, $\beta_{5,n} = 2n^{-1/3}$, $\beta_{6,n} = n^{-1/2}$, and $\beta_{7,n} = n^{-1}$. 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 groups is $K = 4$. In this simple example, $\beta_{k_1}, \ldots, \beta_{k_{p_k}}$ are all scalars, but the general results allow them to be vectors.

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.,

$$\begin{aligned} &\beta_{\mathcal{I}_k} = (\beta_{k_1}', \ldots, \beta_{k_{p_k}}')' \in \mathbb{R}^{d_k} \\ \text{and} \quad &\pi_{\mathcal{I}_k} = (\pi_{k_1}', \ldots, \pi_{k_{p_k}}')' \in \mathbb{R}^{d_{\pi \mathcal{I}_k}}. \end{aligned} \tag{3.4}$$

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

$$\begin{aligned} &\text{between groups}: \quad \|\beta_{\mathcal{I}_{k'},n}\| = o(\|\beta_{\mathcal{I}_k,n}\|) \quad \text{for } k' > k, \\ &\text{within group}: \quad \|\beta_{j,n}\| \asymp \|\beta_{\mathcal{I}_k,n}\| \quad \text{for } j \in \mathcal{I}_k \\ &\qquad\qquad\qquad\quad \text{and } k = 1, \ldots, K - 1, \end{aligned} \tag{3.5}$$

[3] The metric d_{Φ^*} on Φ^* must satisfy: if $\gamma \to \gamma_0$, then (W_i, W_{i+m}) under γ converges in distribution to (W_i, W_{i+m}) under γ_0. Note that Γ is a metric space with metric $d_\Gamma(\gamma_1, \gamma_2) = \|\theta_1 - \theta_2\| + d_{\Phi^*}(\phi_1, \phi_2)$, where $\gamma_j = (\theta_j, \phi_j) \in \Gamma$ for $j = 1, 2$. The same metric is used in Andrews and Cheng (2012).

[4] The constant 1/2 is added to simplify the asymptotic results presented below.

where $\asymp$ represents convergence at the same rate.[5] 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 $Y_t = \beta_1 g(X_t, \pi_1) + \beta_2 g(X_t, \pi_2) + U_t$ and $\beta_1, \beta_2 \in \mathbb{R}$.

(i) If $\beta_{1,n} \to \beta_{1,0} \neq 0$ and $\beta_{2,n} \to \beta_{2,0} \neq 0$, $\mathcal{I}_1 = \{1, 2\}$.

(ii) If $n^{1/2}\beta_{1,n} \to b_1 \in R$, $n^{1/2}\beta_{2,n} \to b_2 \in R$, $\mathcal{I}_1 = \emptyset$, $\mathcal{I}_2 = \{1, 2\}$. Here $\mathcal{I}_1 = \oslash$ because both $\beta_{1,n}$ and $\beta_{2,n}$ have zero limits.

(iii) If $\beta_{1,n} \to 0$, $|n^{1/2}\beta_{1n}| \to \infty$, $\beta_{2,n} \asymp \beta_{1,n}$, $\mathcal{I}_1 = \oslash$ and $\mathcal{I}_2 = \{1, 2\}$.

(iv) If $\beta_{1,n} \to \beta_{1,0} \neq 0$ and $\beta_{2,n} \to 0$, $\mathcal{I}_1 = \{1\}$, $\mathcal{I}_2 = \{2\}$.

(v) If $\beta_{1,n} \to 0$, $\beta_{2,n} \to 0$, $|n^{1/2}\beta_{1n}| \to \infty$, $\beta_{2n}/\beta_{1n} \to 0$, $\mathcal{I}_1 = \oslash$, $\mathcal{I}_2 = \{1\}$, $\mathcal{I}_3 = \{2\}$.

In cases (i)–(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. □

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. With the grouping notations, the model can be equivalently written as

$$Y_t = \sum_{k=1}^{K} g_{\mathcal{I}_k}(X_t, \pi_{\mathcal{I}_k})'\beta_{\mathcal{I}_k} + Z_t'\zeta + U_t. \tag{3.6}$$

Define the first and second order derivatives as

$$g_{\pi_k}(X_t, \pi_{\mathcal{I}_k}) = \frac{\partial}{\partial \pi'_{\mathcal{I}_k}} g_{\mathcal{I}_k}(X_t, \pi_{\mathcal{I}_k}) \in \mathbb{R}^{d_k \times d_{\pi_{\mathcal{I}_k}}} \quad \text{and}$$
$$g_{\pi\pi_k}(X_t, \pi_{\mathcal{I}_k}) = \frac{\partial}{\partial \pi'_{\mathcal{I}_k}} vec(g_{\pi_k}(X_t, \pi_{\mathcal{I}_k})') \in \mathbb{R}^{(d_k d_{\pi_{\mathcal{I}_k}}) \times d_{\pi_{\mathcal{I}_k}}}. \tag{3.7}$$

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.,

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

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.

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 following Assumption 3. 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, $\mathcal{I}_1 = \oslash$ 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 $\widehat{\beta}(\pi_{1^+})$, $\widehat{\zeta}(\pi_{1^+})$, and $\widehat{\pi}_{\mathcal{I}_1}(\pi_{1^+})$. The concentrated criterion function $Q_n(\widehat{\beta}(\pi_{1^+}), \widehat{\zeta}(\pi_{1^+}), \widehat{\pi}_{\mathcal{I}_1}(\pi_{1^+}), \pi_{1^+})$ 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 $k = 2, \ldots, K - 1$ 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 $\widehat{\pi}_{\mathcal{I}_k}(\pi_{k^+})$. Concentrating out $\pi_{\mathcal{I}_k}$, the criterion function $Q_n^c(\widehat{\pi}_{\mathcal{I}_k}(\pi_{k^+}), \pi_{k^+})$ 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 $\widehat{\pi}_{\mathcal{I}_K}$.

(iv) Reverse the order of the procedure. Sequentially plug in the estimators from $\widehat{\pi}_{\mathcal{I}_K}$ to $\widehat{\pi}_{\mathcal{I}_2}$, we obtain $\widehat{\pi}_{\mathcal{I}_{K-1}} = \widehat{\pi}_{\mathcal{I}_{K-1}}\left(\widehat{\pi}_{\mathcal{I}_K}\right)$, $\ldots$, $\widehat{\pi}_{\mathcal{I}_1} = \widehat{\pi}_{\mathcal{I}_1}(\widehat{\pi}_{\mathcal{I}_2}, \ldots, \widehat{\pi}_{\mathcal{I}_K})$, $\widehat{\beta} = \widehat{\beta}(\widehat{\pi}_{\mathcal{I}_2}, \ldots, \widehat{\pi}_{\mathcal{I}_K})$, and $\widehat{\zeta} = \widehat{\zeta}(\widehat{\pi}_{\mathcal{I}_2}, \ldots, \widehat{\pi}_{\mathcal{I}_K})$.

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

$$\widehat{\theta} = (\widehat{\beta}', \widehat{\zeta}', \widehat{\pi}'_{\mathcal{I}_1}, \ldots, \widehat{\pi}'_{\mathcal{I}_K})'. \tag{3.9}$$

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 $\beta_{\mathcal{I}_k,n} \to 0$ for $k > 1$. By an extension of standard arguments for the consistency of extremum estimators, we obtain the uniform consistency for the strongly identified parameters.

Lemma 1 (*Consistency for Strong Identification Groups*). *Suppose Assumptions 1–3 hold. Then, under $\gamma_n \to \gamma_0$,*

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

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 $k = 2, \ldots, K - 1$. We show that, after proper recentering 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

$$\beta_{k^-} = (\beta'_{\mathcal{I}_1}, \ldots, \beta'_{\mathcal{I}_{k-1}})' \quad \text{and} \quad \beta_{k^+} = (\beta'_{\mathcal{I}_{k+1}}, \ldots, \beta'_{\mathcal{I}_K})', \tag{3.10}$$

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

$$\psi_{k^-} = (\beta', \zeta', \pi'_{k^-})' \in \mathbb{R}^{k^-}. \tag{3.11}$$

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

Define

$$\psi^0_{k^-,n} = (\beta'_{k^-,n}, \beta^{0\prime}_{\mathcal{I}_k}, \beta^{0\prime}_{k^+}, \zeta'_n, \pi'_{k^-,n}),$$
$$\text{with } \beta^0_{\mathcal{I}_k} = 0 \text{ and } \beta^0_{k^+} = 0. \tag{3.12}$$

Note that the difference between $\psi^0_{k^-,n}$ and $\psi_{k^-,n}$, the true value of ψ_{k^-}, lies in $\beta_{\mathcal{I}_k}$ and β_{k^+}. To derive the asymptotic distribution of the concentrated criterion function, $Q_n(\widehat{\psi}_{k^-}(\pi_{\mathcal{I}_k}, \pi_{k^+}), \pi_{\mathcal{I}_k}, \pi_{k^+})$ is centered around $Q_n(\psi^0_{k^-,n}, \pi_{\mathcal{I}_k}, \pi_{k^+})$. We set $\beta^0_{\mathcal{I}_k} = 0$ and $\beta^0_{k^+} = 0$ in $\psi^0_{k^-,n}$ so that the centering term $Q_n(\psi^0_{k^-,n}, \pi_{\mathcal{I}_k}, \pi_{k^+})$ does not depend on $(\pi_{\mathcal{I}_k}, \pi_{k^+})$. To make it clear, $Q_n(\psi^0_{k^-,n}, \pi_{\mathcal{I}_k}, \pi_{k^+})$ is abbreviated to $Q_n(\psi^0_{k-1,n})$.

[5] For two sequences of non-zero constants $\{a_n : n \geq 1\}$ and $\{b_n : n \geq 1\}$, we say a_n and b_n converge to 0 at the same rate and write $a_n \asymp b_n$ if and only if $\liminf_{n\to\infty} \frac{a_n}{b_n} \neq 0$ and $\liminf_{n\to\infty} \frac{b_n}{a_n} \neq 0$.

To study the local expansion of the sample criterion function around $\psi^0_{k^-,n}$, define a vector associated with the first order derivative with respect to ψ_{k^-}:

$$\begin{aligned} &d_{\psi_k,t}(\pi,\omega_{k^-}) \\ &\quad = (g(X_t,\pi)', Z_t', \omega_1' g_{\pi_1}(X_t,\pi_{\mathcal{I}_1}),\ldots,\omega_{k-1}' g_{\pi_{k-1}}(X_t,\pi_{\mathcal{I}_{k-1}}))', \end{aligned} \tag{3.13}$$

where

$$\omega_k = \beta_{\mathcal{I}_k}/\|\beta_{\mathcal{I}_k}\| \quad \text{and} \quad \omega_{k^-} = (\omega_1',\ldots,\omega_{k-1}')' \tag{3.14}$$

are the angle parameters for each group. The angle parameters $\omega_1,\ldots,\omega_{k-1}$ show up in (3.13) because the norm $\|\beta_{\mathcal{I}_1}\|,\ldots,\|\beta_{\mathcal{I}_{k-1}}\|$ are taken out for renormalization in the results developed below.

For any $\pi_{\mathcal{I}_k},\widetilde{\pi}_{\mathcal{I}_k}\in\Pi_{\mathcal{I}_k}$, define a covariance matrix

$$H_k(\pi_{\mathcal{I}_k},\widetilde{\pi}_{\mathcal{I}_k}|\pi_{k^+}) = \mathbb{E}_{\gamma_0} d_{\psi_k,t}(\pi_{\mathcal{I}_k},\pi_{k^+}) d_{\psi_k,t}(\widetilde{\pi}_{\mathcal{I}_k},\pi_{k^+})' \tag{3.15}$$

where $d_{\psi_k,t}(\pi_{\mathcal{I}_k},\pi_{k^+})$ abbreviates $d_{\psi_k,t}(\pi,\omega_{k^-})$ when $\pi_{k^-}=\pi_{k^-,0}$ and $\omega_{k^-}=\omega_{k^-,0}$ take the limits of the true values as $n\to\infty$. Assumption 4 is similar to Assumption C4 in Andrews and Cheng (2012).

Assumption 4. $\lambda_{\min}(H_k(\pi_{\mathcal{I}_k},\pi_{\mathcal{I}_k}|\pi_{k^+}))\geq\varepsilon$ for some $\varepsilon>0$ for any $\pi_{\mathcal{I}_k}\in\Pi_{\mathcal{I}_k}$, $\pi_{k^+}\in\Pi_{k^+}$, $\gamma_0\in\Gamma$ for $k=1,\ldots,K$.

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, under $\gamma_n\to\gamma_0$, for $k=2,\ldots,K-1$,*

(a) *the concentrated sample criterion function satisfies*

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

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

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

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

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

$$\|\beta_{\mathcal{I}_k,n}\|^{-1}\begin{pmatrix}\widehat{\beta}_{k^-}(\pi_{k^+})-\beta_{k^-,n}\\ \widehat{\beta}_{\mathcal{I}_k}(\pi_{k^+})-\beta_{\mathcal{I}_k,n}\\ \widehat{\beta}_{k^+}(\pi_{k^+})\\ \widehat{\zeta}-\zeta_n\\ \mathbf{B}^*\left(\beta_{k^-,n}\right)\left(\widehat{\pi}_{k^-}(\pi_{k^+})-\pi_{k^-,n}\right)\end{pmatrix}\to_p 0,$$

where

$$\mathbf{B}^*(\beta_{k^-})=diag\{(1_{d_{\pi_{\mathcal{I}_1}}}\|\beta_{\mathcal{I}_1}\|,\ldots,1_{d_{\pi_{\mathcal{I}_{k-1}}}}\|\beta_{\mathcal{I}_{k-1}}\|)'\}. \tag{3.16}$$

Comments. 1. Part (a) is obtained by a quadratic expansion of $Q_n(\widehat{\psi}_{k^-}(\pi_{\mathcal{I}_k},\pi_{k^+}),\pi_{\mathcal{I}_k},\pi_{k^+})$ around the centering term $Q_n(\psi^0_{k^-,n})$. This expansion relies on the consistency of $\widehat{\psi}_{k^-}(\pi_{\mathcal{I}_k},\pi_{k^+})$, 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 $\psi^0_{k^-,n}$ instead of the true value of ψ_{k^-}. The choice of $\psi^0_{k^-,n}$ 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). For models with only one point of identification failure, Assumption C1 of Andrews and Cheng (2012) suggest centering the criterion function at $\beta=0$. The specification of $\psi^0_{k^-,n}$ generalizes this one-group strategy to cases where we have to consider $\beta_{k^-,n}$, $\beta^0_{\mathcal{I}_k}$, $\beta^0_{k^+}$ for each k, with β_{k^-} at the true value and $\beta_{\mathcal{I}_k}$ and β_{k^+} both at 0. The rate of convergence for the criterion function is based on the group specific identification strength. A similar rate is derived in Lemma 3.2 of Andrews and Cheng (2012) when there is only one group. Second, in this quadratic expansion, both the first and second order derivatives have mixed rate of convergence. This is different from the one-group case in Andrews and Cheng (2012).

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.,

$$\sup_{\pi_K\in\Pi_K}\left\|\widehat{\pi}_{K^-}(\pi_K)-\pi_{K^-,n}\right\|\to_p 0. \tag{3.17}$$

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., $n^{1/2}\beta_{\mathcal{I}_K}\to b_{\mathcal{I}_K}\in R^{d_K}$. (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

$$\Omega(\pi_{\mathcal{I}_K},\widetilde{\pi}_{\mathcal{I}_K})=\mathbb{E}_{\gamma_0}U_t^2 d_{\psi_K,t}(\pi_{\mathcal{I}_K})d_{\psi_K,t}(\widetilde{\pi}_{\mathcal{I}_K})', \tag{3.18}$$

where $d_{\psi_K,t}(\pi_{\mathcal{I}_K})$ abbreviates $d_{\psi_K,t}(\pi_{\mathcal{I}_{K^-},0},\pi_{\mathcal{I}_K},\omega_{K^-,0})$ defined in (3.13) for $k=K$. Building on this Gaussian process, define

$$\begin{aligned} \tau(\pi_{\mathcal{I}_K}) &= \left[H_K(\pi_{\mathcal{I}_K},\pi_{\mathcal{I}_K})\right]^{-1}\left[H_K(\pi_{\mathcal{I}_K},\pi_{\mathcal{I}_K,0})S_{\mathcal{I}_K}b_{\mathcal{I}_K}+G(\pi_{\mathcal{I}_K})\right], \\ \chi(\pi_{\mathcal{I}_K}) &= -\frac{1}{2}\tau(\pi_{\mathcal{I}_K})'H_K(\pi_{\mathcal{I}_K},\pi_{\mathcal{I}_K})\tau(\pi_{\mathcal{I}_K}), \\ \pi^*_{\mathcal{I}_K} &= \arg\min_{\pi_K\in\Pi_K}\chi(\pi_{\mathcal{I}_K}), \end{aligned} \tag{3.19}$$

where $S_{\mathcal{I}_K}=[0_{d_K\times d_{K^-}},I_{d_K},0_{d_K\times(d_\zeta+d_{K^-})}]'$ selects $\beta_{\mathcal{I}_K}$ out of ψ_{K^-}. We assume that each sample path of the non-central chi-square process $\chi(\pi_{\mathcal{I}_K})$ has a unique minimizer with probability one and call this minimizer $\pi^*_{\mathcal{I}_K}$. In the presence of weak identification, Theorem 1 shows that $\chi(\pi_{\mathcal{I}_K})$ 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, $\chi(\pi_{\mathcal{I}_K})$ 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 of $\widehat{\theta}$ in case (b), define covariance matrices

$$\Sigma(\pi,\omega)=H^{-1}(\pi,\omega)\Omega_\theta(\pi,\omega)H^{-1}(\pi,\omega), \tag{3.20}$$

where

$$\begin{aligned}
&H(\pi,\omega)=\mathbb{E}_{\gamma_0}d_{\theta,t}(\pi,\omega)d_{\theta,t}(\pi,\omega)',\\
&\Omega_\theta(\pi,\omega)=\mathbb{E}_{\gamma_0}U_t^2d_{\theta,t}(\pi,\omega)d_{\theta,t}(\pi,\omega)' \quad \text{with}\\
&d_{\theta,t}(\pi,\omega)\\
&\quad=(g(X_t,\pi)',Z_t',\omega_1'g_{\pi_1}(X_t,\pi_{\mathcal{I}_1}),\ldots,\omega_K'g_{\pi_{k-1}}(X_t,\pi_{\mathcal{I}_K}))'. \quad (3.21)
\end{aligned}$$

Assumption 5. (i) $\lambda_{\min}(H(\pi,\omega))\geq\varepsilon$, $\lambda_{\min}(\Omega_\theta(\pi,\omega))\geq\varepsilon$, for some $\varepsilon>0\ \forall\pi\in\Pi$, $\|\omega_k\|=1$, and $\gamma_0\in\Gamma$ for $k=1,\ldots,K$.

