Robust Confidence Intervals in Nonlinear Regression under Weak Identification*

Xu Cheng†
Department of Economics
Yale University

JOB MARKET PAPER

December 21, 2008

Abstract

In this paper, we develop a practical procedure to construct confidence intervals (CIs) in a weakly identified nonlinear regression model. When the coefficient of a nonlinear regressor is small, modelled here as local to zero, the signal from the respective nonlinear regressor is weak, resulting in weak identification of the unknown parameters within the nonlinear regression component. In such cases, standard asymptotic theory can provide a poor approximation to finite-sample behavior and failure to address the problem can produce misleading inferences. This paper seeks to tackle this problem in complementary ways. First, we develop a local limit theory that provides a uniform approximation to the finite-sample distribution irrespective of the strength of identification. Second, standard CIs based on conventional normal or chi-squared approximations as well as subsampling CIs are shown to be prone to size distortions that can be severe. Third, a new confidence interval (CI) is constructed that has good finite-sample coverage probability. Simulation results show that when the nonlinear function is a Box-Cox type transformation, the nominal 95% standard CI and subsampling CI have asymptotic sizes of 53% and 2.3%, respectively. In contrast, the robust CI has correct asymptotic size and a finite-sample coverage probability of 93.4% when sample size is 100.

Keywords: Asymptotic size, confidence interval, model selection, nonlinear regression, subsampling, uniform convergence, weak identification.

*I am very grateful to my advisors, Donald Andrews and Peter Phillips, and committee members, Xiaohong Chen and Yuichi Kitamura, for their continual support and encouragement. I am extremely indebted to Donald Andrews for his generous and thoughtful guidance. I also have benefited from insightful comments made by Martin Browning, Patrik Guggenberger, Bruce Hansen, Oliver Linton, James Morley, Taisuke Otsu, Yixiao Sun, Edward Vytlacil, Qiying Wang, and Yoon-Jae Whang. Financial support from a Carl Arvid Anderson Prize of the Cowles Foundation is acknowledged.

†Email: xu.cheng@yale.edu. Tel: 1 203 535 3093. Comments are welcome.

1 Introduction

This paper studies inference methods under weak identification. In particular, we consider construction of CIs with good finite-sample coverage probability in a nonlinear regression model. The model belongs to a broad class of models in which lack of identification occurs at some point(s) in the parameter space. When the true parameter is close to the point of non-identification, we are confronted with weak identification. As in the weak instruments literature, standard asymptotic approximations to the finite-sample distributions of test statistics can be poor under such weak identification. For example, even though the t statistic has a standard normal distribution asymptotically, use of critical values from the standard normal distribution can lead to large size distortions in weakly identified situations. This paper develops a new asymptotic distribution that provides good approximations to the finite-sample distribution uniformly over the parameter space. This new asymptotic distribution is non-standard, but its quantiles can be approximated by simulation. Using proper quantiles of the new asymptotic distribution, we construct a robust CI that has good finite-sample coverage probability irrespective of the strength of identification. The procedure is developed in a nonlinear regression model estimated by least squares (LS). But the idea carries over to more general weak identification set-ups with other criterion-based estimators.

Economic theory and empirical studies often suggest nonlinear relationships among economic variables. These relationships are commonly specified in a parametric form involving several nonlinear component functions with unknown transformation parameters and loading coefficients that measure the importance of each component. The simplest version of the nonlinear regression model that we consider takes the form

$$Y_i = \beta \cdot g(X_i, \pi) + Z_i'\zeta + U_i \text{ for } i = 1, ..., n, \tag{1.1}$$

where the loading coefficient β and the transformation coefficient π are both scalars. Examples of the nonlinear function $g(\cdot, \pi)$ include Box-Cox type transformation, logistic/exponential transformations in smooth transition models, and neural network models of nonlinear responses. When $\beta = 0$, the nonlinear regressor $g(X_i, \pi)$ does not enter the regression function and the parameter π is not identified. As a result, asymptotic distributions of test statistics are non-standard when $\beta = 0$ and different from those when $\beta \neq 0$. Discontinuities of the asymptotic distributions also occur in some other models such as those with a unit root, moment inequalities or weak instruments, and under post-model-selection inference.

Hypothesis testing when a nuisance parameter is not identified under the null is considered in Davies (1977, 1987), Andrews and Ploberger (1994), and Hansen (1996), among others. While hypothesis testing only considers non-standard asymptotics under the null $\beta = 0$, we need to consider the asymptotic distributions of the test statistics for $\beta = 0$ as well as $\beta \neq 0$ in order to construct a CI with good coverage in both cases.

The conventional way to construct a CI is to use the standard asymptotic distribution obtained when β is fixed at a point different from 0 and the sample size n goes to infinity. Even if the true value of β is different from 0, however, the standard asymptotic distribution does not necessarily give a good approximation to finite-sample behavior. The reason is that approximation by the standard asymptotic theory does not have uniform validity, even for $\beta \neq 0$. In other words, there can be parameter values for which the finite-sample coverage probability is smaller than the nominal level, no matter how large the sample size is. The intuition is that when β is close to the point of non-identification, i.e. $\beta = 0$, the finite-sample behavior of the test statistics is contaminated by the non-standard asymptotics under non-identification. The extent of this contamination decreases as the sample size gets larger, but it is always present in finite samples. As a result, the standard CI is valid only when β is large enough for the model to be strongly identified. Uniformity issues of this sort have been discussed recently in the subsampling and bootstrap literature by Andrews (2000), Andrews and Guggenberger (2007, 2009a, b, c), Mikusheva (2007), and Romano and Shaikh (2006, 2008).

The aim of the paper is to construct CIs with good finite-sample coverage probability *uniformly* over the parameter space. To this end, we first develop a local limit theory that provides a good approximation to the finite-sample behavior in any identification scenario. This uniform approximation is obtained by considering sequences of the loading coefficients β that drift to the non-identification point $\beta = 0$. The asymptotic distribution derived under these drifting sequences is used to approximate the finite-sample distribution for any fixed true value of β. More precisely, the weak identification case is modelled as β being in an $n^{-1/2}$ neighborhood of zero. Approximation devices of this kind also are employed in the weak instruments literature, e.g. Staiger and Stock (1997).