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

A similar condition is used in Assumption C6 of Andrews and Cheng (2012) and some sufficient conditions are discussed.

Theorem 1 (*Asymptotic Distribution of Estimators*). *Suppose Assumptions 1–5 hold. Then, under $\gamma_n\to\gamma_0$,*

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

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

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 $\psi_{K-}=(\beta',\zeta',\pi'_{\mathcal{I}_1},\ldots,\pi'_{\mathcal{I}_{K-1}})'$, $S_{\mathcal{I}_k}=[0_{d_k\times d_{k-}},I_{d_k},0_{d_k\times(d_\zeta+d_{k-})}]'$, and $\mathbf{B}(\beta_{K-})=diag\{(1_{d_\beta+d_\zeta},1_{d_{\pi_{\mathcal{I}_1}}}\|\beta_{\mathcal{I}_1}\|,\ldots,1_{d_{\pi_{\mathcal{I}_{K-1}}}}\|\beta_{\mathcal{I}_{K-1}}\|)'\}$.

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

$$n^{1/2}\mathbf{B}(\beta_n)\left(\widehat{\theta}-\theta_n\right)\to_d N(0,\Sigma(\pi_0,\omega_0)),$$

where $\mathbf{B}(\beta)=diag\{(1_{d_\beta+d_\zeta},1_{d_{\pi_{\mathcal{I}_1}}}\|\beta_{\mathcal{I}_1}\|,\ldots,1_{d_{\pi_{\mathcal{I}_{K-1}}}}\|\beta_{\mathcal{I}_K}\|)'\}$.

Comments. 1. In case (a), $\widehat{\psi}_{K-}=(\widehat{\beta}',\widehat{\zeta}',\widehat{\pi}'_{\mathcal{I}_1},\ldots,\widehat{\pi}'_{\mathcal{I}_{K-1}})'$ is consistent but it has a non-standard asymptotic distribution. The distribution involves the Gaussian process $\tau(\pi_K)$ and the inconsistent estimator $\pi^*_{\mathcal{I}_K}$, which minimizes the sample paths of the non-central chi-squared process $\chi(\pi_{\mathcal{I}_K})$ defined in (3.19). In addition, the rate of convergence of $\widehat{\pi}_{\mathcal{I}_2},\ldots,\widehat{\pi}_{\mathcal{I}_{K-1}}$ 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 $\mathbf{B}(\beta_n)$. Asymptotic distributions with mixed rate of convergence also appear in Antoine and Renault (2012).

Example (*Cont.*). In the example $y_t=\beta_1g_1(X_t,\pi_1)+\beta_2g_2(X_t,\pi_2)+U_t$, consider the distribution of the least squares estimator when $\beta_{1,n}\to0$, $|n^{1/2}\beta_{1,n}|\to\infty$, and $n^{1/2}\beta_{2,n}\to b_2\in\mathbb{R}$. 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

$$\begin{pmatrix} n^{1/2}\left(\widehat{\beta}-\beta_n\right)\\ n^{1/2}\beta_{1n}\left(\widehat{\pi}_1-\pi_{1,n}\right)\\ \widehat{\pi}_2\end{pmatrix}\Rightarrow\begin{pmatrix}\tau(\pi_2^*)-S_2b_2\\ \pi_2^*\end{pmatrix}, \tag{3.22}$$

where $S_2b_2=(0,b_2,0)'$, $G(\pi_2)$, $\tau(\pi_2)$, and π_2^* are as defined in (3.18) and (3.19) with

$$\begin{aligned}
&H_K(\pi_2,\pi_{2,0})=\mathbb{E}_{\gamma_0}d_{\psi_K,t}(\pi_{1,0},\pi_2)d_{\psi_K,t}(\pi_{1,0},\pi_{2,0})',\\
&\Omega(\pi_2,\widetilde{\pi}_2)=\mathbb{E}_{\gamma_0}U_t^2d_{\psi_K,t}(\pi_{1,0},\pi_2)d_{\psi_K,t}(\pi_{1,0},\widetilde{\pi}_2)', \quad \text{where}\\
&d_{\psi_K,t}(\pi_{1,0},\pi_2)=\left(g_1(X_t,\pi_{1,0}),g_2(X_t,\pi_2),g_{\pi_1}(X_t,\pi_{1,0})\right)'. \quad (3.23)
\end{aligned}$$

Note that the angle parameter does not show up in (3.23) because (i) β_1 is a scalar and (ii) β_{1n} instead of $|\beta_{1n}|$ is used for renormalization on the left hand side of (3.22). □

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 $\mathbf{B}^{-1}(\beta_0)\Sigma(\pi_0,\omega_0)\mathbf{B}^{-1}(\beta_0)$ is the asymptotic covariance matrix of the least squares estimator $\widehat{\theta}$. Following the definition of $\Sigma(\pi,\omega)$ in (3.20), we estimate $\Sigma(\pi,\omega)$ by

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

and $\widehat{U}_t$ is the regression residual. The covariance estimator $\widehat{\Sigma}$ is not always consistent because the estimators of π and ω are not always consistent. Its asymptotic distribution is given in (4.20) and (4.21). The standard definition of the Wald statistic for the null hypothesis $H_0:R\theta_n=v_n$ is

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

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 disassembles 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 with 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

$$R = [R_0 : R_1 : \cdots : R_K], \tag{4.3}$$

where R_0 is the submatrix of R associated with (β', ζ') and R_k is the submatrix of R associated with $\pi_{\mathcal{I}_k}$ for $k = 1, \ldots, K$. Thus, R_0 is a $d_r \times (d_\beta + d_\zeta)$ matrix and R_k is a $d_r \times d_{\pi_{\mathcal{I}_k}}$ matrix for $k = 1, \ldots, K$. Let

$$A = [A_0 : A_1 : \cdots : A_K] \in \mathcal{O}(d_r) \tag{4.4}$$

be an orthogonal matrix that satisfies two conditions below:

$$\text{(i) } A'R = \begin{bmatrix} A_0'R_0 & 0 & 0 & 0 & 0 \\ A_1'R_0 & A_2'R_1 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & 0 & 0 \\ A_{K-1}'R_0 & A_{K-1}'R_1 & \cdots & A_{K-1}'R_{K-1} & 0 \\ A_K'R_0 & A_K'R_1 & \cdots & A_K'R_{K-1} & A_K'R_K \end{bmatrix}$$

is block lower triangular (4.5)

and

$$\text{(ii) } R^* = \begin{bmatrix} A_0'R_0 & 0 & 0 & 0 & 0 \\ A_1'R_0 & A_1'R_1 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & A_{K-1}'R_{K-1} & 0 \\ 0 & 0 & 0 & 0 & A_K'R_K \end{bmatrix}$$

has full rank. (4.6)

This rotation matrix A can be obtained as follows. For $k = K$, let $d_K^* = rank(R_K)$ 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 $d_{K-1}^* = rank([R_{K-1} : R_K]) - rank(R_K)$ 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 $k = K - 2, \ldots, 1$. In each step, let

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

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 the $d_r \times d_0^*$ matrix A_0 ensures that A is an orthogonal matrix. When $d_{\pi_{\mathcal{I}_k}} = 0$, A_k disappears from the construction of A. 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

$$W_n(R) = W_n(A'R) = W_n(R^*) + \varepsilon_n, \tag{4.8}$$

where ε_n is explicitly defined as the difference between $W_n(A'R)$ and $W_n(R^*)$. In the proof of Theorem 2, we show that ε_n is asymptotically negligible.[6] Therefore, only the block-diagonal elements remain asymptotically and the asymptotic distribution of $W_n(R)$ is determined by that of $W_n(R^*)$. Note that the term $A_1'R_0$ does not disappear in R^* as the other off diagonal terms because $\pi_{\mathcal{I}_1}$ is the strong identification group and $\pi_{\mathcal{I}_1}$ and (β', ζ') have the same rate of convergence.

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 π_1 is semi-strongly identified (estimated at a rate slower than $n^{-1/2}$) and π_2 is weakly identified,

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

(ii) If π_1 and π_2 are both weakly identified,

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

4.2. Rescaling matrix for asymptotic singularity of covariance matrix

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

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

To deal with the asymptotic singularity of the covariance matrix, we start with the diagonal matrix $\mathbf{B}(\beta) = diag\{(1_{d_\beta + d_\zeta}, 1_{d_{\pi_{\mathcal{I}_1}}} \|\beta_{\mathcal{I}_1}\|, \ldots, 1_{d_{\pi_{\mathcal{I}_{K-1}}}} \|\beta_{\mathcal{I}_K}\|)'\}$. To deal with the asymptotic singularity of $\mathbf{B}(\widehat{\beta})$, define a new diagonal matrix $\mathbf{D}^*(\widehat{\beta})$ as

$$\mathbf{D}^*(\beta) = diag\{(1_{d_0^*}, \|\beta_{\mathcal{I}_1}\| 1_{d_1^*}, \|\beta_{\mathcal{I}_2}\| 1_{d_2^*}, \ldots, \|\beta_{\mathcal{I}_K}\| 1_{d_K^*})'\} \in R^{d_r \times d_r}, \tag{4.13}$$

where d_k^* is defined in (4.7). Note that

$$R^\dagger(\beta) = \mathbf{D}^*(\beta)R^*\mathbf{B}^{-1}(\beta) = \begin{bmatrix} A_0'R_0 & 0 & 0 & 0 & 0 \\ A_1'R_0\|\beta_{\mathcal{I}_1}\| & A_1'R_1 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & A_{K-1}'R_{K-1} & 0 \\ 0 & 0 & 0 & 0 & A_K'R_K \end{bmatrix}, \tag{4.14}$$

which is full rank for any β by construction. Therefore, with probability approaching one,

$$W_n(R^*) = W_n(\mathbf{D}^*(\widehat{\beta})R^*) = \rho_n' V_n^{-1} \rho_n, \tag{4.15}$$

[6] The analysis roughly goes as follows. Under the null, consider $A'R(\widehat{\theta} - \theta_n)$ in $W_n(A'R)$. For $k = 2, \ldots, K$,

$$A_k'R\left(\widehat{\theta} - \theta_n\right) = \sum_{\ell < k} A_k'R_\ell\left(\widehat{\psi}_{k-} - \psi_{k-,n}\right) + A_k'R_k\left(\widehat{\pi}_{\mathcal{I}_k} - \pi_{\mathcal{I}_k,n}\right) \quad \text{and}$$

$$\widehat{\psi}_{k-} - \psi_{k-,n} = o_p(\|\widehat{\pi}_{\mathcal{I}_k} - \pi_{\mathcal{I}_k,n}\|) \tag{4.9}$$

where the first equality holds because $A'R$ is upper block-diagonal and the second equality follows from Theorem 1. The remaining term $A_k'R_k\left(\widehat{\pi}_{\mathcal{I}_k} - \pi_{\mathcal{I}_k,n}\right)$ is the counterpart in $W_n(R^*)$.

where

$$\begin{aligned}\rho_n &= n^{1/2}\mathbf{D}^*(\widehat{\beta})R^*(\widehat{\theta}-\theta_n)\\ &= \left[\mathbf{D}^*(\widehat{\beta})R^*\mathbf{B}^{-1}(\widehat{\beta})\right]\left[n^{1/2}\mathbf{B}(\widehat{\beta})(\widehat{\theta}-\theta_n)\right]\\ &= R^{\dagger}(\widehat{\beta})\xi_n \quad \text{with } \xi_n = n^{1/2}\mathbf{B}(\widehat{\beta})(\widehat{\theta}-\theta_n),\end{aligned} \tag{4.16}$$

and

$$\begin{aligned}V_n &= \mathbf{D}^*(\widehat{\beta})R^*\mathbf{B}^{-1}(\widehat{\beta})\widehat{\Sigma}\mathbf{B}^{-1}(\widehat{\beta})R^{*\prime}\mathbf{D}^*(\widehat{\beta})\\ &= R^{\dagger}(\widehat{\beta})\widehat{\Sigma}R^{\dagger}(\widehat{\beta})'.\end{aligned} \tag{4.17}$$

An important implication of the calculation in (4.17) is that V_n is non-singular asymptotically and V_n^{-1} appears as the rescaling covariance matrix in (4.15). Below we derive the asymptotic distribution of ξ_n and $\widehat{\Sigma}$ under all identification scenarios, which in turn yields the asymptotic distribution of the Wald statistic following (4.15)–(4.17).

4.3. Non-standard distribution of the test statistic

First consider the re-centered and re-scaled parameter ξ_n defined in (4.16). Following the asymptotic distribution in Theorem 1(a), define a function of the Gaussian process $\tau(\pi_K)$:

$$\xi(\pi_{\mathcal{I}_K}) = \begin{pmatrix} \tau(\pi_{\mathcal{I}_K}) - S_{\mathcal{I}_K} b_{\mathcal{I}_K} \\ \|\tau_{\beta_K}(\pi_{\mathcal{I}_K})\| \left(\pi_{\mathcal{I}_K} - \pi_{\mathcal{I}_K,0}\right)\end{pmatrix}, \tag{4.18}$$

where $\tau_{\beta_K}(\pi_{\mathcal{I}_K}) = S'_{\mathcal{I}_K}\tau(\pi_K)$ are the elements of $\tau(\pi_{\mathcal{I}_K})$ associated with $\beta_{\mathcal{I}_K}$. Under weak identification, we show $\xi_n \Rightarrow \xi(\pi^*_{\mathcal{I}_K})$ in the proof of Theorem 2.

To handle $\widehat{\omega}$ in the estimation of $\Sigma(\pi,\omega)$, define

$$\omega(\pi_{\mathcal{I}_K}) = \left(\omega'_{1,0}, \omega'_{2,0}, \ldots, \omega'_{K-1,0}, \frac{\tau_{\beta_K}(\pi_{\mathcal{I}_K})'}{\|\tau_{\beta_K}(\pi_{\mathcal{I}_K})\|}\right)'. \tag{4.19}$$

For the strong and semi-strong identification groups, the angle parameters are estimated consistently. This is the reason that $\omega_{k,0}$ shows up in (4.19) for $k = 1,\ldots,K-1$. For group K, $\tau_{\beta_K}(\pi^*_{\mathcal{I}_K})/\|\tau_{\beta_K}(\pi^*_{\mathcal{I}_K})\|$ characterizes the limit of the angle parameter.

In the proof of Theorem 2, we show that

(a) under weak identification, i.e., $n^{1/2}\beta_{\mathcal{I}_K} \to b_{\mathcal{I}_K} \in R^{d_K}$,

$$\begin{aligned}&\xi_n \Rightarrow \xi(\pi^*_{\mathcal{I}_K}), \qquad \widehat{\omega} \Rightarrow \omega(\pi^*_{\mathcal{I}_K}),\\ &\widehat{\Sigma} \Rightarrow \Sigma(\pi_{K-,0}, \pi^*_{\mathcal{I}_K}, \omega(\pi^*_{\mathcal{I}_K})), \qquad R^{\dagger}(\widehat{\beta}) \to_p R^{\dagger}(\beta_0);\end{aligned} \tag{4.20}$$

(b) without weak identification, i.e., $\|n^{1/2}\beta_{\mathcal{I}_K}\| \to \infty$,

$$\begin{aligned}&\xi_n \to_d \xi \sim N(0, \Sigma(\pi_0,\omega_0)), \qquad \widehat{\omega} \to_p \omega_0,\\ &\widehat{\Sigma} \to_p \Sigma(\pi_0,\omega_0), \qquad R^{\dagger}(\widehat{\beta}) \to_p R^{\dagger}(\beta_0).\end{aligned} \tag{4.21}$$

All convergence holds jointly. Put the distributions in (4.20) and (4.21) together with the decomposition in (4.15)–(4.17), the asymptotic distribution of the Wald statistic is given below.

Theorem 2 (*Wald Statistic with Mixed Identification Strength*). *Suppose Assumptions 1–5 hold. Then, under $\gamma_n \to \gamma_0$,*

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

$$W_n(R) \Rightarrow \mathcal{W}(\pi^*_{\mathcal{I}_K}), \quad \text{where}$$

$$\begin{aligned}\mathcal{W}(\pi_{\mathcal{I}_K}) = &\left[R^{\dagger}(\beta_0)\xi(\pi_{\mathcal{I}_K})\right]'\left[R^{\dagger}(\beta_0)\Sigma(\pi_{\mathcal{I}_K})R^{\dagger}(\beta_0)'\right]^{-1}\\ &\times\left[R^{\dagger}(\beta_0)\xi(\pi_{\mathcal{I}_K})\right],\end{aligned}$$

where $\Sigma(\pi_{\mathcal{I}_K})$ abbreviates $\Sigma(\pi_{K-,0}, \pi_{\mathcal{I}_K}, \omega(\pi_{\mathcal{I}_K}))$.

(b) *without weakly identified parameters: If $\|n^{1/2}\beta_{\mathcal{I}_K}\| \to \infty$, $W_n(R) \to_d \chi^2_{d_r}$.*

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 $R^{\dagger}(\beta_0)$. In $R^{\dagger}(\beta_0)$, the rotation matrices $A_0,\ldots,A_K$ are only specified up to orthogonal rotation. The distribution $\mathcal{W}(\pi_{\mathcal{I}_K})$ is invariant to orthogonal rotations of each of these matrices.