Applying the local limit theory, we provide explicit formulae for the asymptotic sizes of the standard CI and the subsampling CI based on Andrews and Guggenberger (2009a) (hereafter AG). The asymptotic size, defined as the limit of the smallest finite-sample coverage probability over the parameter space, is a good approximation to the finite-sample size when the sample size is large. We show that the asymptotic sizes depend on the specific nonlinear functional form $g(\cdot, \pi)$ and can be simulated from the analytical formulae derived here. Simulation results show that the nominal level 95% standard CIs for β and π based on asymptotic normality have asymptotic sizes of 52.9% and 94.5%, respectively, under a Box-Cox transformation, and 73.2% and 63.6%, respectively, under a logistic transformation. The severe size distortions of the standard method motivate interest in a new robust CI with correct asymptotic size under weak identification.

Here we introduce the idea of the robust CI. The $1-\alpha$ quantile of the finite-sample distribution depends on the strength of identification. In consequence, the asymptotic distribution of the test statistic under a suitable drifting sequence $\beta_n = n^{-1/2}b$ depends heavily on a nuisance parameter b, which indexes the identification strength. The larger is b, the stronger is the identification. In

the extreme case that $b = \pm\infty$, all parameters are strongly identified and the standard asymptotic distribution is justified. However, a larger critical value might be required under weak identification where $b \in R$. One way to deal with this problem is to take the supremum of the $1-\alpha$ quantile over all b. This least favorable CI has correct asymptotic size, but it can be unnecessarily long when the model is strongly identified, i.e. $b = \pm\infty$.

The robust CI introduced here improves upon the least favorable CI by using a model-selection procedure to choose the critical value. The idea is to use the data to determine whether the model is weakly identified. Under weak identification, the least favorable critical value should be employed to get correct asymptotic size. Otherwise, the standard critical value is used. We use a t statistic for β centered at 0 for the purpose of model selection. Specifically, we proceed with weak identification only if the t statistic is smaller than a tuning parameter. The tuning parameter has to be designed so that the model selection procedure is consistent under weak identification. We show that the robust CI has correct asymptotic size provided the tuning parameter diverges to infinity with the sample size. Suitable choices of the divergence rate are investigated by simulation. This model-selection procedure is analogous to the generalized moment selection method in Andrews and Soares (2007).

The paper also develops a sequential procedure to deal with multiple nonlinear regressors under different identification scenarios. It is shown that weak identification of any particular nonlinear regressor can have serious negative implications on inference for all the parameters in the model when standard asymptotics are employed. A new CI that is robust to weak identification is developed to address this difficulty in the multiple regressor context.

The paper also is related to the weak identification issue discussed by Dufour (1997). While Dufour (1997) provides theoretical results on length of the CI, we propose a practical procedure to construct a CI with correct asymptotic size. We do so under the assumption that the parameter space for the potentially nonidentified parameter is bounded. Dufour (1997) also has an extensive discussion of econometric models with weak identification problems, where methods in the present paper can be applied after modification and generalization. Another related literature is that covering the partially identified models, as in Chernozhukov, Hong, and Tamer (2007). However, the standard LS population criterion function is either uniquely minimized by the true value or by the entire parameter space, rather than by a subset of it.

The remainder of the paper is organized as follows. Section 2 describes the model and the estimation methods. To provide basic intuition, Section 3 derives a new local limit theory in a simple model with one nonlinear regressor. Section 4 provides explicit formulae for the asymptotic sizes of the standard CI and the subsampling CI. Simulation results with two specific nonlinear functions are reported. Section 5 introduces the robust CI and presents simulation results to investigate its finite-sample properties. Section 6 generalizes the local limit theory to a model with multiple nonlinear regressors as well as linear regressors. In particular, a sequential procedure is

introduced. Section 7 discusses the standard CI, subsampling CI, and robust CI in the general set-up. Section 8 concludes and provides future research directions. Proofs are collected in the Appendix.

2 Model

The general model we consider is

$$\begin{aligned} Y_i &= g(X_i,\pi)'\beta + Z_i'\zeta + U_i \text{ for } i = 1,...,n, \text{ where} \\ \pi &= (\pi_1,...,\pi_p)' \text{ and } g(X_i,\pi) = (g_1(X_i,\pi_1),...,g_p(X_i,\pi_p))'. \end{aligned} \tag{2.1}$$

In (2.1), $X_i \in R^d$, $Y_i, U_i \in R$, $\pi, \beta \in R^p$, and $Z_i, \zeta \in R^q$. The function $g_j(x,\pi_j) \in R$ is known and nonlinear in x for any given π_j. Examples of the nonlinear function include

Example 1. Box-Cox function: $g(x,\pi) = (x^\pi - 1)/\pi$,

Example 2. logistic function: $g(x,\pi) = (1+\exp(-(x-\pi))^{-1}$.

We assume the support of X_i is contained in a set $\mathcal{X}$. The parameter spaces for $(\beta',\zeta')'$ and π are $B \subset R^{p+q}$ and $\Pi = \Pi_1 \times \cdots \times \Pi_p$, respectively, where $\Pi_j \subset R$ for $j = 1,...,p$. We assume B is bounded and contains $(\beta',\zeta')'$ with β_j arbitrarily close to 0 for any j. We also assume Π is compact. The data and the function $g(X_i,\pi)$ are assumed to satisfy the following assumptions.

Assumption 1. $g(x,\pi)$ is twice continuously differentiable with respect to (wrt) π, $\forall \pi \in \Pi$ and $\forall x \in \mathcal{X}$. We denote the first and second order derivatives of $g_j(X_i,\pi_j)$ wrt π_j by $g_{\pi j}(X_i,\pi_j)$ and $g_{\pi\pi j}(X_i,\pi_j)$, respectively, and write $g_\pi(X_i,\pi) = (g_{\pi 1}(X_i,\pi_1),...,g_{\pi p}(X_i,\pi_p))'$.

Assumption 2. (a) $\{(X_i,Z_i,U_i) : i = 1,...,n\}$ are *i.i.d.*

(b) $E(U_i|X_i,Z_i) = 0$ a.s., $E(U_i^2|X_i,Z_i) = \sigma^2(X_i,Z_i)$ a.s., and $EU_i^4 < \infty$.

(c) $\forall \pi, \overline{\pi} \in \Pi$ with $\pi \neq \overline{\pi}$, $\mathbb{P}(g(X_i,\pi) = kg(X_i,\overline{\pi})) < 1$ for all $k \in R$.

(d) $E\sup_{\pi\in\Pi} g_j^4(X_i,\pi_j) + E\sup_{\pi\in\Pi} g_{\pi j}^4(X_i,\pi_j) + E\sup_{\pi\in\Pi} g_{\pi\pi j}^2(X_i,\pi_j) < \infty$, $\forall j \leq p$.

(e) EZ_iZ_i' is nonsingular and $E\|Z_i\|^4 < \infty$.

Assumption 2(c) is imposed to identify π when $\beta \neq 0$. Assumption 2(d) is needed to ensure uniform convergence used throughout the paper.

We are interested in confidence sets (CSs) for some sub-vectors of β, ζ, and π, and construct them by inverting test statistics based on the LS estimator. When the parameter of interest is a scalar, the CSs become CIs. We first derive the asymptotic distributions of the LS estimators of β, ζ, and π. Then these are used to determine the asymptotic distributions of the test statistics. By definition, the LS estimators are obtained by minimizing the sum of squared regression residuals. This procedure can be viewed in the following way. First, for each given π, we take $g(X_i,\pi)$ as an exogenous regressor and estimate β and ζ by the LS estimator on (2.1), yielding estimates $\widehat{\beta}(\pi)$

and $\widehat{\zeta}(\pi)$. The concentrated LS sample criterion function is the average of the squared residuals

$$Q_n(\pi) = n^{-1}\sum_{i=1}^{n}\widehat{U}_i^2(\pi), \text{ where } \widehat{U}_i(\pi) = Y_i - g(X_i,\pi)'\widehat{\beta}(\pi) - Z_i'\widehat{\zeta}(\pi). \tag{2.2}$$

The LS estimator of π, denoted by $\widehat{\pi}_n$, minimizes $Q_n(\pi)$ over $\pi \in \Pi$.[1] This yields

$$\widehat{\pi}_n = \arg\min_{\pi\in\Pi} Q_n(\pi), \ \widehat{\beta}_n = \widehat{\beta}(\widehat{\pi}_n), \text{ and } \widehat{\zeta}_n = \widehat{\zeta}(\widehat{\pi}_n). \tag{2.3}$$

3 Simple Model and Asymptotic Results

In order to explain the finite-sample and asymptotic results as clearly as possible, we first analyze a simple model in which there is only one nonlinear regressor, i.e. $p = 1$, and there are no linear regressors. The model then becomes

$$Y_i = \beta \cdot g(X_i,\pi) + U_i, \tag{3.1}$$

where $\beta \in R$ and $\pi \in R$. This simple model sheds light on how the magnitude of β affects the finite-sample and asymptotic properties of the LS estimators, the test statistics, and the LS-based CIs.

3.1 Asymptotic Size

Let θ be a generic notation for any parameter in the model. It can be β or π. We construct a CI for θ by inverting a test statistic $T_n(\theta_0)$ for $H_0 : \theta = \theta_0$. The nominal level $1-\alpha$ CI for θ is $CI_n = \{\theta : T_n(\theta) \le c_{n,1-\alpha}(\theta)\}$, where $c_{n,1-\alpha}(\theta)$ is the critical value. The critical value choice introduced in this paper can depend on the true value of θ as well as the sample size n. When $\gamma = (\beta,\pi)'$ is the true parameter, the coverage probability of a CI is

$$P_\gamma(\theta \in CI_n) = P_\gamma(T_n(\theta) \le c_{n,1-\alpha}(\theta)). \tag{3.2}$$

This paper focuses on the smallest finite-sample coverage probability of a CI over the parameter space, i.e. the finite-sample size. It is approximated by the asymptotic size defined as

$$AsyCS = \liminf_{n\to\infty}\inf_{\gamma\in\Gamma} P_\gamma(\theta \in CI_n) = \liminf_{n\to\infty}\inf_{\gamma\in\Gamma} P_\gamma(T_n(\theta) \le c_{n,1-\alpha}(\theta)), \tag{3.3}$$

where Γ is the parameter space of γ. Note that in the definition of the asymptotic size the operation $\inf_{\gamma\in\Gamma}$ is taken before the operation $\liminf_{n\to\infty}$. Without the uniformity over $\gamma \in \Gamma$ before the

[1] We actually search for the value of π on a set that is slightly larger than the parameter space Π to avoid the problems that occur when the parameter is on the boundary of the parameter space. For details, see Andrews (1999).

limit operation, only a pointwise result is obtained. A pointwise result does not provide a good approximation to the finite-sample size because the contamination on the finite-sample behavior by non-identification exists for any given sample size n. Only a uniform result can capture that the extent of this contamination decreases as sample size gets larger.

A key implication from the definition of the asymptotic size is that the true parameter value γ at which the smallest finite-sample coverage probability attains can vary with the sample size. Therefore, in order to investigate the asymptotic size we need to derive the asymptotic distributions of the test statistics $T_n(\theta)$ under sequences of true parameters $\gamma_n = (\beta_n, \pi_n)'$, where the subscript n denotes that the true value may change with the sample size n. The sequences that we consider are the ones that determine the asymptotic sizes of the CI based on results in AG. These are also the sequences under which asymptotic distributions provide good uniform approximations to the finite-sample distributions of the test statistics. This is consistent with the fact that the asymptotic size is defined to approximate the finite-sample size.

Specifically, the sequences of parameters we consider are characterized by a localization parameter

$$h = (b, \pi_0)', \text{ where } n^{1/2}\beta_n \to b \text{ and } \pi_n \to \pi_0. \tag{3.4}$$

The parameter space for h is $H = R_{[\pm\infty]} \times \Pi$, where $R_{[\pm\infty]} = R \cup \{\pm\infty\}$.