2. Theorem 2 shows that the Wald statistic has a non-standard asymptotic null 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 asymptotic null distribution 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 null distribution of the Wald statistic. The Wald statistic for tests with linear restrictions is self-corrected when all parameters are consistently estimated. A similar self-correction result for the Wald statistic also is obtained by Antoine and Renault (2012) when parameters have mixed rates of convergence.

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

$$t_n(R) = \frac{n^{1/2}\left(R\widehat{\theta} - v_n\right)}{\sqrt{R\mathbf{B}^{-1}(\widehat{\beta})\widehat{\Sigma}\mathbf{B}^{-1}(\widehat{\beta})R'}}. \tag{4.22}$$

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, under $\gamma_n \to \gamma_0$,*

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

$$t_n(R) \Rightarrow \mathcal{T}(\pi^*_{\mathcal{I}_K}), \quad \text{where } \mathcal{T}(\pi_{\mathcal{I}_K}) = \frac{R^{\dagger}(\beta_0)\xi(\pi_{\mathcal{I}_K})}{\sqrt{R^{\dagger}(\beta_0)\Sigma(\pi_{\mathcal{I}_K})R^{\dagger}(\beta_0)'}};$$

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

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 $|n^{1/2}\beta_{1,n}| \to \infty$, 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 (4.18) with elements of $\tau(\pi_2)$ given in (3.23), $S_2 = (0, 1, 0)'$, $b_2 = 0$.

(2) If $n^{1/2}\beta_{1,n} \to b_1 \in R$, $\mathcal{I}_K = \{1, 2\}$ 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 (3.18) and (3.19) with

$$\begin{aligned}&H_K(\pi,\pi_0) = \mathbb{E}_{\gamma_0} d_{\psi_K,t}(\pi) d_{\psi_K,t}(\pi_0)',\\ &\Omega(\pi,\widetilde{\pi}) = \mathbb{E}_{\gamma_0} U_t^2 d_{\psi_K,t}(\pi) d_{\psi_K,t}(\widetilde{\pi}_2)', \quad \text{where}\\ &d_{\psi_K,t}(\pi) = (g_1(X_t,\pi_1), g_2(X_t,\pi_2))',\end{aligned} \tag{4.23}$$

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

$$S_{\mathcal{I}_K} b_{\mathcal{I}_K} = b = (b_1, 0)', \qquad \tau_{\beta_K}(\pi_{\mathcal{I}_K}) = \tau(\pi). \quad \square \tag{4.24}$$

4.4. Asymptotic distribution of the wald statistic under the alternative

Next, we consider the asymptotic distribution of the Wald test under local and fixed alternatives. Consider the null hypothesis: $H_0 : R\theta_n = v_n^{null}$, where $R\theta_n \neq v_n^{null}$. The null value v_n^{null} is allowed to depend on n. Similar to (4.15)–(4.17), we can show that in this case

$$W_n(R) = \left(R^\dagger(\widehat{\beta})\xi_n + \Delta_n\right)' \left[R^\dagger(\widehat{\beta})\widehat{\Sigma}R^\dagger(\widehat{\beta})'\right]^{-1} \times \left(R^\dagger(\widehat{\beta})\xi_n + \Delta_n\right) + o_p(\|\Delta_n\|^2) + o_p(1), \tag{4.25}$$

where

$$\Delta_n = n^{1/2}\mathbf{D}^*(\widehat{\beta})A'\left(R\theta_n - v_n^{null}\right) \tag{4.26}$$

is the additional term that appears under the alternative.[7] The asymptotic distribution of $R^\dagger(\widehat{\beta})$, ξ_n, and $\widehat{\Sigma}$ are discussed in (4.20) and (4.21). Local alternatives are defined by values of θ_n such that Δ_n is stochastically bounded and non-degenerate. This depends on the identification scenario and the restriction matrix R.

To be more specific on the appropriate local alternatives, we discuss the following cases. First, consider $R\theta = R_\beta\beta$, i.e., the test is on β. In this case, $\Delta_n = n^{1/2}(R_\beta\beta_n - v_n^{null})$. Under the local alternative $\Delta_n \to d \in R^{d_r}$, the asymptotically distribution of the Wald statistic is given by that in Theorem 2(a) with $R^\dagger(\beta_0)\xi(\pi_{\mathcal{I}_K})$ replaced by $R^\dagger(\beta_0)\xi(\pi_{\mathcal{I}_K}) + d$ under weak identification and the asymptotic distribution becomes a non-central $\chi^2_{d_r}$ distribution with noncentrality parameter $d'[R^\dagger(\beta_0)\widehat{\Sigma}R^\dagger(\beta_0)]^{-1}d$ without weak identification. Under the fixed alternative $R_\beta\beta_n - v_n^{null} \to d_0 \neq 0$, the Wald statistic diverges to ∞ in probability with or without weak identification.

Next, consider $R\theta = R_\pi\pi_1$, i.e., the test is on π_1. In this case, $\Delta_n = n^{1/2}\|\widehat{\beta}_1\|(R_\pi\pi_1 - v_n^{null})$. The appropriate local alternative varies with the identification strength of π_1. (i) If $n^{1/2}\|\beta_{1,n}\| \to b_1 \in R^{\beta_1}$, we have $n^{1/2}\|\widehat{\beta}_1\| = O_p(1)$. Under any local alternative $R_\pi\pi_{1,n} - v_n^{null} \to 0$, the asymptotic distribution of the Wald statistic is the same as that under the null $R_\pi\pi_1 = v_n^{null}$. Under the fixed alternative $R_\pi\pi_1 - v_n^{null} \to d_0 \neq 0$, we have $\Delta_n \to_p \|\tau_{\beta_1}(\pi_K^*)\|d_0$, where $\tau_{\beta_1}(\pi_K)$ is a subvector of $\tau(\pi_{\mathcal{I}_K})$ associated with β_1. In this case, the Wald statistic has the same limit as in Theorem 2(a) with $R^\dagger(\beta_0)\xi(\pi_{\mathcal{I}_K})$ replaced by $R^\dagger(\beta_0)\xi(\pi_{\mathcal{I}_K}) + \|\tau_{\beta_1}(\pi_K^*)\|d_0$. (ii) If $\|n^{1/2}\beta_{1,n}\| \to \infty$, the appropriate local alternative is defined by $n^{1/2}\|\beta_{1,n}\|(R_\pi\pi_{1,n} - v_n^{null}) \to d \in R^{d_{\pi_1}}$. In this case, $\Delta_n \to_d d$. The asymptotic distribution of the Wald statistic is given by that in Theorem 2(a) with $R^\dagger(\beta_0)\xi(\pi_{\mathcal{I}_K})$ replaced by $R^\dagger(\beta_0)\xi(\pi_{\mathcal{I}_K}) + d$ under weak identification and the asymptotic distribution becomes a non-central $\chi^2_{d_r}$ distribution with noncentrality parameter $d'[R^\dagger(\beta_0)\widehat{\Sigma}R^\dagger(\beta_0)]^{-1}d$ without weak identification. Under the fixed alternative $R_\pi\pi_1 - v_n^{null} \to d_0 \neq 0$, the Wald statistic diverges to ∞ in probability with or without weak identification. To sum up, the appropriate non-degenerate local alternative depend on both the parameter of interest and the identification scenarios.

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

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

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

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

which is the lim inf 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

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

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

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

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

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

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

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

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

For a standard Wald test, the $1-\alpha$ quantile of $\chi^2_{d_r}$, denoted by $\chi^2_{d_r,1-\alpha}$, 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) The distribution function (df) of $\mathcal{W}(h)$ is continuous at $\chi^2_{d_r,1-\alpha}\forall h \in H$.

(ii) The df function of $|\mathcal{T}(h)|$ 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* $\sup_{h\in H(v)} \Pr(\mathcal{W}(h) > \chi^2_{d_r,1-\alpha})$;

(b) *the asymptotic size of a standard Wald confidence set is* $\inf_{h\in H} \Pr(\mathcal{W}(h) \leq \chi^2_{d_r,1-\alpha})$;

(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 $\chi^2_{d_r,1-\alpha}$ 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. The results in Theorem 3 combine the pointwise results in Theorem 2 to obtain the uniform results of asymptotic size in (5.1) and (5.2). Roughly speaking, the supremum or infimum in the definition of the asymptotic size of a test or confidence set is achieved along certain convergent subsequences and we show that these limits can be represented by those of the sequences indexed by $h \in H$. The proof applies the generic results in Andrews et al. (2011). If Assumption CV1 does not hold, the asymptotic size can be replaced by bounds following the method in Andrews and Guggenberger (2010) and Andrews et al. (2011).

[7] Details of the arguments are provided in the Appendices.

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 $\mathcal{I}_2, \ldots, \mathcal{I}_{K-1}$ 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 than one obtained by all possible group specifications without selection.

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

$$ICS_{j,n} = \left(n\widehat{\beta}_j'(\widehat{\Sigma}_j)^{-1}\widehat{\beta}_j/d_{\beta_j}\right)^{1/2}, \tag{5.7}$$

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

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

where $\{\kappa_{j,n} : n \geq 1\}$ is a sequence of constants such that $\kappa_{j,n} \to \infty$ and $\kappa_{j,n}/n^{1/2} \to 0$ for $j = 1, \ldots, p$.[8] 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

$$\widehat{H} = \{h \in H : \mathcal{I}_K = \widehat{\mathcal{I}}_W,\ \omega_{\mathcal{I}_k} = \widehat{\beta}_{\mathcal{I}_k}/\|\widehat{\beta}_{\mathcal{I}_k}\| \text{ and } \pi_{\mathcal{I}_k} = \widehat{\pi}_k \text{ for } k < K\}.\text{[9]} \tag{5.9}$$

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

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

Because $\widehat{H}$ is a subset of H, $\widehat{c}_{n,1-\alpha}$ 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 $\widehat{c}_{n,1-\alpha}(v)$ is obtained by replacing $\widehat{H}$ with $\widehat{H}(v) = \widehat{H} \cap H(v)$. 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 the test and confidence set based on this plug-in critical value the robust test and robust confidence set.

In empirical implementation, the first step is to specify $\widehat{H}$ 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 are given in Hansen (1996). Finally, obtain the plug-in critical value following (5.10). The difficulty in 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 weakly-identified nonlinear regressors in group $\mathcal{I}_K$ increases. For this reason, the current paper suggests a simple data-dependent rule in (5.8). The smooth-transition method considered by Andrews and Barwick (2012) and the type 2 robust critical value of Andrews and Cheng (2012) can be applied as well but the computation is more intensive.

The critical value in (5.10) treats the unknown parameter h by reducing its parameter space from H to $\widehat{H}$ and take the supremum over $\widehat{H}$. Alternatively, one can consider the Bonferroni method, which constructs a confidence set for h and takes the supremum over this confidence set. McCloskey (2012) studies the Bonferroni method in non-standard problems and its various refinements.

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

(ii) The df function of $\mathcal{W}(h)$ is continuous at $\mathcal{W}_{1-\alpha}(h)$ 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 $1-\alpha$.*

Example (*Cont.*)**.** Fig. 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 Fig. 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 Fig. 1 and the robust critical value in Fig. 2. The left panel of Fig. 2 is obtained by drawing the t statistic and the *ICS* statistic from their asymptotic distributions.[10] Both figures demonstrate how the null rejection probability of the test changes with the true values of β_1 and β_2.

Table 1 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 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 $b_1/\sqrt{n}$

[8] To see the requirement $\kappa_{j,n}/n^{1/2} \to 0$, consider a strong identification case where $\beta_{j,n}$ is bounded away from 0, say $\beta_{j,n} = 1$. In this case, $\widehat{\beta}_j$ converges to 1 in probability so that $ICS_{j,n}$ diverges to infinity at rate $n^{1/2}$. To ensure $ICS_{j,n}$ is larger than $\kappa_{j,n}$ with probability approaching 1, we need $\kappa_{j,n}$ diverge at a rate slower than $n^{1/2}$, which leads to $\kappa_{j,n}/n^{1/2} \to 0$. This upper bound for $\kappa_{j,n}$ is given by the strong identification case, whereas the lower bound $\kappa_{j,n} \to \infty$ is given by the weak identification case.

[9] The 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.

[10] The asymptotic distribution of the t statistic and the *ICS* statistic are given in Corollary 1 and (C.12) in the Appendix B. The *ICS* statistics are non-centered t statistics. Thus, their asymptotic distributions follow the same arguments for the t statistic.

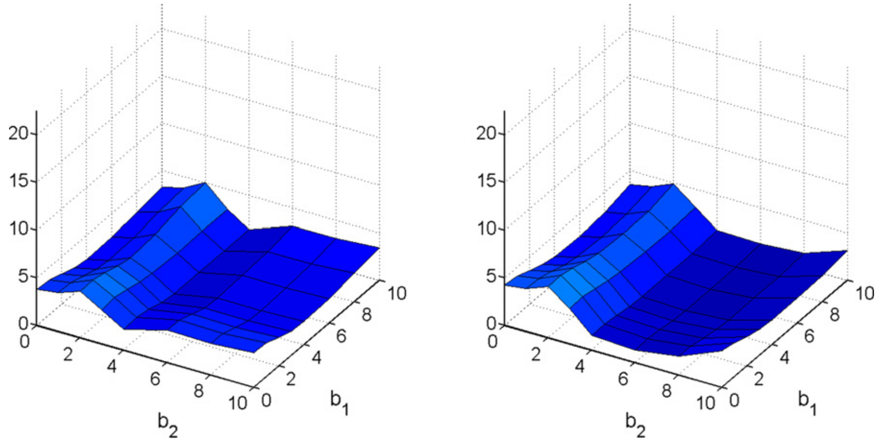

Fig. 2. Robust Test: Asymptotic (left) and Finite-Sample (right, $n = 500$) Rejection Probability ($\times 100$) for $H_0 : \beta_2 = \beta_{2,0}$. Notes: DGP is the same as that for Fig. 1, nominal size $\alpha = 5\%$; the true values of β_1 and β_2 are $\beta_{1,0} = b1/\sqrt{500}$ and $\beta_{2,0} = b2/\sqrt{500}$ in the right panel.

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	5.4	5.3	21.0	19.9	19.7
	1	6.1	5.2	5.0	20.0	19.1	18.8
	2	5.7	4.8	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	5.5	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

and $b_2/\sqrt{n}$, respectively, for finite-sample results with sample size n.[11] The true values of π_1 and π_2 are both 0 for Figs. 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 Fig. 2 is comparable to the right panel of Fig. 1 with the standard test replaced by the robust test. The left panel of Fig. 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 Fig. 2.

In Fig. 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

[11] In 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.

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 Fig. 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 Fig. 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 Fig. 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%.

Table C.1 in the Appendix C reports the power of the robust and standard tests under the local alternative $\beta_{2n} = n^{-1/2} b_2$ for b_2 from 1 to 10. As in Table 1, we also consider b_1 from 1 to 10. Because the robust test and the standard test have different projection probability under the null, we adjust the rejection probability of the standard test by a constant such that the null rejection probability of the robust test and the standard test are the same for any (b_1, b_2). Table C.1 shows that the robust test is less powerful than the standard test but the power loss is mild and it mainly occurs for small values of b_2. □

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 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 et al. (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

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.

Appendix 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) *Assumptions* 2(i) *or Assumption* 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) $\sup_{\theta\in\Theta} \|n^{-1}\sum_{t=1}^{n} s(W_t, \theta) - E_{\gamma_0} s(W_t, \theta)\| \to_p 0$ *under any sequence of true parameters* $\{\gamma_n \in \Gamma : n \geq 1\}$ *and* $\gamma_n \to \gamma_0 \in \Gamma$. (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* $\limsup_{n\to\infty} P [\sup_{\theta_1,\theta_2\in\Theta:\|\theta_1-\theta_2\|<\delta} \|\nu_n s(\theta_1) - \nu_n s(\theta_2)\| > \eta] < \varepsilon \forall \gamma_0 \in \Gamma$.

(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,* $n^{-1/2}\sum_{t=1}^{n}(s(W_t) - E_{\gamma_n} s(W_t)) \to_d N(0, V_s(\gamma_0))$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0) \forall \gamma_0 \in \Gamma$, *where* $V_s(\gamma_0) = \sum_{m=-\infty}^{\infty} Cov_{\gamma_0}(s(W_t), s(W_{t+m}))$.

(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.

Appendix B. Proofs for asymptotic distributions of estimators and test statistics

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

$$Q_n(\theta) = n^{-1}\sum_{t=1}^{n} U_t^2(\theta)/2, \quad \text{where}$$

$$\begin{aligned} U_t(\theta) &= Y_t - g(X_t, \pi)'\beta - Z_t'\zeta \\ &= U_t + g(X_t, \pi_n)'\beta_n + Z_t'\zeta_n - g(X_t, \pi)'\beta - Z_t'\zeta. \end{aligned} \tag{B.1}$$

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

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

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

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

Let ψ_n denote the true value of ψ for sample size n and $\psi_n \to \psi_0$. We write the criterion function $Q(\theta)$ as $Q(\psi, \pi_{\mathcal{I}_1}|\pi_{1+})$ and analyze $Q(\psi, \pi_{\mathcal{I}_1}|\pi_{1+})$ as a function of $(\psi, \pi_{\mathcal{I}_1})$ for a fixed value of π_{1+}.

Now we show that for any π_{1+}, $Q(\psi, \pi_{\mathcal{I}_1}|\pi_{1+})$ is uniquely minimized by $(\psi_0, \pi_{\mathcal{I}_1,0})$. Note that $\beta_{\mathcal{I}_k,0} = 0$ for $k > 1$ by the grouping rule. Therefore, $Q(\psi_0, \pi_{\mathcal{I}_1,0}|\pi_{1+}) = \mathbb{E}_{\gamma_0} U_t^2/2$. For fixed π_{1+},

$$\begin{aligned} &Q(\psi, \pi_{\mathcal{I}_1}|\pi_{1+}) - Q(\psi_0, \pi_{\mathcal{I}_1,0}|\pi_{1+}) \\ &\quad = \mathbb{E}_{\gamma_0}\Bigg[g_{\mathcal{I}_1}(X_t, \pi_{\mathcal{I}_1,0})'\beta_{\mathcal{I}_1,0} - g_{\mathcal{I}_1}(X_t, \pi_{\mathcal{I}_1})'\beta_{\mathcal{I}_1} \\ &\qquad - \sum_{k>1}^{K} g_{\mathcal{I}_k}(X_t, \pi_{\mathcal{I}_k})'\beta_{\mathcal{I}_k} + Z_t'(\zeta_0 - \zeta)\Bigg]^2/2. \end{aligned} \tag{B.4}$$

By Assumption 3,

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

for any $a \neq 0$ and $\pi \neq \pi_0$. Because $\beta_{\mathcal{I}_1,0} \neq 0$, the rhs of (B.4) is greater than 0 for any $\pi_{\mathcal{I}_1} \neq \pi_{\mathcal{I}_1,0}$. When $\pi_{\mathcal{I}_1} = \pi_{\mathcal{I}_1,0}$, (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 $Q(\psi, \pi_{\mathcal{I}_1}|\pi_{1+})$ is uniquely minimized by $(\psi_0, \pi_{\mathcal{I}_1,0})$ for any π_{1+}, (ii) $Q(\psi, \pi_{\mathcal{I}_1}|\pi_{1+})$ is continuous, and (iii) the parameter spaces are all compact, we have the identification uniqueness condition

$$\inf_{\pi_{1+}\in\Pi_{1+}} \inf_{\psi\in\Psi, \pi_{\mathcal{I}_1}\in\Pi_1} \left\{Q(\psi, \pi_{\mathcal{I}_1}|\pi_{1+}) - Q(\psi_0, \pi_{\mathcal{I}_1,0}|\pi_{1+})\right\} > 0 \qquad \text{(B.6)}$$

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. □

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

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

$$\sup_{\pi_{1+}\in\Pi_{1+}} \left\|\widehat{\pi}_{\mathcal{I}_1}(\pi_{1+}) - \pi_{\mathcal{I}_1,n}\right\| \to_p 0, \qquad \text{(B.7)}$$

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

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

uniformly over π_{1+}, which follows from Lemma 1, $\beta_{\mathcal{I}_1,n} \to \beta_{\mathcal{I}_1,0} \neq 0$, and $\beta_{\mathcal{I}_k,n} \to \beta_{\mathcal{I}_k,0} = 0$ for $k > 1$.

Step 2. Suppose Lemma 2 holds for $k-1$. For $\psi_{k-} = (\beta', \zeta', \pi'_{\mathcal{I}_1}, \ldots, \pi'_{\mathcal{I}_{k-1}})'$, the result for $k-1$ yields uniform consistency of $\widehat{\psi}_{k-}(\pi_{\mathcal{I}_k}, \pi_{k+})$ over $(\pi_{\mathcal{I}_k}, \pi_{k+})$. Now we show Lemma 2 holds for k.

For $k = 1, \ldots, K$, $g_{\mathcal{I}_k}(X_t, \pi_{\mathcal{I}_k})$ is the collection of regressors in group k. The model can be equivalently written as

$$Y_t = \sum_{k=1}^{K} g_{\mathcal{I}_k}(X_t, \pi_{\mathcal{I}_k})'\beta_{\mathcal{I}_k} + Z_t'\zeta + U_t. \qquad \text{(B.9)}$$

Define the first and second order derivatives as

$$g_{\pi_k}(X_t, \pi_{\mathcal{I}_k}) = \frac{\partial}{\partial \pi'_{\mathcal{I}_k}} g_{\mathcal{I}_k}(X_t, \pi_{\mathcal{I}_k}) \in R^{d_k \times d_{\pi_{\mathcal{I}_k}}} \quad \text{and}$$

$$g_{\pi\pi_k}(X_t, \pi_{\mathcal{I}_k}) = \frac{\partial}{\partial \pi'_{\mathcal{I}_k}} vec(g_{\pi_k}(X_t, \pi_{\mathcal{I}_k})') \in R^{(d_k d_{\pi_{\mathcal{I}_k}}) \times d_{\pi_{\mathcal{I}_k}}}, \qquad \text{(B.10)}$$

where d_k is the dimension of $g_{\mathcal{I}_k}(X_t, \pi_{\mathcal{I}_k})$ and $\beta_{\mathcal{I}_k}$. The angle parameters are

$$\omega_k = \beta_{\mathcal{I}_k}/\|\beta_{\mathcal{I}_k}\| \quad \text{and} \quad \omega_{k-} = (\omega_1', \ldots, \omega_{k-1}')'. \qquad \text{(B.11)}$$

Let $D^1_{\psi_k}(\theta)$ and $D^2_{\psi_k}(\theta)$ denote the first and second order partial derivatives of $Q_n(\theta)$ wrt ψ_{k-}, where $\theta = (\psi'_{k-}, \pi'_{\mathcal{I}_k}, \pi'_{k+})'$. The first order derivative wrt $\psi_k = (\beta', \zeta', \pi'_{\mathcal{I}_1}, \ldots, \pi'_{\mathcal{I}_{k-1}})$ is

$$\begin{aligned} D^1_{\psi_k}(\theta) &= -n^{-1}\sum_{t=1}^{n} \begin{pmatrix} g(X_t, \pi) \\ Z_t \\ g_{\pi_1}(X_t, \pi_{\mathcal{I}_1})'\beta_{\mathcal{I}_1} \\ \vdots \\ g_{\pi_{k-1}}(X_t, \pi_{\mathcal{I}_{k-1}})'\beta_{\mathcal{I}_{k-1}} \end{pmatrix} U_t(\theta) \\ &= -n^{-1}\sum_{t=1}^{n} \begin{pmatrix} g(X_t, \pi) \\ Z_t \\ g_{\pi_1}(X_t, \pi_{\mathcal{I}_1})'\omega_1\|\beta_{\mathcal{I}_1}\| \\ \vdots \\ g_{\pi_{k-1}}(X_t, \pi_{\mathcal{I}_{k-1}})'\omega_{k-1}\|\beta_{\mathcal{I}_{k-1}}\| \end{pmatrix} U_t(\theta) \\ &= -n^{-1}\sum_{t=1}^{n} \mathbf{B}(\beta_{k-}) d_{\psi_k,t}(\theta) U_t(\theta), \end{aligned} \qquad \text{(B.12)}$$

where by definition

$$\begin{aligned} &d_{\psi_k,t}(\pi, \omega_{k-}) \\ &\quad = (g(X_t, \pi)', Z_t', \omega_1' g_{\pi_1}(X_t, \pi_{\mathcal{I}_1}), \ldots, \omega'_{k-1} g_{\pi_{k-1}}(X_t, \pi_{\mathcal{I}_{k-1}}))', \end{aligned}$$

and

$$\mathbf{B}(\beta_{k-}) = diag\{(1_{d_\beta + d_\zeta}, 1_{d_{\pi_{\mathcal{I}_1}}}\|\beta_{\mathcal{I}_1}\|, \ldots, 1_{d_{\pi_{\mathcal{I}k-1}}}\|\beta_{\mathcal{I}_{k-1}}\|)'\}. \qquad \text{(B.13)}$$

The second order derivative wrt $\psi_k = (\beta', \zeta', \pi'_{\mathcal{I}_1}, \ldots, \pi'_{\mathcal{I}_{k-1}})'$ is

$$\begin{aligned} D^2_{\psi_k}(\theta) \;=\; \mathbf{B}(\beta_{k-}) \Bigg(& n^{-1}\sum_{t=1}^{n} d_{\psi_k,t}(\pi, \omega_{k-}) d_{\psi_k,t}(\pi, \omega_{k-})' \\ & - \; n^{-1}\sum_{t=1}^{n} d^*_{\psi_k,t}(\theta) U_t(\theta) \Bigg) \mathbf{B}(\beta_{k-}), \quad \text{where} \end{aligned}$$

$$d^*_{\psi_k,t}(\theta) = \begin{pmatrix} 0_{d_\beta \times d_\beta} & 0_{d_\beta \times d_\zeta} & \delta^\pi_{k-1}(X_t, \theta) \\ 0_{d_\zeta \times d_\beta} & 0_{d_\zeta \times d_\zeta} & 0_{d_\zeta \times d_\pi} \\ \delta^\pi_{k-1}(X_t, \theta)' & 0_{d_\pi \times d_\zeta} & \delta^{\pi\pi}_{k-1}(X_t, \theta) \end{pmatrix} \qquad \text{(B.14)}$$

and by definition

$$\begin{aligned} &\delta^\pi_{k-1}(X_t, \theta) \\ &\quad = \begin{pmatrix} \|\beta_{\mathcal{I}_1}\|^{-1} g_{\pi_1}(X_t, \pi_{\mathcal{I}_1}) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \|\beta_{\mathcal{I}_{k-1}}\|^{-1} g_{\pi_{k-1}}(X_t, \pi_{\mathcal{I}_{k-1}}) \end{pmatrix} \in R^{d_\beta \times d_\pi} \end{aligned} \qquad \text{(B.15)}$$

and

$$\begin{aligned} &\delta^{\pi\pi}_{k-1}(X_t, \theta) \\ &\quad = \begin{pmatrix} h_1(X_t, \theta) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & h_{k-1}(X_t, \theta), \end{pmatrix} \in R^{d_\pi \times d_\pi}, \quad \text{where} \end{aligned}$$

$$\begin{aligned} h_\ell(X_t, \theta) \;&=\; \|\beta_{\mathcal{I}_\ell}\|^{-1} \left(\omega'_\ell \otimes I_{d_{\pi_\ell}}\right) \frac{\partial}{\partial \pi'_{\mathcal{I}_\ell}} vec(g_{\pi_\ell}(X_t, \pi_{\mathcal{I}_\ell})') \\ &=\; \|\beta_{\mathcal{I}_\ell}\|^{-1} \left(\omega'_\ell \otimes I_{d_{\pi_\ell}}\right) g_{\pi\pi_\ell}(X_t, \pi_{\mathcal{I}_\ell}). \end{aligned} \qquad \text{(B.16)}$$

Recall that

$$\begin{aligned} &\psi^0_{k-,n} = (\beta_{k-,n}, \beta^0_{\mathcal{I}_k}, \beta^0_{k+}, \zeta_n, \pi_{k-,n}), \\ &\quad \text{where } \beta^0_{\mathcal{I}_k} = 0 \text{ and } \beta^0_{k+} = 0. \end{aligned} \qquad \text{(B.17)}$$

We set $\beta^0_{\mathcal{I}_k} = 0$ and $\beta^0_{k+} = 0$ in $\psi^0_{k-,n}$ so that the criterion function $Q_n(\theta)$ does not depend on $(\pi_{\mathcal{I}_k}, \pi_{k+})$ when evaluated at $\psi^0_{k-,n}$. Hence, we write $Q_n(\psi^0_{k-,n}) = Q_n(\psi^0_{k-,n}, \pi_{\mathcal{I}_k}, \pi_{k+})$.

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

$$\begin{aligned} 0 &= D^1_{\psi_k}(\widehat{\psi}_{k-}(\pi_{\mathcal{I}_k}, \pi_{k+}), \pi_{\mathcal{I}_k}, \pi_{k+}) \\ &= D^1_{\psi_k}(\psi^0_{k-,n}, \pi_{\mathcal{I}_k}, \pi_{k+}) \\ &\quad + D^2_{\psi_k}(\psi^*_{k-,n}, \pi_{\mathcal{I}_k}, \pi_{k+}) \left(\widehat{\psi}_{k-}(\pi_{\mathcal{I}_k}, \pi_{k+}) - \psi^0_{k-,n}\right), \end{aligned} \qquad \text{(B.18)}$$

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

$$\begin{aligned}&\widehat{\psi}_{k^-}(\pi_{\mathcal{I}_k}, \pi_{k^+}) - \psi^0_{k^-,n}\\ &\quad = -\left[D^2_{\psi_k}(\psi^*_{k^-,n}, \pi_{\mathcal{I}_k}, \pi_{k^+})\right]^{-1} D^1_{\psi_k}(\psi^0_{k^-,n}, \pi_{\mathcal{I}_k}, \pi_{k^+}).\end{aligned}\tag{B.19}$$

We first study the first-order partial derivative in (B.19). Normalize it by $\left[\mathbf{B}(\beta_{k^-,n})\right]^{-1}$,

$$\begin{aligned}&\left[\mathbf{B}(\beta_{k^-,n})\right]^{-1} D^1_{\psi_k}(\psi^0_{k^-,n}, \pi_{\mathcal{I}_k}, \pi_{k^+})\\ &\quad = -n^{-1}\sum_{t=1}^{n} d_{\psi_k,t}(\pi_{k^-,n}, \pi_{\mathcal{I}_k}, \pi_{k^+}, \omega_{k^-,n})\\ &\qquad \times \left[g_k(X_t, \pi_{\mathcal{I}_k,n})'\beta_{\mathcal{I}_k,n} + g_{k^+}(X_t, \pi_{k^+,n})'\beta_{k^+,n} + U_t\right].\end{aligned}\tag{B.20}$$

We normalize both sides of (B.20) by $\|\beta_{\mathcal{I}_k,n}\|^{-1}$ and obtain

$$\begin{aligned}&\|\beta_{\mathcal{I}_k,n}\|^{-1}\left(\left[\mathbf{B}(\beta_{k^-,n})\right]^{-1} D^1_{\psi_k}(\psi^0_{k^-,n}, \pi_{\mathcal{I}_k}, \pi_{k^+})\right)\\ &\quad \to_p \; -\Phi_k(\pi_{\mathcal{I}_k}, \pi_{\mathcal{I}_k,0}|\pi_{k^+})\omega_{k,0}, \quad \text{where}\\ &\Phi_k(\pi_{\mathcal{I}_k}, \pi_{\mathcal{I}_k,0}|\pi_{k^+})\\ &\quad = \mathbb{E}_{\gamma_0} d_{\psi_k,t}(\pi_{k^-,0}, \pi_{\mathcal{I}_k}, \pi_{k^+}, \omega_{k^-,0}) g_k(X_t, \pi_{\mathcal{I}_k,0})'.\end{aligned}\tag{B.21}$$

The convergence follows from (i) applying Lemma A.1 to $n^{-1}\sum_{t=1}^n d_{\psi_k,t}(\pi_{k^-,n}, \pi_{\mathcal{I}_k}, \pi_{k^+}, \omega_{k^-,n})\ \ g_k(X_t, \pi_{\mathcal{I}_k,n})'$ and $n^{-1}\sum_{t=1}^n d_{\psi_k,t}(\pi_{k^-,n}, \pi_{\mathcal{I}_k}, \pi_{k^+}, \omega_{k^-,n}) g_{k^+}(X_t, \pi_{k^+,n})'$, (ii) applying Lemmas A.2 and A.3 to the empirical process $n^{-1/2}\sum_{t=1}^n d_{\psi_k,t}(\pi_{k^-,n}, \pi_{\mathcal{I}_k}, \pi_{k^+}, \omega_{k^-,n})U_t$, (iii) $\beta_{k^+,n} = o(\|\beta_{\mathcal{I}_k,n}\|)$, and (iv) $\|n^{1/2}\beta_{\mathcal{I}_k,n}\| \to \infty$. Note that $\Phi_k(\pi_{\mathcal{I}_k}, \pi_{\mathcal{I}_k,0}|\pi_{k^+}) = H_k(\pi_{\mathcal{I}_k}, \pi_{\mathcal{I}_k,0}|\pi_{k^+})S_k$, where $H_k(\pi_{\mathcal{I}_k}, \pi_{\mathcal{I}_k,0}|\pi_{k^+})$ is defined in (3.15) and S_k is a selector matrix such that $g_k(X_t, \pi_{\mathcal{I}_k,0}) = S_k' d_{\psi_k,t}(\pi_{k^-,0}, \pi_{\mathcal{I}_k}, \pi_{k^+}, \omega_{k^-,0})$.

Next we study the second-order partial derivative in (B.19). Pre- and post-multiply $D^2_{\psi_k}(\theta)$ by $[\mathbf{B}(\beta_{k^-})]^{-1}$,

$$\begin{aligned}&[\mathbf{B}(\beta_{k^-})]^{-1}D^2_{\psi_k}(\theta)[\mathbf{B}(\beta_{k^-})]^{-1}\\ &\quad = n^{-1}\sum_{t=1}^n d_{\psi_k,t}(\pi)d_{\psi_k,t}(\pi)' - n^{-1}\sum_{t=1}^n d^*_{\psi_k,t}(\theta)U_t(\theta).\end{aligned}\tag{B.22}$$

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.,

$$n^{-1}\sum_{t=1}^n d^*_{\psi_k,t}(\theta)U_t(\theta) = o_p(1) \quad \text{at } \theta = (\psi^{*\prime}_{k^-,n}, \pi'_{\mathcal{I}_k}, \pi'_{k^+})',\tag{B.23}$$

uniformly over $(\pi_{\mathcal{I}_k}, \pi_{k^+})$, where $\psi^*_{k^-,n}$ is between $\widehat{\psi}_{k^-}(\pi_{\mathcal{I}_k}, \pi_{k^+})$ and $\psi^0_{k^-,n}$. Given the definition of $d^*_{\psi_k,t}(\theta)$, it is sufficient to show that for $j = 1, \ldots, k-1$,

$$n^{-1}\sum_{t=1}^n[g_{\pi_j}(X_t, \pi_{\mathcal{I}_j}) + g_{\pi\pi_j}(X_t, \pi_{\mathcal{I}_j})]U_t(\theta)/\|\beta_{\mathcal{I}_j}\| = o_p(1)\tag{B.24}$$

uniformly over $(\pi_{\mathcal{I}_k}, \pi_{k^+})$ when evaluated at $\theta = (\psi^{*\prime}_{k^-,n}, \pi'_{\mathcal{I}_k}, \pi'_{k^+})'$.