The localization parameter in (3.4) is a general representation that includes various identification situations. First, weak identification is modelled as $b \in R$ such that β_n converges to 0 at rate $n^{-1/2}$. The values of b characterize the specific paths of these drifting sequences. These sequences also correspond to those considered in the weak instruments asymptotics of Staiger and Stock (1997). Second, hypothesis testing for non-identification, i.e. $\beta = 0$, corresponds to $b = 0$. Third, standard asymptotics that are based on a fixed parameter $\beta \neq 0$ correspond to $b = \pm\infty$. Hence, standard asymptotics fail to reveal the finite-sample behavior under weak identification for $b \in R$.

3.2 Asymptotic Distributions

We now derive a local limit theory under these drifting sequences to investigate the asymptotic sizes of the CIs. The procedure roughly goes as follows. To analyze the LS-based test statistics, we start with the LS sample criterion function, which is the average squared residuals defined in (2.2). The advantage of working with this concentrated sample criterion function is that $\widehat{\beta}(\pi)$ has a closed form for any fixed π. We properly re-center and re-scale the sample criterion function $Q_n(\pi)$ to derive its asymptotic distribution. The limit is denoted by $Q(\pi)$, which may be stochastic. Invoking the continuous mapping theorem (CMT), the minimizer of $Q_n(\pi)$, i.e. $\widehat{\pi}_n$, is expected to converge (in a stochastic sense) to the minimizer of $Q(\pi)$. Finally, the limit of $\widehat{\pi}_n$ is plugged into the closed form $\widehat{\beta}(\pi)$ to obtain the asymptotic distributions of $\widehat{\beta}_n$ and the LS-based test statistics.

A key step in derivation of the local limit theory is to obtain different forms of $Q(\pi)$ under

various identification scenarios. In a standard set-up where β is fixed and bounded away from 0, the sample criterion function $Q_n(\pi)$ converges in probability to a non-random population criterion function $Q(\pi)$ uniformly over Π. This non-random function $Q(\pi)$ is uniquely minimized at the true value of π. However, under weak identification, where β_n is in an $n^{-1/2}$ neighborhood of 0, the limit of $Q_n(\pi)$ is random after proper re-centering and re-scaling. Our method is to view $Q_n(\pi)$ as an empirical process indexed by π, and to show that a centered and scaled version of $Q_n(\pi)$ converges weakly to a stochastic process $Q(\pi)$.

To define the stochastic limit of $Q_n(\pi)$ under weak identification, we let $S(\pi)$ be a mean zero Gaussian process with covariance kernel $\Omega(\pi,\overline{\pi}) = EU_i^2 g(X_i,\pi) g(X_i,\overline{\pi})$. Define the function $\Phi(\pi,\overline{\pi}) = Eg(X_i,\pi) g(X_i,\overline{\pi})$. For any given π and $\overline{\pi}$, $\Phi(\pi,\overline{\pi})$ is the covariance. Let $Y = (Y_1,...,Y_n)'$.

Assumption 3a. $\Phi(\pi,\pi) \geq \varepsilon \ \forall \pi \in \Pi$ for some $\varepsilon > 0$.

Lemma 3.1 *Suppose Assumptions* 1, 2, *and* 3a *hold.*
(**a**) *When* $n^{1/2}\beta_n \to b \in R$ *and* $\pi_n \to \pi_0$, $n\left(Q_n(\pi) - n^{-1}Y'Y\right) \Rightarrow -\Phi^{-1}(\pi,\pi)\left(S(\pi) + \Phi(\pi,\pi_0)b\right)^2$.
(**b**) *When* $\left|n^{1/2}\beta_n\right| \to \infty$ *and* $\pi_n \to \pi_0$, $\beta_n^{-2}\left(Q_n(\pi) - n^{-1}Y'Y\right) \to_p -\Phi^{-1}(\pi,\pi)\Phi^2(\pi,\pi_0)$ *uniformly over* Π.

Comments: 1. In Lemma 3.1(a), $Q_n(\pi) - n^{-1}Y'Y$ converges in probability to 0 at rate n^{-1}. Note that the centering term $n^{-1}Y'Y$ does not depend on π. Hence, the asymptotic distribution of $\widehat{\pi}_n$ only depends on the non-central chi-square process on the rhs of Lemma 3.1(a).[2]

2. In Lemma 3.1(b), β_n is bounded away from 0 or converges to 0 slower than $n^{-1/2}$. In the former case, the model is standard. In the latter case, $Q_n(\pi) - n^{-1}Y'Y \to_p 0$ at rate β_n^2. This rate of convergence is slower than that in Lemma 3.1(a). As in part (a), the centering variable does not depend on π, so the probability limit of $\widehat{\pi}_n$ only depends on the rhs of Lemma 3.1(b), which is uniquely minimized at π_0 by the Cauchy-Schwarz inequality under Assumption 2(c). The rhs of Lemma 3.1(a) is related to that of Lemma 3.1(b). As b diverges to plus or minus infinity, $\Phi(\pi,\pi_0)b$ dominates $S(\pi)$.

By invoking the CMT, we show below that $\widehat{\pi}_n$ converges in distribution to the minimizer of the rhs of Lemma 3.1(a) when $\beta_n = O(n^{-1/2})$ and converges in probability to π_0 when β_n is of a larger order than $n^{-1/2}$, represented by $\beta_n \gg O(n^{-1/2})$. Before applying the CMT, however, conditions are needed to ensure the argmin functions are continuous here.

Lemma 3.2 *Under Assumption* 2, *the sample paths of the non-central chi-square process* $-\Phi^{-1}(\pi,\pi)\left(S(\pi) + \Phi(\pi,\pi_0)b\right)^2$ *have unique minima over* $\pi \in \Pi$ *with probability one.*

[2] For fixed π, $\Phi^{-1}(\pi,\pi)\left(S(\pi) + \Phi(\pi,\pi_0)b\right)^2$ has a non-central chi-square distribution under homoskedasticity, although not under heteroskedasticity. Nevertheless, for simplicity, we call its opposite $-\Phi^{-1}(\pi,\pi)(S(\pi) + \Phi(\pi,\pi_0)b)^2$ a non-central chi-square process throughout the paper.

The uniqueness of the minimizer in Lemma 3.2 ensures that the argmin function on a compact support is a continuous function of the sample paths of the chi-square process with probability one. This property is used to derive the asymptotic distribution of $\widehat{\pi}_n$ in Lemma 3.3 below. In the proof of Lemma 3.2, we use arguments analogous to those in Kim and Pollard (1990) regarding the unique maximizer of a Gaussian process.

Lemma 3.3 *Suppose Assumptions* 1, 2, *and* 3a *hold.*
(**a**) *When* $n^{1/2}\beta_n \to b \in R$ *and* $\pi_n \to \pi_0$,

$$\begin{aligned} \widehat{\pi}_n &\Rightarrow \pi^*(h), \text{ where} \\ \pi^*(h) &= \arg\min_{\pi\in\Pi}(-\{\Phi^{-1}(\pi,\pi)(S(\pi)+\Phi(\pi,\pi_0)b)^2\}) \text{ and } h=(b,\pi_0)'. \end{aligned}$$

(**b**) *When* $\left|n^{1/2}\beta_n\right| \to \infty$ *and* $\pi_n \to \pi_0$, $\widehat{\pi}_n - \pi_n \to_p 0$.

The intuition behind Lemma 3.3 is that whether π_n can be estimated consistently depends on the strength of the signal from $g(X_i,\pi_n)\beta_n$ relative to the noise from the errors. The strength of the signal is proportional to the magnitude of β_n. Under weak identification (or non-identification), i.e. $\beta_n = O(n^{-1/2})$, $\widehat{\pi}_n$ converges in distribution to a random variable that minimizes the sample paths of a non-central chi-square process. The randomness comes from the noise in the errors. Under strong identification, i.e. $\beta_n \gg O(n^{-1/2})$, $\widehat{\pi}_n$ is consistent because the noise is of a smaller order compared with the signal.

Next, we derive the asymptotic distributions of the LS estimator $\widehat{\beta}_n$ under different identification scenarios. When $\widehat{\pi}_n$ is inconsistent, we can show that $n^{1/2}(\widehat{\beta}(\pi)-\beta_n)$, regarded as an empirical process indexed by π, converges weakly to a stochastic process. The asymptotic distribution of $n^{1/2}(\widehat{\beta}_n-\beta_n)$ can be obtained by plugging the random limit of $\widehat{\pi}_n$ into the limit process of $n^{1/2}(\widehat{\beta}(\pi)-\beta_n)$, because both of them are continuous functions of the Gaussian process $S(\pi)$. When $\widehat{\pi}_n$ is consistent, the asymptotic normality of $n^{1/2}(\widehat{\beta}_n-\beta_n)$ can be established using standard arguments.

To specify the asymptotic distributions, we define

$$\begin{aligned} m_i(\pi) &= (g(X_i,\pi), g_\pi(X_i,\pi))', \; G(\pi) = Em_i(\pi)m_i(\pi)', \\ V(\pi) &= EU_i^2 m_i(\pi)m_i(\pi)', \text{and } \Sigma(\pi) = G^{-1}(\pi)V(\pi)G^{-1}(\pi). \end{aligned} \tag{3.5}$$

The vector $m_i(\pi)$ equals the vector of partial derivatives of $g(X_i,\pi)\beta$ wrt to the parameter vector $(\beta,\pi)'$ except that the latter has $g_\pi(X_i,\pi)\beta$ in place of $g_\pi(X_i,\pi)$. The standard asymptotic covariance matrix of the LS estimator, which is based on the partial derivative vector, involves β and becomes singular when the true parameter β_n drifts to 0. To avoid the problem of singularity, we

have to employ an asymptotic covariance matrix based on $m_i(\pi)$ and make the true parameter β_n part of the rate of convergence of the LS estimator of π, as shown in Lemma 3.4(b) below.

Assumption 4a. $G(\pi) \geq \varepsilon \ \forall \ \pi \in \Pi$ for some $\varepsilon > 0$.

Lemma 3.4 *Suppose Assumptions* 1, 2, 3a, *and* 4a *hold.*
(**a**) *When* $n^{1/2}\beta_n \to b \in R$ *and* $\pi_n \to \pi_0$,

$$n^{1/2}\left(\widehat{\beta}_n - \beta_n\right) \Rightarrow \tau\left(\pi^*(h), h\right), \text{ where } \tau(\pi, h) = \Phi^{-1}(\pi, \pi)\left(S(\pi) + \Phi(\pi, \pi_0)\, b\right) - b.$$

(**b**) *When* $\left|n^{1/2}\beta_n\right| \to \infty$ *and* $\pi_n \to \pi_0$,

$$\begin{pmatrix} n^{1/2}\left(\widehat{\beta}_n - \beta_n\right) \\ n^{1/2}\beta_n\left(\widehat{\pi}_n - \pi_n\right) \end{pmatrix} \Rightarrow N\left(0, \Sigma(\pi_0)\right).$$

Comments: 1. Under weak identification in Lemma 3.4(a), $n^{1/2}(\widehat{\beta}_n - \beta_n)$ is not asymptotically normal. Instead, it is characterized by a Gaussian process $\tau(\pi, h)$ indexed by π and a random variable $\pi^*(h)$, defined in Lemma 3.3(a). The non-standard asymptotic distribution of $n^{1/2}(\widehat{\beta}_n - \beta_n)$ is determined by the finite localization parameter h.

2. In Lemma 3.4(b), where $\beta_n \gg O(n^{-1/2})$, both $\widehat{\beta}_n$ and $\widehat{\pi}_n$ have asymptotic normal distributions. However, the convergence rate of $\widehat{\pi}_n$ depends on β_n. Specifically, its convergence rate is $n^{-1/2}\beta_n^{-1}$, which is slower than $n^{-1/2}$ if β_n converges to 0. The reason is that when β_n converges to 0, the signal-to-noise ratio of the regressor $g(X_i, \pi_n)\beta_n$ also converges to 0. As a result, in order for $\widehat{\pi}_n$ to achieve the same level of accuracy as in a standard set-up, more data is required to compensate for the weak signal.