Next we show (B.24) holds for $j = 1, \ldots, k-1$. For $j \le k-1$ and $\ell = k-1$, we have the following results:

$$\begin{aligned}&\frac{\|\beta_{\mathcal{I}_\ell,n}\|}{\|\beta_{\mathcal{I}_j,n}\|} = O(1) \quad \text{and}\\ &\frac{\|\beta_{\mathcal{I}_\ell,n}\|}{\widehat{\beta}_{\mathcal{I}_j,n}(\pi_{\mathcal{I}_k}, \pi_{k^+})} = \left(\frac{\widehat{\beta}_{\mathcal{I}_j,n}(\pi_{\mathcal{I}_k}, \pi_{k^+}) - \beta_{\mathcal{I}_j,n}}{\|\beta_{\mathcal{I}_\ell,n}\|} + \frac{\beta_{\mathcal{I}_j,n}}{\|\beta_{\mathcal{I}_\ell,n}\|}\right)^{-1}\\ &\qquad = O_p(1),\end{aligned}\tag{B.25}$$

because (i) the coefficients in β are grouped in a decreasing order and (ii) Lemma 2(c) applies to $\ell = k-1$. Given (B.25), we have

$$\frac{\|\beta_{\mathcal{I}_\ell,n}\|}{\beta_{\mathcal{I}_j}} = O_p(1)\tag{B.26}$$

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

$$\begin{aligned}U_t(\theta) = &\left[U_t + g_{\ell^-}(X_t, \pi_{\ell^-,n})'\beta_{\ell^-,n} + g_\ell(X_t, \pi_{\mathcal{I}_\ell,n})'\beta_{\mathcal{I}_\ell,n}\right.\\ &\left. + \ g_{\ell^+}(X_t, \pi_{\ell^+,n})\beta_{\ell^+,n}\right]\\ &- \left[g_{\ell^-}(X_t, \pi_{\ell^-})'\beta_{\ell^-} + g_\ell(X_t, \pi_{\mathcal{I}_\ell})'\beta_{\mathcal{I}_\ell}\right.\\ &\left. + \ g_{\ell^+}(X_t, \pi_{\ell^+})\beta_{\ell^+}\right],\end{aligned}\tag{B.27}$$

where the subscript ℓ^- and ℓ^+ represent the groups before and after group ℓ. Using this expansion, write

$$n^{-1}\sum_{t=1}^n g_{\pi_j}(X_t, \pi_{\mathcal{I}_j})U_t(\theta)/\left\|\beta_{\mathcal{I}_j}\right\| = \left(A_j + B_j + C_j\right)\frac{\|\beta_{\mathcal{I}_\ell,n}\|}{\left\|\beta_{\mathcal{I}_j}\right\|},\tag{B.28}$$

where $\|\beta_{\mathcal{I}_\ell,n}\|/\left\|\beta_{\mathcal{I}_j}\right\| = O_p(1)$ following (B.26) and A_j, B_j, C_j are specified as follows. The first term is

$$A_j = \frac{n^{-1/2}\sum\limits_{t=1}^n g_{\pi_j}(X_t, \pi_{\mathcal{I}_j})U_t}{n^{1/2}\|\beta_{\mathcal{I}_\ell,n}\|}.\tag{B.29}$$

The second term is

$$\begin{aligned}B_j &= n^{-1}\sum_{t=1}^n g_{\pi_j}(X_t, \pi_{\mathcal{I}_j})g_{\ell^-}(X_t, \pi_{\ell^-,n})'\frac{\beta_{\ell^-,n}}{\|\beta_{\mathcal{I}_\ell,n}\|}\\ &\quad - n^{-1}\sum_{t=1}^n g_{\pi_j}(X_t, \pi_{\mathcal{I}_j})g_{\ell^-}(X_t, \pi_{\ell^-})'\frac{\beta_{\ell^-}}{\|\beta_{\mathcal{I}_\ell,n}\|}\\ &= n^{-1}\sum_{t=1}^n g_{\pi_j}(X_t, \pi_{\mathcal{I}_j})\left(g_{\ell^-}(X_t, \pi_{\ell^-,n})\right.\\ &\quad \left. - \ g_{\ell^-}(X_t, \pi_{\ell^-})\right)'\frac{\beta_{\ell^-,n}}{\|\beta_{\mathcal{I}_\ell,n}\|}\\ &\quad - n^{-1}\sum_{t=1}^n g_{\pi_j}(X_t, \pi_{\mathcal{I}_j})g_{\ell^-}(X_t, \pi_{\ell^-})'\frac{\beta_{\ell^-} - \beta_{\ell^-,n}}{\|\beta_{\mathcal{I}_\ell,n}\|}.\end{aligned}\tag{B.30}$$

The third term is

$$\begin{aligned}C_j &= n^{-1}\sum_{t=1}^n g_{\pi_j}(X_t, \pi_{\mathcal{I}_j})g_\ell(X_t, \pi_{\mathcal{I}_\ell,n})'\frac{\beta_{\mathcal{I}_\ell,n}}{\|\beta_{\mathcal{I}_\ell,n}\|}\\ &\quad - n^{-1}\sum_{t=1}^n g_{\pi_j}(X_t, \pi_{\mathcal{I}_j})g_\ell(X_t, \pi_{\mathcal{I}_\ell})'\frac{\beta_{\mathcal{I}_\ell}}{\|\beta_{\mathcal{I}_\ell,n}\|}\\ &\quad + n^{-1}\sum_{t=1}^n g_{\pi_j}(X_t, \pi_{\mathcal{I}_j})g_{\ell^+}(X_t, \pi_{\ell^+,n})'\frac{\beta_{\ell^+,n}}{\|\beta_{\mathcal{I}_\ell,n}\|}\\ &\quad - n^{-1}\sum_{t=1}^n g_{\pi_j}(X_t, \pi_{\mathcal{I}_j})g_{\ell^+}(X_t, \pi_{\ell^+})'\frac{\beta_{\ell^+}}{\|\beta_{\mathcal{I}_\ell,n}\|}.\end{aligned}\tag{B.31}$$

Now we show $A_j, B_j, C_j = o_p(1)$. Note that the rate of convergence in Lemma 2(c) holds when $\widehat{\psi}_{k^-}(\pi_{\mathcal{I}_k}, \pi_{k^+})$ is replaced by ψ_n. Hence, it also holds for any ψ_{k^-} between $\widehat{\psi}_{k^-}(\pi_{\mathcal{I}_k}, \pi_{k^+})$ and ψ_n. First, $A_j = o_p(1)$ because (i) $n^{-1/2}\sum_{t=1}^n g_{\pi_j}(X_t, \pi_{\mathcal{I}_j})U_t = O_p(1)$ uniformly over $\pi_{\mathcal{I}_j}$ by Lemmas A.2 and A.3 and (ii) $n^{1/2}\|\beta_{\mathcal{I}_\ell,n}\| \to \infty$. Second, $B_j = o_p(1)$ because $\|\beta_{\mathcal{I}_j,n}\|(\widehat{\pi}_{\mathcal{I}_j}(\pi_{\mathcal{I}_k}, \pi_{k^+}) - \pi_{\mathcal{I}_j,n})$ and $\widehat{\beta}_{\mathcal{I}_j}(\pi_{\mathcal{I}_k}, \pi_{k^+}) - \beta_{\mathcal{I}_j,n}$ both converge to 0 faster than $\|\beta_{\mathcal{I}_\ell,n}\|$ for $j < \ell$ by Lemma 2(c). Third, $C_j = o_p(1)$ holds because (i) for ψ_{k^-} between

$\widehat{\psi}_{k-}(\pi_{\mathcal{I}_k},\pi_{k+})$ and ψ_n, $\beta_{\mathcal{I}_\ell,n}/\|\beta_{\mathcal{I}_\ell,n}\| \to \omega_{\ell,0}$, $\beta_{\mathcal{I}_\ell}/\|\beta_{\mathcal{I}_\ell,n}\| \to \omega_{\ell,0}$, $\beta_{\ell+,n}/\|\beta_{\mathcal{I}_\ell,n}\| \to 0$, $\beta_{\ell+}/\|\beta_{\mathcal{I}_\ell,n}\| \to 0$ and (ii) the sample means are $O_p(1)$ by the ULLN in Lemma A.1. Similarly, (B.28) holds when $g_{\pi_j}(X_t,\pi_{\mathcal{I}_j})$ is replaced by $g_{\pi\pi,j}(X_t,\pi_{\mathcal{I}_j})$. This proves (B.24), which in turn implies (B.23).

It follows from (B.22) and (B.23) that, for $\theta = (\psi'_{k-},\pi'_{\mathcal{I}_k},\pi'_{k+})'$, where ψ_{k-} is between $\widehat{\psi}_{k-}(\pi_{\mathcal{I}_k},\pi_{k+})$ and $\psi^0_{k-,n}$, the normalized second order partial derivative satisfies

$$[\mathbf{B}(\beta_{k-})]^{-1}D^2_{\psi_k}(\theta)[\mathbf{B}(\beta_{k-})]^{-1} \to_p H_k(\pi_{\mathcal{I}_k},\pi_{\mathcal{I}_k}|\pi_{k+}), \quad \text{where} \tag{B.32}$$

$$H_k(\pi_{\mathcal{I}_k},\pi_k|\pi_{k+}) = \mathbb{E}_{\gamma_0} d_{\psi_k,t}(\pi_{k-,0},\pi_{\mathcal{I}_k},\pi_{k+},\omega_{k-,0}) \times d_{\psi_k,t}(\pi_{k-,0},\pi_{\mathcal{I}_k},\pi_{k+},\omega_{k-,0})'.$$

Next we show

$$[\mathbf{B}(\beta_{k-,n})]^{-1}[\mathbf{B}(\beta_{k-})] \to_p I_{d_\beta+d_\zeta+d_{k-}}, \tag{B.33}$$

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 $j = 1,\ldots,k-1$,

$$\left|\frac{\|\widehat{\beta}_{\mathcal{I}_j}(\pi_{\mathcal{I}_k},\pi_{k+})\|}{\|\beta_{\mathcal{I}_j,n}\|} - 1\right| \le \frac{\|\widehat{\beta}_{\mathcal{I}_j}(\pi_{\mathcal{I}_k},\pi_{k+}) - \beta_{\mathcal{I}_j,n}\|}{\|\beta_{\mathcal{I}_{k-1},n}\|}\frac{\|\beta_{\mathcal{I}_{k-1},n}\|}{\|\beta_{\mathcal{I}_j,n}\|} \to 0 \tag{B.34}$$

by applying Lemma 2(c) to $k-1$. This implies that for $j = 1,\ldots,k-1$, $\|\beta_{\mathcal{I}_j}\|/\|\beta_{\mathcal{I}_j,n}\| \to 1$ for any $\beta_{\mathcal{I}_j}$ between $\widehat{\beta}_{\mathcal{I}_j}(\pi_{\mathcal{I}_k},\pi_{k+})$ and $\beta_{\mathcal{I}_j,n}$, which further implies the desired result in (B.33).

Normalizing the equality in (B.19), we obtain

$$\begin{aligned}
&\mathbf{B}(\beta_{k-,n})\left(\widehat{\psi}_{k-}(\pi_{\mathcal{I}_k},\pi_{k+}) - \psi^0_{k-,n}\right)\\
&= -\left\{[\mathbf{B}(\beta_{k-,n})]^{-1}D^2_{\psi_k}(\psi^*_{k-,n},\pi_{\mathcal{I}_k},\pi_{k+})[\mathbf{B}(\beta_{k-,n})]^{-1}\right\}^{-1}\\
&\quad\times\left\{\left[\mathbf{B}(\beta_{k-,n})\right]^{-1}D^1_{\psi_k}(\psi^0_{k-,n},\pi_{\mathcal{I}_k},\pi_{k+})\right\}.
\end{aligned} \tag{B.35}$$

Applying (B.21), (B.32), and (B.33) to (B.35) yields

$$\begin{aligned}
&\|\beta_{k,n}\|^{-1}\left(\mathbf{B}(\beta_{k-,n})\left(\widehat{\psi}_{k-}(\pi_{\mathcal{I}_k},\pi_{k+}) - \psi^0_{k-,n}\right)\right)\\
&\to_p \left[H_k(\pi_{\mathcal{I}_k},\pi_{\mathcal{I}_k}|\pi_{k+})\right]^{-1}\Phi_k(\pi_{\mathcal{I}_k},\pi_{\mathcal{I}_k,0}|\pi_{k+})\omega_{k,0}\\
&= \left[H_k(\pi_{\mathcal{I}_k},\pi_{\mathcal{I}_k}|\pi_{k+})\right]^{-1}H_k(\pi_{\mathcal{I}_k},\pi_{\mathcal{I}_k,0}|\pi_{k+})\Delta_k
\end{aligned} \tag{B.36}$$

uniformly over $(\pi_{\mathcal{I}_k},\pi_{k+})$, where $\Delta_k = S_k\omega_{k,0}$ by definition.

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

$$\begin{aligned}
&Q^c_n(\pi_{\mathcal{I}_k},\pi_{k+}) - Q_n(\psi^0_{k-1,n})\\
&= D^1_{\psi_{k-}}(\psi^0_{k-,n},\pi_k,\pi_{k+})'\left(\widehat{\psi}_{k-}(\pi_{\mathcal{I}_k},\pi_{k+}) - \psi^0_{k-,n}\right)\\
&\quad+\frac{1}{2}\left(\widehat{\psi}_{k-}(\pi_{\mathcal{I}_k},\pi_{k+}) - \psi^0_{k-,n}\right)'D^2_{\psi_{k-}}(\psi^{**}_{k-,n},\pi_k,\pi_{k+})\\
&\quad\times\left(\widehat{\psi}_{k-}(\pi_{\mathcal{I}_k},\pi_{k+}) - \psi^0_{k-,n}\right)\\
&= \left(D^1_{\psi_{k-}}(\psi^0_{k-,n},\pi_k,\pi_{k+})'\left[\mathbf{B}(\beta_{k-,n})\right]^{-1}\right)\\
&\quad\times\left(\mathbf{B}(\beta_{k-,n})\left(\widehat{\psi}_{k-}(\pi_{\mathcal{I}_k},\pi_{k+}) - \psi^0_{k-,n}\right)\right)\\
&\quad+\frac{1}{2}\left(\mathbf{B}(\beta_{k-,n})\left(\widehat{\psi}_{k-}(\pi_{\mathcal{I}_k},\pi_{k+}) - \psi^0_{k-,n}\right)\right)'\\
&\quad\times\left([\mathbf{B}(\beta_{k-,n})]^{-1}D^2_{\psi_{k-}}(\psi^{**}_{k-,n},\pi_2)[\mathbf{B}(\beta_{k-,n})]^{-1}\right)\\
&\quad\times\left(\mathbf{B}(\beta_{k-,n})\left(\widehat{\psi}_{k-}(\pi_{\mathcal{I}_k},\pi_{k+}) - \psi^0_{k-,n}\right)\right)
\end{aligned} \tag{B.37}$$

for some $\psi^{**}_{k-,n}$ between $\widehat{\psi}_{k-}(\pi_{\mathcal{I}_k},\pi_{k+})$ and $\psi^0_{k-,n}$. Applying the results for the first and second order derivatives in (B.21) and (B.32) and the results for $\mathbf{B}(\beta_{k-,n})(\widehat{\psi}_{k-}(\pi_{\mathcal{I}_k},\pi_{k+}) - \psi^0_{k-,n})$ in (B.35) and (B.36), we obtain the desired result in part (a).

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

$$\begin{aligned}
&\mathbb{P}_\gamma\left(\left[g_k(X_t,\pi_{\mathcal{I}_k,0})'\omega_{k,0}\right]a\right.\\
&\quad+\left[g_{k-}(X_t,\pi_{k-,0})',g_k(X_t,\pi_{\mathcal{I}_k})',g_{k+}(X_t,\pi_{k+})',\right.\\
&\quad\left.\left.Z'_t,g_{\pi_{k-}}(X_t,\pi_{k-,0})'\right]b = 0\right)\\
&\le 1-\varepsilon
\end{aligned} \tag{B.38}$$

for $\pi_{\mathcal{I}_k} \neq \pi_{\mathcal{I}_k,0}$. The desired result in (B.38) is implied by Assumption 3 and the grouping rule. Thus, part (b) follows from part(a), the argmax CMT (Theorem 3.2.2 in van der Vaart and Wellner (1996)), and $\pi_{\mathcal{I}_k,n} \to \pi_{\mathcal{I}_k,0}$ as $n \to \infty$.

Part (c). Part (c) follows from (B.36), the consistency in part (b), and replacing $\beta^0_{\mathcal{I}_k,n}$, which is a vector of zeros, with $\beta_{\mathcal{I}_k,n}$ in the centering term.