3. When β_n is fixed at a point different from 0, the model is standard and the LS estimator is $n^{1/2}$ consistent. In this case, we can move β_n from the normalization of $\widehat{\pi}_n$ on the lhs of Lemma 3.4(b) to the rhs. With this adjustment, the rhs becomes a standard covariance matrix of the LS estimator.

Because both β and π are scalars in the simple model, t statistics are employed to construct the CIs. Define

$$\begin{aligned} \widehat{\Sigma}(\pi) &= \widehat{G}^{-1}(\pi)\,\widehat{V}(\pi)\,\widehat{G}^{-1}(\pi), \text{ where} \\ \widehat{G}(\pi) &= n^{-1}\sum_{i=1}^{n} m_i(\pi)\, m_i(\pi)', \ \widehat{V}(\pi) = n^{-1}\sum_{i=1}^{n} \widehat{U}_i^2(\pi)\, m_i(\pi)\, m_i(\pi)', \end{aligned} \tag{3.6}$$

$m_i(\pi)$ is defined in (3.5), and $\widehat{U}_i(\pi)$ is defined in (2.2). The t statistics for β and π are

$$T_{\beta,n} = n^{1/2}(\widehat{\beta}_n - \beta_n)/\widehat{\sigma}_\beta \text{ and } T_{\pi,n} = n^{1/2}\left(\widehat{\pi}_n - \pi_n\right)/\widehat{\sigma}_\pi, \tag{3.7}$$

where $\widehat{\sigma}_\beta = \widehat{\Sigma}_n(\widehat{\pi}_n)_{11}^{1/2}$ and $\widehat{\sigma}_\pi = \widehat{\Sigma}_n(\widehat{\pi}_n)_{22}^{1/2}\widehat{\beta}_n^{-1}$. Note that $\widehat{\sigma}_\pi$ takes into account the fact that the normalization factor for $\widehat{\pi}_n$ is $n^{1/2}\beta_n$ in Lemma 3.4(b). Both $T_{\beta,n}$ and $T_{\pi,n}$ are equivalent to the standard definitions of the t statistics.

Theorem 3.1 *Suppose Assumptions* 1, 2, 3a, *and* 4a *hold.*
(**a**) *When* $n^{1/2}\beta_n \to b \in R$ *and* $\pi_n \to \pi_0$,

$$T_{\beta,n} \Rightarrow T_\beta(\pi^*(h), h) \text{ and } T_{\pi,n} \Rightarrow T_\pi(\pi^*(h), h), \text{ where}$$
$$T_\beta(\pi, h) = \tau(\pi, h)(\Sigma(\pi)_{11})^{-1/2} \text{ and } T_\pi(\pi, h) = (\tau(\pi, h) + b)(\pi - \pi_0)(\Sigma(\pi)_{22})^{-1/2}.$$

(**b**) *When* $\left|n^{1/2}\beta_n\right| \to \infty$ *and* $\pi_n \to \pi_0$,

$$T_{\beta,n} \Rightarrow N(0,1) \text{ and } T_{\pi,n} \Rightarrow N(0,1).$$

Comment: Under weak identification in Theorem 3.1(a), the asymptotic distributions of the t statistics are characterized by the Gaussian processes $T_\beta(\pi, h)$ and $T_\pi(\pi, h)$ together with the limit random variable $\pi^*(h)$, all of which are determined by the finite localization parameter h. With the analytical formulae derived, proper quantiles of these non-standard asymptotic distributions can be simulated. In Lemma 3.1(b), where $\beta_n \gg O(n^{-1/2})$, both of the t statistics have standard normal distributions as expected. It is worth repeating our claim that under weak identification the non-standard asymptotic distribution in part (a) provides a much better approximation to the finite-sample behavior than the standard asymptotic distribution in part (b) does. Simulation results given in Section 5.3 corroborate this.

4 Standard and Subsampling Confidence Intervals

As in Section 3.1, let θ be a generic notation for any parameter in the model and $T_n(\theta)$ be a test statistic for θ. The nominal level $1-\alpha$ CI for θ is $CI_n = \{\theta : T_n(\theta) \leq c_{n,1-\alpha}(\theta)\}$, where $c_{n,1-\alpha}(\theta)$ is the critical value. The standard CI and the subsampling CI are obtained by different choices of their critical values. The critical value for a standard CI is the $1-\alpha$ quantile of the asymptotic distribution derived under strong identification. This is the standard normal distribution in a simple model with t statistics.[3] In this case, the standard critical value for a nominal level $1-\alpha$ symmetric two-sided CI is $z_{1-\alpha/2}$, which is the $1-\alpha/2$ quantile of the standard normal distribution.

We now define the subsampling critical value. The idea of subsampling is to use the empirical distribution of the subsample statistics to approximate the finite-sample distribution of the full-sample statistics. Let n_s denote the subsample size when the full-sample size is n. For the asymptotic

[3] When symmetric two-sided CI is constructed for a scalar parameter, the asymptotic distribution of the test statistic is $|Z|$, where $Z \sim N(0,1)$.

results we assume that $n_s \to \infty$ and $n_s/n \to 0$ as $n \to \infty$. The number of subsamples of length n_s is $q_n = n!/\left((n-n_s)!n_s!\right)$ with *i.i.d.* observations. The subsample statistics used to construct the subsampling critical values are $\{T_{n,n_s,j}(\theta) : j = 1, ..., q_n\}$, where $T_{n,n_s,j}(\theta)$ is a subsample statistic defined exactly as $T_n(\theta)$ but based on the j^{th} subsample of size n_s rather than the full sample. The subsampling critical value $c_{n,n_s}(1-\alpha)$ is the $1-\alpha$ sample quantile of $\{T_{n,n_s,j}(\theta) : j = 1, ..., q_n\}$.