Proof of Theorem 1. Part (a). For $k = K$, normalizing (B.20) by $n^{1/2}$ yields

$$\begin{aligned}
&n^{1/2}\left[\mathbf{B}(\beta_{K-,n})\right]^{-1}D^1_{\psi_K}(\psi^0_{K-,n},\pi_K)\\
&= -n^{-1}\sum_{t=1}^n d_{\psi_K,t}(\pi_{K-,n},\pi_K,\omega_{K-,n})g_K(X_t,\pi_{K,n})'\left(n^{1/2}\beta_{K,n}\right)\\
&\quad-n^{-1/2}\sum_{t=1}^n U_td_{\psi_K,t}(\pi_{K-,n},\pi_K,\omega_{K-,n})\\
&\Rightarrow -\left[H_K(\pi_K,\pi_{K,0})S_Kb_K + G(\pi_K)\right]
\end{aligned} \tag{B.39}$$

following Lemmas A.1–A.3 and $n^{1/2}\beta_{K,n} \to b_K$. For $k = K$, (B.32) yields

$$[\mathbf{B}(\beta_{K-})]^{-1}D^2_{\psi_K}(\theta)[\mathbf{B}(\beta_{K-})]^{-1} \to_p H_K(\pi_K,\pi_K) \tag{B.40}$$

for any $\theta = (\psi'_{K-},\pi'_K)'$ where ψ_{K-} is between $\widehat{\psi}_{K-}(\pi_K)$ and $\psi^0_{K,n}$. In addition, (B.33) gives

$$[\mathbf{B}(\beta_{K-,n})]^{-1}[\mathbf{B}(\beta_{K-})] \to_p I_{d_\beta+d_\zeta+d_{K-}}. \tag{B.41}$$

For $k = K$, normalizing (B.35) by $n^{1/2}$, we obtain

$$\begin{aligned}
&n^{1/2}\mathbf{B}(\beta_{K-,n})\left(\widehat{\psi}_{K-}(\pi_K) - \psi^0_{K-,n}\right)\\
&= -\left([\mathbf{B}(\beta_{K-,n})]^{-1}D^2_{\psi_K}(\psi^*_{K-,n},\pi_K)[\mathbf{B}(\beta_{K-,n})]^{-1}\right)^{-1}\\
&\quad\times n^{1/2}\left[\mathbf{B}(\beta_{K-,n})\right]^{-1}D^1_{\psi_K}(\psi^0_{K-,n},\pi_K)
\end{aligned} \tag{B.42}$$

for $\psi^*_{K-,n}$ between $\widehat{\psi}_{K-}(\pi_K)$ and $\psi^0_{K,n}$. Combining (B.39)–(B.42) yields

$$n^{1/2}\mathbf{B}(\beta_{K-,n})\left(\widehat{\psi}_{K-}(\pi_K) - \psi^0_{K-,n}\right) \Rightarrow \tau(\pi_K), \quad \text{where}$$

$$\tau(\pi_K) = [H_K(\pi_K,\pi_K)]^{-1}\left[H_K(\pi_K,\pi_{K,0})S_Kb_K + G(\pi_K)\right]. \tag{B.43}$$

Applying (B.37) to $k = K$ and normalizing the criterion function by n, we obtain

$$\begin{aligned} & n\left(Q_n^c(\pi_K) - Q_n(\psi_{K-,n}^0)\right) \\ & = \left(n^{1/2} D_{\psi_K}^1(\psi_{K-,n}^0, \pi_K)' \left[\mathbf{B}(\beta_{K-,n})\right]^{-1}\right) \\ & \quad \times \left(n^{1/2}\mathbf{B}(\beta_{K-,n}) \left(\widehat{\psi}_{K-}(\pi_K) - \psi_{K-,n}^0\right)\right) \\ & \quad + \frac{1}{2}\left(n^{1/2}\mathbf{B}(\beta_{K-,n}) \left(\widehat{\psi}_{K-}(\pi_K) - \psi_{K-,n}^0\right)\right)' \\ & \quad \times \left([\mathbf{B}(\beta_{K-,n})]^{-1} D_{\psi_K}^2(\psi_{K-,n}^{**}, \pi_K)[\mathbf{B}(\beta_{K-,n})]^{-1}\right) \\ & \quad \times \left(n^{1/2}\mathbf{B}(\beta_{K-,n}) \left(\widehat{\psi}_{K-}(\pi_K) - \psi_{K-,n}^0\right)\right) \qquad \text{(B.44)} \\ & \Rightarrow -\frac{1}{2}\left[H_K(\pi_K, \pi_{K,0})S_K b_K + G_K(\pi_K)\right]' \left[H_K(\pi_K, \pi_K)\right]^{-1} \\ & \quad \times \left[H_K(\pi_K, \pi_{K,0})S_K b_K + G_K(\pi_K)\right] \end{aligned}$$

following (B.39), (B.40), and (B.43). Because $\widehat{\pi}_K$ minimizes $Q_n^c(\pi_K)$, applying the argmax CMT, we obtain

$$\widehat{\pi}_K \Rightarrow \pi_K^*. \qquad \text{(B.45)}$$

Because $\widehat{\psi}_{K-}(\widehat{\pi}_K) = \widehat{\psi}_{K-}$, the CMT and (B.43) yield

$$\begin{aligned} & n^{1/2}\mathbf{B}(\beta_{K-,n}) \left(\widehat{\psi}_{K-} - \psi_{K-,n}\right) \\ & = n^{1/2}\mathbf{B}(\beta_{K-,n}) \left(\widehat{\psi}_{K-}(\widehat{\pi}_K) - \psi_{K-,n}^0\right) \\ & \quad - n^{1/2}\mathbf{B}(\beta_{K-,n}) \left(\psi_{K-,n} - \psi_{K-,n}^0\right) \\ & \Rightarrow \tau_K(\pi_K^*) - S_K b_K, \qquad \text{(B.46)} \end{aligned}$$

where $S_K b_K$ is a vector of the same size as ψ_{K-} but with the subvector of β_K replaced by b_K and the rest replaced by zeros. The convergence in (B.45) and (B.46) hold jointly because there are both functionals of the same underlying stochastic processes. This completes the proof. □

Part (b). When $\|n^{1/2}\beta_{K,n}\| \to \infty$, 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

$$\begin{aligned} & D_\theta^1(\theta) = -n^{-1}\sum_{t=1}^n \mathbf{B}(\beta) d_{\theta,t}(\pi, \omega) U_t(\theta), \quad \text{with} \\ & \mathbf{B}(\beta) = diag\{(1_{d_\beta + d_\zeta}, 1_{d_{\pi_{\jmath_1}}}\|\beta_{\jmath_1}\|, \ldots, 1_{d_{\pi_{\jmath K}}}\|\beta_{\jmath_K}\|)'\}, \\ & d_{\theta,t}(\pi, \omega) \\ & \quad = (g(X_t, \pi)', Z_t', \omega_1' g_{\pi_1}(X_t, \pi_{\jmath_1}), \ldots, \omega_K' g_{\pi_K}(X_t, \pi_{\jmath_K}))' \qquad \text{(B.47)} \end{aligned}$$

and

$$\begin{aligned} D_\theta^2(\theta) \; = \; & \mathbf{B}(\beta)\left(n^{-1}\sum_{t=1}^n d_{\theta,t}(\pi, \omega) d_{\theta,t}(\pi, \omega)' \right. \\ & \left. - \; n^{-1}\sum_{t=1}^n U_t(\theta) d_{\theta,t}^*(\theta)\right)\mathbf{B}(\beta), \quad \text{where} \qquad \text{(B.48)} \end{aligned}$$

$$d_{\theta,t}^*(\theta) = \begin{pmatrix} 0_{d_\beta \times d_\beta} & 0_{d_\beta \times d_\zeta} & \delta_K^\pi(X_t, \theta) \\ 0_{d_\zeta \times d_\beta} & 0_{d_\zeta \times d_\zeta} & 0_{d_\zeta \times d_\pi} \\ \delta_K^\pi(X_t, \theta)' & 0_{d_\pi \times d_\zeta} & \delta_K^{\pi\pi}(X_t, \theta) \end{pmatrix}$$

and $\delta_K^\pi(X_t, \theta)$ and $\delta_K^{\pi\pi}(X_t, \theta)$ follow the definitions in (B.15) and (B.16).

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

$$\widehat{\theta} - \theta_n = -\left[D_\theta^2(\theta^*)\right]^{-1} D_\theta^1(\theta_n) \qquad \text{(B.49)}$$

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

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

$$n^{1/2}\left[\mathbf{B}(\beta_n)\right]^{-1} D_\theta^1(\theta_n) \to_d N(0, \Omega_\theta(\pi_0, \omega_0)). \qquad \text{(B.50)}$$

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

$$\begin{aligned} [\mathbf{B}(\beta)]^{-1} D_\theta^2(\theta)[\mathbf{B}(\beta)]^{-1} \; = \; & n^{-1}\sum_{t=1}^n d_{\theta,t}(\pi, \omega) d_{\theta,t}(\pi, \omega)' \\ & - n^{-1}\sum_{t=1}^n U_t(\theta) d_{\theta,t}^*(\theta), \qquad \text{(B.51)} \end{aligned}$$

where we have

$$n^{-1}\sum_{t=1}^n U_t(\theta) d_{\theta,t}^*(\theta) = o_p(1) \quad \text{at } \theta = \theta^*, \qquad \text{(B.52)}$$

for any θ^* between $\widehat{\theta}$ and θ_n following the arguments used to show (B.23). It follows that

$$[\mathbf{B}(\beta)]^{-1} D_\theta^2(\theta)[\mathbf{B}(\beta)]^{-1} \to_p H(\pi_0, \omega_0) \qquad \text{(B.53)}$$

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

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

$$\begin{aligned} & n^{1/2}\mathbf{B}(\beta_n)\left(\widehat{\theta} - \theta_n\right) \\ & \quad = -\left([\mathbf{B}(\beta_n)]^{-1} D_\theta^2(\theta^*)[\mathbf{B}(\beta_n)]^{-1}\right)^{-1} n^{1/2}\left[\mathbf{B}(\beta_n)\right]^{-1} D_\theta^1(\theta_n) \\ & \quad \to_d N(0, H(\pi_0, \omega_0)^{-1}\Omega_\theta(\pi_0, \omega_0) H(\pi_0, \omega_0)^{-1}). \qquad \text{(B.54)} \end{aligned}$$

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

$$\begin{aligned} W_n(R) \; = \; & n\left[R\left(\widehat{\theta} - \theta_n\right)\right]' \left[R\mathbf{B}^{-1}(\widehat{\beta})\widehat{\Sigma}_n\mathbf{B}^{-1}(\widehat{\beta})R'\right]^{-1} \\ & \times \left[R\left(\widehat{\theta} - \theta_n\right)\right]. \qquad \text{(B.55)} \end{aligned}$$

We first show

$$\varepsilon_n = W_n(R) - W_n(R^*) = o_p(1). \qquad \text{(B.56)}$$

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

$$A'R = R^* + \varepsilon_R^*, \qquad \text{(B.57)}$$

where $\varepsilon_R^* = A'R - R^*$ is defined implicitly. Using this decomposition, we have

$$\begin{aligned} & W_n(\mathbf{D}^*(\widehat{\beta})A'R) = \overline{\varrho}'\left[\overline{R}\widehat{\Sigma}_n\overline{R}'\right]^{-1}\overline{\varrho}, \quad \text{where} \\ & \overline{\rho} = n^{1/2}\mathbf{D}^*(\widehat{\beta})\left(R^* + \varepsilon_R^*\right)\left(\widehat{\theta}_n - \theta_n\right) \\ & \overline{R} = \mathbf{D}^*(\widehat{\beta})\left(R^* + \varepsilon_R^*\right)\mathbf{B}^{-1}(\widehat{\beta}). \qquad \text{(B.58)} \end{aligned}$$

Following the definition of $R^\dagger(\beta)$ in (4.14),

$$\overline{R} = R^\dagger(\widehat{\beta}) + \mathbf{D}^*(\widehat{\beta})\varepsilon_R^*\mathbf{B}^{-1}(\widehat{\beta}), \qquad \text{(B.59)}$$

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

$$\begin{aligned} & \overline{\rho} = \rho_n + n^{1/2}\mathbf{D}^*(\widehat{\beta})\varepsilon_R^*(\widehat{\theta}_n - \theta_n), \quad \text{where} \\ & \rho_n = n^{1/2}\mathbf{D}^*(\widehat{\beta})R^*(\widehat{\theta}_n - \theta_n). \qquad \text{(B.60)} \end{aligned}$$

The second term $n^{1/2}\mathbf{D}^*(\widehat{\beta})\varepsilon_R^*(\widehat{\theta}_n-\theta_n)=o_p(1)$ because its components are $n^{1/2}\|\widehat{\beta}_k\|\left(A_k'R_j\right)(\widehat{\pi}_{\mathcal{I}_j}-\pi_{\mathcal{I}_j,n})$ for $j<k$. By Theorem 1, the convergence rate of $\widehat{\pi}_{\mathcal{I}_j}$ is $n^{1/2}\|\beta_{\mathcal{I}_j,n}\|$, which is an order of magnitude larger than $n^{1/2}\|\widehat{\beta}_{\mathcal{I}_k}\|$ for $j<k$. Putting together (B.58)–(B.60), we have

$$\begin{aligned} W_n(R) &= W_n(\mathbf{D}^*(\widehat{\beta})A'R) \\ &= \left(\rho_n+o_p(1)\right)'\left[\left(R^\dagger(\widehat{\beta})+o_p(1)\right)\widehat{\Sigma}_n\left(R^\dagger(\widehat{\beta})+o_p(1)\right)'\right]^{-1} \\ &\quad\times\left(\rho_n+o_p(1)\right). \\ &= \rho_n'V_n^{-1}\rho_n+\varepsilon_n \\ &= W_n(R^*)+\varepsilon_n, \end{aligned} \tag{B.61}$$

where ε_n is implicitly defined by the third equality. Comparing the second and the third lines of (B.61), $\varepsilon_n=o_p(1)$ provided that (i) $\rho_n=O_p(1)$, (ii) $\widehat{\Sigma}_n=O_p(1)$, and (iii) $\lambda_{\min}(\widehat{\Sigma}_n^{-1})>0$ w.p.a.1., given that $R^\dagger(\widehat{\beta})$ has full rank by construction. We investigate these terms below.

We first consider weak identification in part (a). Following (4.16), $\rho_n=R^\dagger(\widehat{\beta})\xi_n$, 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 $\pi_{\mathcal{I}_K}$:

$$\xi_n(\pi_{\mathcal{I}_K})=\begin{pmatrix} n^{1/2}\mathbf{B}(\widehat{\beta}_{K-,n}(\pi_{\mathcal{I}_K}))\left(\widehat{\psi}_{K-}(\pi_{\mathcal{I}_K})-\psi_{K-,n}\right) \\ n^{1/2}\left\|\widehat{\beta}_{\mathcal{I}_K}(\pi_{\mathcal{I}_K})\right\|\left(\pi_{\mathcal{I}_K}-\pi_{\mathcal{I}_K,n}\right)\end{pmatrix}. \tag{B.62}$$

Applying (B.33) with $k=K$, we have $\mathbf{B}(\widehat{\beta}_{K-,n}(\pi_{\mathcal{I}_K}))[\mathbf{B}(\beta_{K-,n})]^{-1}=I_{d_{K-}}+o_p(1)$. Applying it together with Theorem 1(a) and the CMT yields

$$\xi_n=\xi_n(\widehat{\pi}_{\mathcal{I}_K})\Rightarrow\xi(\pi^*_{\mathcal{I}_K}), \tag{B.63}$$

where

$$\xi(\pi_{\mathcal{I}_K})=\begin{pmatrix}\tau_K(\pi_{\mathcal{I}_K})-S_Kb_K \\ \left\|\tau_{\beta_K}(\pi_{\mathcal{I}_K})\right\|\left(\pi_{\mathcal{I}_K}-\pi_{\mathcal{I}_K,0}\right)\end{pmatrix}. \tag{B.64}$$

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

$$\begin{aligned}\widehat{\omega}_k &= \frac{\widehat{\beta}_{\mathcal{I}_k}}{\|\widehat{\beta}_{\mathcal{I}_k}\|} \\ &= \frac{\widehat{\beta}_{\mathcal{I}_k}-\beta_{\mathcal{I}_k,n}}{\|\beta_{\mathcal{I}_k,n}\|}\frac{\|\beta_{\mathcal{I}_k,n}\|}{\|\widehat{\beta}_{\mathcal{I}_k}\|}+\frac{\beta_{\mathcal{I}_k,n}}{\|\beta_{\mathcal{I}_k,n}\|}\frac{\|\beta_{\mathcal{I}_k,n}\|}{\|\widehat{\beta}_{\mathcal{I}_k}\|}\to_p\omega_{k,0}.\end{aligned} \tag{B.65}$$

For the last group,

$$\widehat{\omega}_K=n^{1/2}\widehat{\beta}_{\mathcal{I}_K}/\|n^{1/2}\widehat{\beta}_{\mathcal{I}_k}\|\Rightarrow\frac{\tau_{\beta_K}(\pi^*_{\mathcal{I}_K})}{\|\tau_{\beta_K}(\pi^*_{\mathcal{I}_K})\|} \tag{B.66}$$

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

$$\widehat{\omega}\Rightarrow\omega(\pi^*_{\mathcal{I}_K})=\left(\omega_{1,0},\ldots,\omega_{K-1,0},\frac{\tau_{\beta_K}(\pi^*_{\mathcal{I}_K})}{\|\tau_{\beta_K}(\pi^*_{\mathcal{I}_K})\|}\right). \tag{B.67}$$

The covariance matrix is $\widehat{\Sigma}=[\widehat{H}(\widehat{\pi},\widehat{\omega})]^{-1}\widehat{\Omega}_\theta(\widehat{\theta})[\widehat{H}(\widehat{\pi},\widehat{\omega})]^{-1}$. Define

$$\widehat{H}(\pi,\omega)=n^{-1}\sum_{t=1}^n d_{\theta,t}(\pi,\omega)d_{\theta,t}(\pi,\omega)'. \tag{B.68}$$

Lemma A.1 implies that

$$\widehat{H}(\pi,\omega)\to_p H(\pi,\omega) \tag{B.69}$$

uniformly over (π,ω), which implies that

$$\widehat{H}\Rightarrow H(\pi_{K-,0},\pi^*_{\mathcal{I}_K},\omega(\pi^*_{\mathcal{I}_K})). \tag{B.70}$$

For the other term, we have

$$\begin{aligned}\widehat{\Omega}_\theta &= n^{-1}\sum_{t=1}^n\widehat{U}_t^2 d_{\theta,t}(\widehat{\pi},\widehat{\omega})d_{\theta,t}(\widehat{\pi},\widehat{\omega})' \\ &= n^{-1}\sum_{t=1}^n U_t^2 d_{\theta,t}(\widehat{\pi},\widehat{\omega})d_{\theta,t}(\widehat{\pi},\widehat{\omega})' \\ &\quad+2n^{-1}\sum_{t=1}^n U_t\left(\sum_{k=1}^K\left(g_{\mathcal{I}_k}(X_t,\pi_{\mathcal{I}_k,n})'\beta_{\mathcal{I}_k,n}\right.\right. \\ &\quad\left.\left.-\ g_{\mathcal{I}_k}(X_t,\widehat{\pi}_{\mathcal{I}_k})'\widehat{\beta}_{\mathcal{I}_k}\right)\right)d_{\theta,t}(\widehat{\pi},\widehat{\omega})d_{\theta,t}(\widehat{\pi},\widehat{\omega})' \\ &\quad+n^{-1}\sum_{t=1}^n\left(\sum_{k=1}^K\left(g_{\mathcal{I}_k}(X_t,\pi_{\mathcal{I}_k,n})'\beta_{\mathcal{I}_k,n}\right.\right. \\ &\quad\left.\left.-\ g_{\mathcal{I}_k}(X_t,\widehat{\pi}_{\mathcal{I}_k})'\widehat{\beta}_{\mathcal{I}_k}\right)\right)^2 d_{\theta,t}(\widehat{\pi},\widehat{\omega})d_{\theta,t}(\widehat{\pi},\widehat{\omega})' \\ &\Rightarrow\Omega_\theta(\pi_{K-,0},\pi^*_{\mathcal{I}_K},\omega(\pi^*_{\mathcal{I}_K})),\quad\text{where }\Omega_\theta(\pi,\omega) \\ &= \mathbb{E}_{\gamma_0}\left[U_t^2 d_{\theta,t}(\pi,\omega)d_{\theta,t}(\pi,\omega)'\right].\end{aligned} \tag{B.71}$$

Thus,

$$\widehat{\Sigma}\Rightarrow\Sigma(\pi_{K-,0},\pi^*_{\mathcal{I}_K},\omega(\pi^*_{\mathcal{I}_K}))=\Sigma\left(\pi^*_{\mathcal{I}_K}\right). \tag{B.72}$$

Putting together (B.63) and (B.72), we obtain $\varepsilon_R=o_p(1)$ by (B.61). Furthermore, these results hold jointly. Therefore,

$$\begin{aligned}W_n(R) &= \rho_n'V_n^{-1}\rho_n+o_p(1) \\ &= \left(R^\dagger(\widehat{\beta})\xi_n\right)'\left[R^\dagger(\widehat{\beta})\widehat{\Sigma}R^\dagger(\widehat{\beta})'\right]^{-1}\left(R^\dagger(\widehat{\beta})\xi_n\right)+o_p(1) \\ &\Rightarrow\left(R^\dagger(\beta_0)\xi(\pi^*_{\mathcal{I}_K})\right)'\left[R^\dagger(\beta_0)\Sigma(\pi^*_{\mathcal{I}_K})R^\dagger(\beta_0)'\right]^{-1} \\ &\quad\times\left(R^\dagger(\beta_0)\xi(\pi^*_{\mathcal{I}_K})\right),\end{aligned} \tag{B.73}$$

where the first equality follows from (B.61) 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.63) and (B.72).

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

$$\xi_n(\widehat{\pi}_K)\to_d\xi\sim N(0,\,\Sigma(\pi_0,\omega_0)) \tag{B.74}$$

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

$$\widehat{\omega}\to_p\omega_0=(\omega_{1,0}',\ldots,\omega_{K,0}')'\quad\text{and}\quad\widehat{\Sigma}\to_p\Sigma(\pi_0,\omega_0) \tag{B.75}$$

following the arguments in (B.65) for $k=K$ and the consistency of $\widehat{\pi}_{\mathcal{I}_K}$ in this case. Therefore, $\varepsilon_n=o_p(1)$ following the calculation in (B.61). Furthermore, the Wald statistic satisfies

$$\begin{aligned}W_n(R) &\to_d \left[R^\dagger(\beta_0)\xi\right]'\left[R^\dagger(\beta_0)\Sigma_0R^\dagger(\beta_0)'\right]^{-1}\left[R^\dagger(\beta_0)\xi\right] \\ &\sim\ \chi^2_{d_r}\end{aligned} \tag{B.76}$$

because $R^\dagger(\beta_0)$ and Σ_0 both have full rank. This completes the proof. □

Corollary 1. *follows directly from Theorem 2.*

Proof to show (4.25). When testing the null hypothesis $H_0 : R\theta_n = v_n^{null}$, the Wald statistic $W_n(R)$ can be written as

$$\begin{aligned} W_n(R) &= n\left[R\left(\widehat{\theta} - \theta_n\right) + \left(R\theta_n - v_n^{null}\right)\right]' \\ &\quad \times \left[R\mathbf{B}^{-1}(\widehat{\beta})\widehat{\Sigma}_n\mathbf{B}^{-1}(\widehat{\beta})R'\right]^{-1} \\ &\quad \times \left[R\left(\widehat{\theta} - \theta_n\right) + \left(R\theta_n - v_n^{null}\right)\right]. \end{aligned} \tag{B.77}$$

Because $\mathbf{D}^*(\widehat{\beta})A'$ is a full rank matrix w.p.a.1, we have

$$\begin{aligned} W_n(R) &= n\left[\mathbf{D}^*(\widehat{\beta})A'R\left(\widehat{\theta} - \theta_n\right) + \mathbf{D}^*(\widehat{\beta})A'\left(R\theta_n - v_n^{null}\right)\right]' \\ &\quad \times \left[\mathbf{D}^*(\widehat{\beta})A'R\mathbf{B}^{-1}(\widehat{\beta})\widehat{\Sigma}_n\mathbf{B}^{-1}(\widehat{\beta})R'A\mathbf{D}^*(\widehat{\beta})'\right]^{-1} \\ &\quad \times \left[\mathbf{D}^*(\widehat{\beta})A'R\left(\widehat{\theta} - \theta_n\right) + \mathbf{D}^*(\widehat{\beta})A'\left(R\theta_n - v_n^{null}\right)\right] \\ &= (\overline{\rho} + \Delta_n)'\left[\overline{R}\widehat{\Sigma}_n\overline{R}'\right]^{-1}(\overline{\rho} + \Delta_n), \end{aligned} \tag{B.78}$$

where

$$\Delta_n = n^{1/2}\mathbf{D}^*(\widehat{\beta})A'\left(R\theta_n - v_n^{null}\right) \tag{B.79}$$

and $\overline{\rho}$ and $\overline{R}$ are defined in (B.58). In the proof of Theorem 2, we have shown

$$\begin{aligned} &\overline{\rho} = \rho_n + o_p(1) = R^{\dagger}(\widehat{\beta})\xi_n + o_p(1), \\ &\quad \text{where } \xi_n = n^{1/2}\mathbf{B}(\widehat{\beta})(\widehat{\theta} - \theta_n) \end{aligned} \tag{B.80}$$

and

$$\overline{R} = R^{\dagger}(\widehat{\beta}) + o_p(1) \tag{B.81}$$

in all identification scenarios. In addition, $\rho_n = O_p(1)$ and $\overline{R} = O_p(1)$ in all cases. This shows the results in (4.25). □

Appendix C. Proofs for the asymptotic size

Proof of Theorem 3. In the original model in (1.1), the DGP is determined by $\beta = (\beta_1', \ldots, \beta_p')'$, ζ, $\pi = (\pi_1', \ldots, \pi_p')'$, and ϕ. Because the identification strength of π_j is determined by $\|\beta_j\|$, we parameterize β_j as $(\|\beta_j\|, \sigma_j)$, where

$$\sigma_j = \beta_j/\|\beta_j\|. \tag{C.1}$$

Without loss of generality, define $\sigma_j = 1_{d_{\beta_j}}$ if $\beta_j = 0$. Note that this angle parameter σ_j is different from ω_j defined above. The former is based on the original parameterization β_j whereas the latter is based on the grouping result $\beta_{\mathcal{I}_j}$.

The DGP is determined by

$$\lambda = (\|\beta_1\|, \ldots, \|\beta_p\|, \sigma_1', \ldots, \sigma_p', \zeta', \pi', \phi) \in \Lambda. \tag{C.2}$$

Define a function

$$\begin{aligned} &h_n(\lambda_n) \\ &\quad = \left(n^{1/2}\|\beta_{1,n}\|, \ldots, n^{1/2}\|\beta_{p,n}\|, g(\|\beta_{1,n}\|, \ldots, \|\beta_{p,n}\|), \lambda_n\right), \end{aligned} \tag{C.3}$$

where

$$\begin{aligned} g(\|\beta_1\|, \ldots, \|\beta_p\|) &= \left(\frac{\|\beta_j\|}{\|\beta_\ell\|}\right)_{j \neq \ell} \\ &= \left(\frac{\|\beta_1\|}{\|\beta_2\|}, \ldots, \frac{\|\beta_1\|}{\|\beta_p\|}, \ldots, \frac{\|\beta_p\|}{\|\beta_1\|}, \ldots, \frac{\|\beta_p\|}{\|\beta_{p-1}\|}\right) \end{aligned} \tag{C.4}$$

where $\|\beta_j\|/\|\beta_\ell\| \in \mathbb{R}_+ \cup \{\infty\}$ and, by definition, $\|\beta_j\|/\|\beta_\ell\| = \infty$ if $\beta_\ell = 0$. In (C.3), $g(\|\beta_{1,n}\|, \ldots, \|\beta_{p,n}\|)$ determines the relative convergence rate, which is needed to specify the grouping result $\mathcal{I}$.

Recall that in (5.3), we define

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

It is a one-to-one transformation between h and the limit of $h_n(\lambda_n)$ because $(\mathcal{I}, b_{\mathcal{I}_K}, \omega_0)$ determines the limit of $n^{1/2}\|\beta_{1,n}\|, \ldots, n^{1/2}\|\beta_{p,n}\|$ and $g(\|\beta_{1,n}\|, \ldots, \|\beta_{p,n}\|)$, and vice versa.

For any sequences of true parameters $\{\lambda_n : n \geq 1\}$ for which the limit of $h_n(\lambda_n)$ can be parameterized as $h \in H$, Theorem 2 shows that $W_n(R) \rightarrow_d \mathcal{W}(h)$. In other words, the limit distribution is index by h. (For convenience, if $h_n(\lambda_n)$ converges to a limit that can be reparameterized as $h \in H$, below we also say $h_n(\lambda_n) \rightarrow h$.) Under Assumption CV1, $\mathcal{W}(h)$ is continuous at $\chi^2_{d_r,1-\alpha} \forall h \in H$. Therefore, the coverage probability satisfies

$$\begin{aligned} CP_n(\lambda_n) &= \Pr(W_n(R) \leq \chi^2_{d_r,1-\alpha}) \rightarrow \Pr(\mathcal{W}(h) \leq \chi^2_{d_r,1-\alpha}) \\ &= CP(h). \end{aligned} \tag{C.6}$$

The generic results in ACG provide a link between the pointwise results in (C.6) along $\{\lambda_n : n \geq 1\}$ and the uniform results for the asymptotic size in (5.1) and (5.2). Take the asymptotic size for a confidence set for an example. The arguments in ACG roughly go as follows. By the definition of inf and liminf, $\liminf_{n\rightarrow\infty} \inf_{\lambda\in\Lambda} CP_n(\lambda) = \lim_{n\rightarrow\infty} CP_{p_n}(\lambda_{p_n})$ for some subsequence $\{p_n\}$ of $\{n\}$. Theorems 2.1 and 2.2 of ACG prove that, for a confidence set, $AsySz = \inf_{h\in H} CP(h)$ if "For any subsequence $\{p_n\}$ of $\{n\}$ and any sequence $\{\lambda_{p_n} \in \Lambda : n \geq 1\}$ for which $h_{p_n}(\lambda_{p_n}) \rightarrow h \in H$, $CP_{p_n}(\lambda_{p_n}) \rightarrow CP(h)$ for some $CP(h) \in [0, 1]$." (This is Assumption B of ACG with $CP(h) = CP^-(h) = CP^+(h)$.) This statement is analogous to (C.6), except that (C.6) is established for the full sequences $\{\lambda_n : n \geq 1\}$, rather than for the subsequences $\{\lambda_{p_n}\}$. The full sequence result in (C.6) verifies Assumptions B1 and C1 in ACG. Lemma 2.1 of ACG shows that a missing link between the subsequence result and the full sequence result is Assumption B2 of ACG, which states "For any subsequence $\{p_n\}$ of $\{n\}$ and any sequence $\{\lambda_{p_n} : n \geq 1\}$ for which $h_{p_n}(\lambda_{p_n}) \rightarrow h \in H$, there exists a sequence $\{\lambda_n^* \in \Lambda : n \geq 1\}$ such that $h_n(\lambda_n^*\} \rightarrow h \in H$ and $\lambda^*_{p_n} = \lambda_{p_n}, \forall n \geq 1$." In other words, Assumption B2 of ACG ensures that the set of subsequence limits along $h_{p_n}(\lambda_{p_n})$ is the same as the set of full sequence limits along $h_n(\lambda_n)$, the latter of which is given in (C.6). Therefore, it remains to verify Assumption B2 of ACG to complete the proof.

Now we verify Assumption B2 of ACG, "For any subsequence $\{p_n\}$ of $\{n\}$ and any sequence $\{\lambda_{p_n} : n \geq 1\}$ for which $h_{p_n}(\lambda_{p_n}) \rightarrow h \in H$, there exists a sequence $\{\lambda_n^* \in \Lambda : n \geq 1\}$ such that $h_n(\lambda_n^*\} \rightarrow h \in H$ and $\lambda^*_{p_n} = \lambda_{p_n}, \forall n \geq 1$." To be clear with the notation, let us call this new full sequence $\{\lambda_k^* : k \geq 1\}$. We aim to construct a full sequence $\{\lambda_k^*\}$ which is the same as λ_{p_n} for $k = p_n$ and $h_k(\lambda_k^*) \rightarrow h$ as $k \rightarrow \infty$. The question is how to fill in the sequence for $k \neq p_n$ for any n. This new sequence $\{\lambda_k^* = (\|\beta^*_{1,k}\|, \ldots, \|\beta^*_{p,k}\|, \sigma^{*\prime}_{1,k}, \ldots, \sigma^{*\prime}_{p,k}, \zeta_k^{*\prime}, \pi_k^{*\prime}, \phi_k^*) : k \geq 1\}$ is defined as follows: (i) $\forall k = p_n$ define $\lambda_k^* = \lambda_{p_n} \in \Lambda$, and (ii) $\forall k \in (p_n, p_{n+1})$, define

$$\begin{aligned} &\|\beta^*_{j,k}\| = \begin{cases} \dfrac{\sqrt{p_n}\|\beta_{j,p_n}\|}{\sqrt{k}} & \text{if (a) } \sqrt{p_n}\|\beta_{j,p_n}\| \rightarrow h_{1,j} \in \mathbb{R} \\ \|\beta_{j,p_n}\| & \text{if (b) } \sqrt{p_n}\|\beta_{j,p_n}\| \rightarrow \infty. \end{cases} \\ &\quad \text{for } j = 1, \ldots, p, \\ &\sigma^*_{j,k} = \sigma_{j,p_n} \\ &\quad \text{for } j = 1, \ldots, p,\ \zeta_k^* = \zeta_{p_n},\ \pi_k^* = \pi_{p_n},\ \phi_k^* = \phi_{p_n}. \end{aligned} \tag{C.7}$$

For k between p_n and p_{n+1}, $\|\beta_{j,k}\|$ for the weak identification group is constructed in a way such that the limit of $\sqrt{k}\|\beta_{j,k}\|$ is the same as $\sqrt{p_n}\|\beta_{j,p_n}\|$. (Note that p denotes the number of nonlinear regressors in the original model, whereas p_n indexes the subsequence to be consistent with the notation in ACG.) For k large enough, we can always construct λ_k^* as proposed because the parameter space is a product space that contains a neighborhood of β arbitrarily close to 0.

It remains to show that $h_k(\lambda_k^*) \rightarrow h$ if $h_{p_n}(\lambda_{p_n}) \rightarrow h$. (i) It is clear that $\sqrt{k}\|\beta_{j,k}^*\|$ and $\sqrt{p_n}\|\beta_{j,p_n}\|$ have the same limit by construction in (C.7). (ii) λ_k^* and λ_{p_n} also have the same limit by construction in (C.7) because $\|\beta_{j,k}^*\|$ converges to 0 if and only if $\|\beta_{j,p_n}\|$ converges to 0. (iii) We now show the limits of $g(\|\beta_{1,k}^*\|, \ldots, \|\beta_{p,k}^*\|)$ and $g(\|\beta_{1,p_n}\|, \ldots, \|\beta_{p,p_n}\|)$ are the same by showing $\|\beta_{j,k}^*\|/\|\beta_{\ell,k}^*\|$ and $\|\beta_{j,p_n}\|/\|\beta_{\ell,p_n}\|$ have the same limits for all $\ell \neq j$. In (C.7), we have cases (a) and (b) for $\|\beta_{j,p_n}\|$ depending on its rate of convergence. Now we discuss three cases. (1) If $\|\beta_{j,p_n}\|$ and $\|\beta_{\ell,p_n}\|$ are both in case (a) or both in case (b), $\|\beta_{j,k}^*\|/\|\beta_{\ell,k}^*\| = \|\beta_{j,p_n}\|/\|\beta_{\ell,p_n}\|$. (2) If $\|\beta_{j,p_n}\|$ is in case (a) and $\|\beta_{\ell,p_n}\|$ is in case (b), we have $\|\beta_{j,p_n}\|/\|\beta_{\ell,p_n}\| \rightarrow 0$. Then, $\|\beta_{j,k}^*\|/\|\beta_{\ell,k}^*\| = (\sqrt{p_n}/\sqrt{k})\|\beta_{j,p_n}\|/\|\beta_{\ell,p_n}\| \rightarrow 0$. (3) If $\|\beta_{j,p_n}\|$ is in case (b) and $\|\beta_{\ell,p_n}\|$ is in case (a), we have $\|\beta_{j,p_n}\|/\|\beta_{\ell,p_n}\| \rightarrow \infty$. Then, $\|\beta_{j,k}^*\|/\|\beta_{\ell,k}^*\| = (\sqrt{k}/\sqrt{p_n})\|\beta_{j,p_n}\|/\|\beta_{\ell,p_n}\| \rightarrow \infty$. This shows that $h_k(\lambda_k^*)$ and $h_{p_n}(\lambda_{p_n})$ have the same limit as desired, which in turn verifies Assumption B2 in ACG. As explained above, the results in (C.6) verifies Assumptions B1 and C1 in ACG. The desired result follows directly from Lemma 2.1, Theorem 2.2, and Theorem 2.1(c) of ACG. □

Proof of Theorem 4. We first introduce some notations. For a sequence of constants $\{c_n : n \geq 1\}$, let $c_n \rightarrow [c_1, c_2]$ denote $c_1 \leq \liminf_{n\rightarrow\infty} c_n \leq \limsup_{n\rightarrow\infty} c_n \leq c_2$.