4.1 Explicit Formulae for Asymptotic Sizes

In order to provide explicit formulae for the asymptotic sizes of the standard CI and the subsampling CI, we first characterize their critical values with the new local limit theory. Let J_h be the asymptotic distribution of $T_n(\theta)$ when h is the localization parameter associated with the drifting sequence of true parameter defined in (3.4). Analytical formulae of J_h for θ being β and π are given in Theorem 3.1. Let $c_h(1-\alpha)$ be the $1-\alpha$ quantile of J_h. The standard critical value is $c_\infty(1-\alpha)$, which is obtained when $b = \pm\infty$ under strong identification.[4]

Under the localization parameter h, the subsampling critical value is denoted by $c_l(1-\alpha)$. According to AG, the relationship between l and h is given in the set LH defined below. The basic idea is that, due to the smaller sample size, the sampling distribution of the subsample test statistic is affected more by the non-identification than that of the full-sample statistic. As a result, the subsample statistic behaves as if the true value is closer to the non-identification point. Under the assumption $n_s/n \to 0$, which is required for asymptotic validity of the subsampling method, the relationship between the subsample and full-sample statistics are characterized by the set

$$\begin{aligned} LH = \big\{&(l,h) \in H \times H : l = (l_b, \pi_0),\ h = (b, \pi_0),\ \text{and (i) } l_b = 0 \text{ if } |b| < \infty, \\ &\text{(ii) } l_b \in R_{+,\infty} \text{ if } b = +\infty,\ \text{and (iii) } l_b \in R_{-,\infty} \text{ if } b = -\infty.\big\} \end{aligned} \tag{4.1}$$

where $R_{+,\infty} = \{x \in R : x \geq 0\} \cup \{\infty\}$ and $R_{-,\infty} = \{x \in R : x \leq 0\} \cup \{-\infty\}$.[5]

By verifying the high-level assumptions in AG with the non-standard asymptotic distribution derived in Theorem 3.1, we establish asymptotic sizes of the standard CI and the subsampling CI for θ in the following theorem. Here θ can be either β or π.

Theorem 4.1 *Suppose Assumptions* 1, 2, 3a, *and* 4a *hold. Then,*
(**a**) $AsyCS = \inf_{h \in H} J_h(c_\infty(1-\alpha))$ *for the standard CI and*
(**b**) $AsyCS = \inf_{(l,h) \in LH} J_h(c_l(1-\alpha))$ *for the subsampling CI.*

Comments: As is evident in J_h, the asymptotic sizes depend on the specific functional form of $g(\cdot, \pi)$. Using these analytical formulae and the asymptotic distributions obtained in Theorem 3.1, we can simulate the asymptotic sizes of the standard CI and the subsampling CI. Simulation results

[4] When $b = \pm\infty$, π_0 does not affect the limit distribution J_h.
[5] Note that l and L correspond to g and G in AG.

for two specific nonlinear functions are reported in Table 4.1 on page 15. Moreover, the asymptotic sizes also depend on the parameter space Π, as the inconsistent estimator $\pi^*(h)$, defined in Lemma 3.3, is involved in J_h.

4.2 Simulations for Standard and Subsampling CIs

Graphs of $c_h(1-\alpha)$ as a function of b for fixed π_0 are informative regarding the behavior of the standard CI and the subsampling CI. A CI has asymptotic size greater than or equal to $1-\alpha$ only if the probability limit of the critical value is greater than or equal to $c_h(1-\alpha)$. Hence, for a standard CI to have correct asymptotic size, one needs $c_\infty(1-\alpha) \geq c_h(1-\alpha)$, for all $h \in H$. For example, this occurs if the graph is increasing in b for $b \geq 0$ and decreasing in b for $b < 0$ for each π_0. On the other hand, for a subsampling CI to have correct asymptotic size, one needs $c_l(1-\alpha) \geq c_h(1-\alpha)$, for all $(l,h) \in LH$. This occurs if the graph is decreasing in b for $b \geq 0$ and increasing in b for $b < 0$ for each π_0. Other cases where the quantile function $c_h(1-\alpha)$ is non-monotonic in $|b|$ are discussed in AG.

We now investigate the symmetric two-sided CIs for the scalar parameters in the following two examples. Both of them include intercepts and linear regressors in addition to a nonlinear regressor to mimic empirical applications with nonlinear regressions. Theoretical results in a general model of this sort are analogous to those derived in the simple model above. The general model is discussed in Sections 6 and 7 below.

Example 1—Cont.: The first results are based on the model

$$Y_i = \zeta_0 + \zeta_1 X_i + \beta g(X_i, \pi) + U_i, \text{ where } g(X_i, \pi) = (X_i^\pi - 1)/\pi. \tag{4.2}$$

The distributions of X_i and U_i are $X_i \sim N(6, 0.25)$ and $U_i \sim N(0, 0.25)$, respectively.[6] The parameter space for π is $[1.5, 4]$. The quantile functions of the test statistics for the symmetric two-sided CIs for β and π are presented in Figure 4.1 on the next page. Since these quantile graphs are symmetric wrt b around 0, we only report the graphs for $b \geq 0$.[7] For any fixed π_0, the graph for β first slopes up and then slopes down, with the maximum above the value 1.96. Hence, these quantile graphs imply that neither the standard CI nor the subsampling CI for β has correct asymptotic size. For any fixed π_0, the graph for π slopes up and always stays below the value 1.96. These graphs indicate that the standard CI for π has correct asymptotic size, while the subsampling CI does not.

[6] In the nonlinear regression model, the results are not invariant to the scales of X_i and U_i. The true values of ζ_0 and ζ_1 do not affect the results.

[7] The asymptotic distributions J_h in Theorem 3.1 are odd functions of b. Hence, asymptotic distributions of the test statistics for symmetric two-sided CIs have quantiles that are even functions of b. This is not true for one-sided CIs.