Below we show (i) the pointwise convergence result in Assumption B1 of ACG hold for the robust test "For any sequence $\{h_n(\lambda_n) \rightarrow h \in H, CP(\lambda_n) \rightarrow [CP^-(h), CP^+(h)] \in [0, 1]\}$" and (ii) the lower bound is achieved as in Assumption C1 of ACG "$CP^-(h_L) = CP^+(h_L)$ for some $h_L \in H$ such that $CP^-(h_L) = \inf_{h\in H} CP^-(h)$." As in the proof of Theorem 3, we invoke Theorem 2.1(c) 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.

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

$$\Pr(W_n(R) \leq \widehat{c}_{n,1-\alpha}) \rightarrow [CP^-(h_0), CP^+(h_0)] \tag{C.8}$$

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, Theorem 2.1(c) of ACG implies that the asymptotic size is $1-\alpha$.

For a given $h_0 \in H$, its corresponding elements are $\mathcal{I}_{K,0}$, $\omega_{\mathcal{I}_k,0}$, $\pi_{\mathcal{I}_k,0}$, and γ_0. We define an infeasible critical value under h_0 as

$$\bar{c}_{1-\alpha}(h_0) = \sup_{h\in H_0} \mathcal{W}_{1-\alpha}(h), \quad \text{where}$$

$$H_0 = \{h \in H : \mathcal{I}_K = \mathcal{I}_{K,0},\ \omega_{\mathcal{I}_k} = \omega_{\mathcal{I}_k,0},\ \pi_{\mathcal{I}_k} = \pi_{\mathcal{I}_k,0} \text{ for } k < K\}. \tag{C.9}$$

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

$$\bar{c}_{1-\alpha}(h_0) \geq \mathcal{W}_{1-\alpha}(h_0). \tag{C.10}$$

Recall the plug-in critical value defined as

$$\widehat{c}_{n,1-\alpha} = \sup_{h\in\widehat{H}} \mathcal{W}_{1-\alpha}(h), \quad \text{where}$$

$$\widehat{H} = \{h \in H : \mathcal{I}_K = \widehat{\mathcal{I}}_W,\ \omega_{\mathcal{I}_k} = \widehat{\beta}_{\mathcal{I}_k}/\|\widehat{\beta}_{\mathcal{I}_k}\| \text{ and } \pi_{\mathcal{I}_k} = \widehat{\pi}_{\mathcal{I}_k} \text{ for } k < K\}. \tag{C.11}$$

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 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 $\widehat{c}_{n,1-\alpha} \geq \chi^2_{d_r,1-\alpha}$ by construction.

Consider $j \in \mathcal{I}_{K,0}$, Theorem 1 and (B.72) imply

$$\begin{aligned} ICS_{j,n} &= \left(n\widehat{\beta}_j'(\widehat{\Sigma}_j)^{-1}\widehat{\beta}_j/d_{\beta_j}\right)^{1/2} \\ &\rightarrow_d \left(\tau_{\beta_j}(\pi_K^*)'(\Sigma_j(\pi_K^*))^{-1}\tau_{\beta_j}(\pi_K^*)/d_{\beta_j}\right)^{1/2}, \end{aligned} \tag{C.12}$$

where $\tau_{\beta_j}(\pi_K)$ is the subvector of $\tau(\pi)$ associated with β_j and $\Sigma_j(\pi_K)$ is a submatrix of $\Sigma(\pi)$ associated with β_j, for both of which $\pi_1, \ldots, \pi_{K-1}$ 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

$$\Pr(\mathcal{I}_{K,0} \subseteq \widehat{\mathcal{I}}_W) \rightarrow 1. \tag{C.13}$$

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,

$$\bar{c}_{1-\alpha}(h_0) \leq \widehat{c}_{n,1-\alpha} + o_p(1). \tag{C.14}$$

Combining it with (C.10), w.p.a.1, we have

$$\mathcal{W}_{1-\alpha}(h_0) \leq \widehat{c}_{n,1-\alpha} + o_p(1). \tag{C.15}$$

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,

$$\begin{aligned} &\Pr\left(W_n(R) \leq \widehat{c}_{n,1-\alpha}\right) \\ &\geq \Pr(W_n(R) + o_p(1) \leq \mathcal{W}_{1-\alpha}(h_0) \ \ \& \\ &\mathcal{W}_{1-\alpha}(h_0) \leq \widehat{c}_{n,1-\alpha} + o_p(1)) \\ &= \Pr(W_n(R) + o_p(1) \leq \mathcal{W}_{1-\alpha}(h_0)) \\ &\quad - \Pr(W_n(R) + o_p(1) \leq \mathcal{W}_{1-\alpha}(h_0) \ \ \& \\ &\mathcal{W}_{1-\alpha}(h_0) > \widehat{c}_{n,1-\alpha} + o_p(1)) \\ &\geq \Pr(W_n(R) + o_p(1) \leq \mathcal{W}_{1-\alpha}(h_0)) - \Pr(\mathcal{W}_{1-\alpha}(h_0) \\ &> \widehat{c}_{n,1-\alpha} + o_p(1)) \\ &\rightarrow 1-\alpha, \end{aligned} \tag{C.16}$$

where the convergence follows from $W_n(R) \rightarrow_d \mathcal{W}(h_0)$, the Slutsky's theorem, and (C.15). Therefore, for any $h_0 \in H$, (C.8) 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 $j = 1, \ldots, p$. In this case,

$$\kappa_n^{-1} ICS_{j,n} = \left(\kappa_n^{-1}n^{1/2}\right)\left(\widehat{\beta}_j'(\widehat{\Sigma}_j)^{-1}\widehat{\beta}_j/d_{\beta}\right)^{1/2} \rightarrow \infty \tag{C.17}$$

Table C.1
Size-adjusted power ($\times 100$) for $H_0 : \beta_2 = 0$ when $\beta_{2n} = n^{-1/2}b_2$, $\beta_{1n} = n^{-1/2}b_1$.

	$b_2 = 0$	$b_2 = 1$	$b_2 = 2$	$b_2 = 3$	$b_2 = 4$	$b_2 = 6$	$b_2 = 8$	$b_2 = 10$
b_1	Robust							
0	5.4	10.7	26.3	50.1	73.3	97.2	99.9	100.0
1	5.2	9.5	23.7	46.2	70.9	97.2	100.0	100.0
2	4.8	7.9	21.0	44.3	70.4	97.3	100.0	100.0
3	4.5	7.1	20.3	44.6	71.6	97.6	100.0	100.0
4	4.6	7.2	21.4	46.5	73.7	97.9	100.0	100.0
6	5.0	9.1	25.4	51.5	77.1	98.3	100.0	100.0
8	5.4	10.2	27.1	53.0	77.7	98.3	100.0	100.0
10	5.5	10.4	27.2	53.0	77.8	98.3	100.0	100.0
b_1	Standard							
0	5.4	13.8	35.0	57.8	74.3	84.9	85.5	85.5
1	5.2	11.8	31.8	55.9	74.3	85.7	86.2	86.2
2	4.8	10.5	31.1	56.8	76.0	87.4	87.9	87.9
3	4.5	10.5	32.0	58.4	77.6	88.7	89.2	89.2
4	4.6	11.2	33.1	59.8	78.6	89.4	89.8	89.8
6	5.0	12.5	34.7	61.0	79.3	89.9	90.4	90.4
8	5.4	13.3	35.5	61.6	79.8	90.3	90.8	90.8
10	5.5	13.5	35.8	61.7	80.0	90.4	90.9	90.9

Note: For each (b_1, b_2), the rejection probabilities for the standard test are adjusted such that the robust test and the standard test have the same rejection probabilities under the null. $n = 500$, $\pi_{1,0} = 0$.

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

$$\Pr\left(W_n(R) \leq \widehat{c}_{n,1-\alpha}\right) \to \Pr(\mathcal{W}(h_0) \leq \chi^2_{d_r,1-\alpha}) = 1 - \alpha. \qquad \text{(C.18)}$$

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.18) 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 t statistic. □

References

Andrews, D.W.K., 1994. Asymptotics for semiparametric econometric models via stochastic equicontinuity. Econometrica 62 (1), 43–72.

Andrews, D.W.K., Barwick, P.J., 2012. Inference for parameters defined by moment inequalities: a recommended moment selection procedure. Econometrica 80 (6), 2805–2826.

Andrews, D.W.K., Cheng, X., 2012. Estimation and inference with weak, semi-strong, and strong identification. Econometrica 80 (5), 2153–2211.

Andrews, D.W.K., Cheng, X., 2013. Maximum likelihood estimation and uniform inference with sporadic identification failure. J. Econometrics 173 (1), 36–56.

Andrews, D.W.K., Cheng, X., 2014. GMM estimation and uniform subvector inference with possible identification failure. Econometric Theory 30, 287–333.

Andrews, D.W.K., Cheng, X., Guggenberger, P., 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., Guggenberger, P., 2009a. Hybrid and size-corrected subsampling methods. Econometrica 77 (3), 721–762.

Andrews, D.W.K., Guggenberger, P., 2009b. Validity of subsampling and plug-in asymptotic inference for parameters defined by moment inequalities. Econometric Theory 25 (03), 669–709.

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

Andrews, D.W.K., Guggenberger, P., 2014a. Asymptotic Size of Kleibergen's LM and Conditional LR Tests for Moment Condition Models, Discussion paper. Yale University and Pennsylvania State University.

Andrews, D.W.K., Guggenberger, P., 2014b. Identification—and Singularity-Robust Inference for Moment Condition Models, Discussion paper. Yale University and Pennsylvania State University.

Andrews, D.W.K., Moreira, M.J., Stock, J.H., 2006. Optimal two-sided invariant similar tests for instrumental variables regression. Econometrica 74 (3), 715–752.

Andrews, D.W.K., Ploberger, W., 1994. Optimal tests when a nuisance parameter is present only under the alternative. Econometrica 62 (6), 1383–1414.

Andrews, D.W.K., Soares, G., 2010. Inference for parameters defined by moment inequalities using generalized moment selection. Econometrica 78 (1), 119–157.

Andrews, D.W.K., Stock, J.H., 2007. Testing with many weak instruments. J. Econometrics 138 (1), 24–46.

Andrews, I., 2013. Conditional Linear Combination Tests for Weakly Identified Models, Discussion paper. MIT.

Andrews, I., Mikusheva, A., 2012. A geometric approach to weakly identified econometric models, Discussion paper. MIT.

Andrews, I., Mikusheva, A., 2015. Maximum likelihood inference in weakly identified DSGE models. Quant. Econ. 6, 123–152.

Antoine, B., Renault, E., 2009. Efficient GMM with nearly-weak instruments. Econom. J. 12 (s1), S135–S171.

Antoine, B., Renault, E., 2012. Efficient minimum distance estimation with multiple rates of convergence. J. Econometrics 170 (2), 350–367.

Bec, F., Salem, M.B., Carrasco, M., 2010. Detecting mean reversion in real exchange rates from a multiple regime STAR model. Ann. Econ. Stat. (99/100), 395–427.

Billingsley, P., 1968. Convergence of Probability Measures. In: Wiley Series in probability and Mathematical Statistics: Tracts on probability and statistics. Wiley.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. J. R. Stat. Soc. Ser. B Stat. Methodol. 26 (2), 211–252.

Caner, M., 2010. Testing, estimation in GMM and CUE with nearly-weak identification. Econometric Rev. 29 (3), 330–363.

Caves, D.W., Christensen, L.R., Tretheway, M.W., 1980. Flexible cost functions for multiproduct firms. Rev. Econ. Stat. 62 (3), 477–481.

Chaudhuri, S., Zivot, E., 2011. A new method of projection-based inference in GMM with weakly identified nuisance parameters. J. Econometrics 164 (2), 239–251.

Chen, X., Ponomareva, M., Tamer, E., 2014. Likelihood inference in some finite mixture models. J. Econometrics 182 (1), 87–99.

Choi, I., Phillips, P.C.B., 1992. Asymptotic and finite sample distribution theory for IV estimators and tests in partially identified structural equations. J. Econometrics 51 (1–2), 113–150.

Clark, J.A., 1984. Estimation of economies of scale in banking using a generalized functional form. J. Money. Credit. Bank 16 (1), 53–68.

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., Taamouti, M., 2005. Projection-based statistical inference in linear structural models with possibly weak instruments. Econometrica 73 (4), 1351–1365.

Dufour, J.-M., Taamouti, M., 2007. Further results on projection-based inference in IV regressions with weak, collinear or missing instruments. J. Econometrics 139 (1), 133–153.

Elliott, G., Müller, U.K., Watson, M.W., 2012. Nearly Optimal Tests when a Nuisance Parameter is Present Under the Null Hypothesis, Discussion paper. UCSD and Princeton University.

Giannakas, K., Tran, K.C., Tzouvelekas, V., 2000. Efficiency, technological change and output growth in Greek olive growing farms: a Box–Cox approach. Appl. Econ 32 (7), 909–916.

Granger, C.W.J., Terasvirta, T., 1993. Modelling Non-Linear Economic Relationships. Oxford University Press, no. 9780198773207 in OUP Catalogue.

Guerron-Quintana, P., Inoue, A., Kilian, L., 2013. Frequentist inference in weakly identified dynamic stochastic general equilibrium models. Quant. Econ. 4 (2), 197–229.

Guggenberger, P., Kleibergen, F., Mavroeidis, S., Chen, L., 2012. On the asymptotic sizes of subset AndersonCRubin and Lagrange multiplier tests in linear instrumental variables regression. Econometrica 80 (6), 2649–2666.

Guggenberger, P., Smith, R.J., 2005. Generalized empirical likelihood estimators and tests under partial, weak, and strong identification. Econometric Theory 21 (04), 667–709.

Hahn, J., Kuersteiner, G., 2002. Discontinuities of weak instrument limiting distributions. Econom. Lett. 75 (3), 325–331.

Hansen, B.E., 1996. Inference when a nuisance parameter is not identified under the null hypothesis. Econometrica 64 (2), 413–430.

Kitamura, Y., Phillips, P.C.B., 1997. Fully modified IV, GIVE and GMM estimation with possibly non-stationary regressors and instruments. J. Econometrics 80 (1), 85–123.

Kleibergen, F., 2002. Pivotal statistics for testing structural parameters in instrumental variables regression. Econometrica 70 (5), 1781–1803.

Kleibergen, F., 2005. Testing parameters in GMM without assuming that they are identified. Econometrica 73 (4), 1103–1123.

Kleibergen, F., 2014. Efficient size correct subset inferencen in linear instrumental variables regression, Discussion paper. Brown University.

Kuan, C.-M., White, H., 1994. Artificial neural networks: an econometric perspective. Econometric Rev. 13 (1), 1–91.

Lee, L.-f., 2005. Classical inference with ML and GMM estimates with various rates of convergence, Discussion paper. Ohio State University.

Lee, L.-f., 2010. Pooling estimates with different rates of convergence: a minimum X2 approach with emphasis on a social interactions model. Econometric Theory 26 (01), 260–299.

Luukkonen, R., Saikkonen, P., Tersvirta, T., 1988. Testing linearity against smooth transition autoregressive models. Biometrika 75 (3), 491–499.

Ma, J., Nelson, C.R., 2010. Valid Inference for a Class of Models Where Standard Inference Performs Poorly: Including Nonlinear Regression, ARMA, GARCH, and Unobserved Components. In: Economics Series, vol. 256. Institute for Advanced Studies.

McAleer, M., Medeiros, M.C., 2008. A multiple regime smooth transition Heterogeneous Autoregressive model for long memory and asymmetries. J. Econometrics 147 (1), 104–119.

McCloskey, A., 2012. Bonferroni-Based Size-Correction for Nonstandard Testing Problems, Working Papers 2012-16, Brown University, Department of Economics.

Montiel Olea, J.L., 2013. Efficient Conditionally Similar Tests: Finite-Sample Theory and Large-Sample Applications, Discussion paper. New York University.

Moreira, M.J., 2003. A conditional likelihood ratio test for structural models. Econometrica 71 (4), 1027–1048.

Nelson, C.R., Startz, R., 1990. Some further results on the exact small sample properties of the instrumental variable estimator. Econometrica 58 (4), 967–976.

Nelson, C.R., Startz, R., 2007. The zero-information-limit condition and spurious inference in weakly identified models. J. Econometrics 138 (1), 47–62.

Phillips, P., 1989. Partially identified econometric models. Econometric Theory 5 (02), 181–240.

Phillips, P.C.B., Park, J.Y., 1988. On the formulation of wald tests of nonlinear restrictions. Econometrica 56 (5), 1065–1083.

Qu, Z., 2014. Inference in DSGE models with possible weak identification. Quant. Econ. 5, 457–494.

Radchenko, P., 2008. Mixed-rates asymptotics. Ann. Statist. 36 (1), 287–309.

Sargan, J.D., 1983. Identification and lack of identification. Econometrica 51 (6), 1605–1633.

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. Cambridge University Press, pp. 184–230.

Shi, X., Phillips, P.C., 2012. Nonlinear cointegrating regression under weak identification. Econometric Theory 28 (03), 509–547.

Shintani, M., Terada-Hagiwara, A., Yabu, T., 2013. Exchange rate pass-through and inflation: A nonlinear time series analysis. J. Internat. money. Financ 32 (0), 512–527.

Sims, C.A., Stock, J.H., Watson, M.W., 1990. Inference in linear time series models with some unit roots. Econometrica 58 (1), 113–144.

Staiger, D., Stock, J.H., 1997. Instrumental variables regression with weak instruments. Econometrica 65 (3), 557–586.

Stock, J.H., Wright, J., 2000. GMM with weak identification. Econometrica 68 (5), 1055–1096.

Terasvirta, T., 1994. Specification, estimation, and evaluation of smooth transition autoregressive models. J. Amer. Statist. Assoc. 89 (425), 208–218.

Tripathi, G., 1999. A matrix extension of the Cauchy–Schwarz inequality. Econom. Lett. 63 (1), 1–3.

van der Vaart, A., Wellner, J., 1996. Weak convergence and empirical processes. In: Springer series in statistics. Springer.

van Dijk, D., Franses, P.H., 1999. Modeling multiple regimes in the business cycle. Macroecon. Dyn. 3 (03), 311–340.

White, H., 1989. An additional hidden unit test for neglected nonlinearity in multilayer feedforward networks. In: Neural Networks, 1989. IJCNN., International Joint Conference on. IEEE, pp. 451–455.