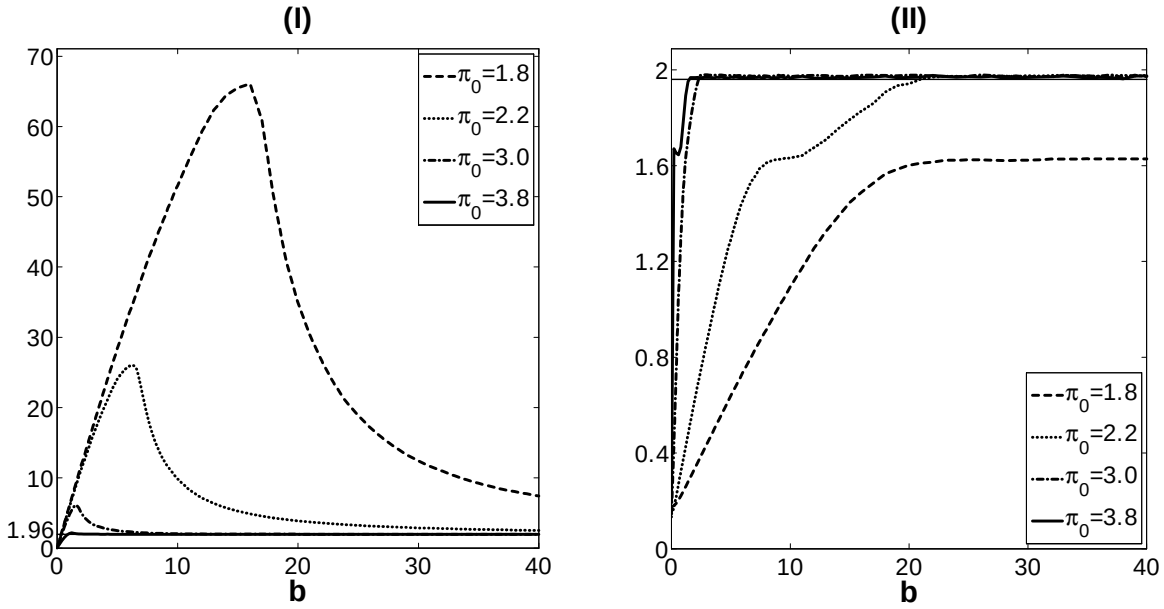

Figure 4.1: .95 Asymptotic Quantiles of the Test Statistics for Symmetric Two-sided CIs in Example 1 with the Box-Cox Function: (I) CI for β and (II) CI for π.

The asymptotic and finite-sample sizes of the standard CIs and the subsampling CIs for ζ_1, β, and π are reported in Table 4.1 on the following page.[8] The asymptotic sizes in Theorem 4.1 were computed by simulation. The nominal 95% standard CIs for ζ_1 and β have asymptotic sizes of 72.6% and 52.9%, respectively, while their counterparts for the subsampling CIs are 17.7% and 2.3%, respectively. The standard CI for π has asymptotic size close to the nominal level, while the subsampling CI does not. The finite-sample sizes are close to the asymptotic sizes and consistent with the quantile graphs in Figure 4.1.

Example 2—Cont.: The smooth transition model considered is of the form

$$Y_i = Z_i'\zeta_0 + \zeta_1 X_i + \beta g\left(X_i, \pi\right) + U_i, \text{ where } g\left(X_i, \pi\right) = X_i(1 + \exp(-(X_i - \pi))^{-1}, \tag{4.3}$$

where Z_i is a 3×1 vector of exogenous variables including an intercept. The distributions of X_i and U_i are $X_i \sim N\left(6.6, 3.6\right)$ and $U_i \sim N\left(0, 0.09\right)$, respectively.[9] The parameter space for π is $[4.6, 8.3]$, where the lower and upper ends are 15% and 85% quantiles of X_i, respectively. The quantile graphs of the test statistics for the symmetric two-sided CIs for β and π are presented in Figure 4.2 on

[8] The results in Tables 4.1 and 5.1 are based on 10,000 simulation repetitions. For example 1, the search over b to determine the Min is done on the interval $[0, 150]$ with stepsize 0.2 on [0,10], stepsize 1 on [10,50], and stepsize 10 on [50,150]. The search over π_0 to determine the Min is done on the set {1,8, 2, 2,2, 2.6, 3.0, 3.4, 3.8}. For example 2, the search over b to determine the Min is done on the interval [0,10], with stepsize 0.05 on [0,2.8] and [6.5,7], stepsize 0.01 on [2.8, 6.5], and stepsize 0.2 on [7,10]. These grids are refined based on previous simulations with coarse grids. The search over π_0 to determine the Min is done on the set [4.8, 7.8] with step size 0.5 and at the point 8.2.

[9] The parameter values used in the simulation are designed to mimic those in an empirical application.

Table 4.1: Asymptotic and Finite-sample Sizes of Nominal 95% Symmetric Two-sided Standard and Subsampling CIs

	Standard CI				Sub CI
	$n = 100$	$n = 250$	$n = 500$	Asy	Asy
Example 1. Box-Cox Function					
ζ_1	75.1	74.1	72.7	72.6	17.7
β	53.7	53.5	52.5	52.9	2.3
π	94.3	94.4	94.8	94.5	9.1
Example 2. Logistic Smooth Transition Function					
ζ_1	72.6	73.6	73.2	74.0	75.7
β	71.8	73.1	72.2	73.2	76.0
π	61.1	62.6	62.6	63.6	90.1

page 16. For any fixed π_0, the graphs are non-monotonic with the maximum above the value 1.96. Hence, these quantile graphs imply that the standard CIs and the subsampling CIs for β and π all suffer from size distortions.

The asymptotic and finite-sample sizes of the standard CIs and subsampling CIs are reported in Table 4.1. As predicted by the quantile graphs, the nominal 95% standard CIs and subsampling CIs for ζ_1, β, and π all under-cover. The standard CIs for ζ_1, β, and π have asymptotic sizes of 74.0%, 73.2%, and 63.6%, respectively. Their counterparts for the subsampling CIs are 75.7%, 76.0%, and 90.1%, respectively. Finite-sample simulations confirm that these are good approximations to the finite-sample coverage probabilities.

5 A New Robust Confidence Interval

When a larger critical value is needed under weak identification than under strong identification and non-identification, as in the two examples above, both the standard CI and the subsampling CI under-cover, sometimes severely. This motivates the introduction of a new CI that has correct asymptotic size. The robust CI introduced here is particularly useful when both the standard CI and subsampling CI fail. However, it remains valid when either of them has correct asymptotic size.

5.1 Description of the Robust CI

The idea behind the robust CI is as follows. The finite-sample distribution of the test statistic $T_n(\theta)$ depends on the identification strength of the model, which is characterized by the localization parameter b. A larger critical value might be required under weak identification than under strong identification. One way to deal with this problem is to construct a critical value that is large enough for all identification situations, but this least favorable CI may be too large and not informative

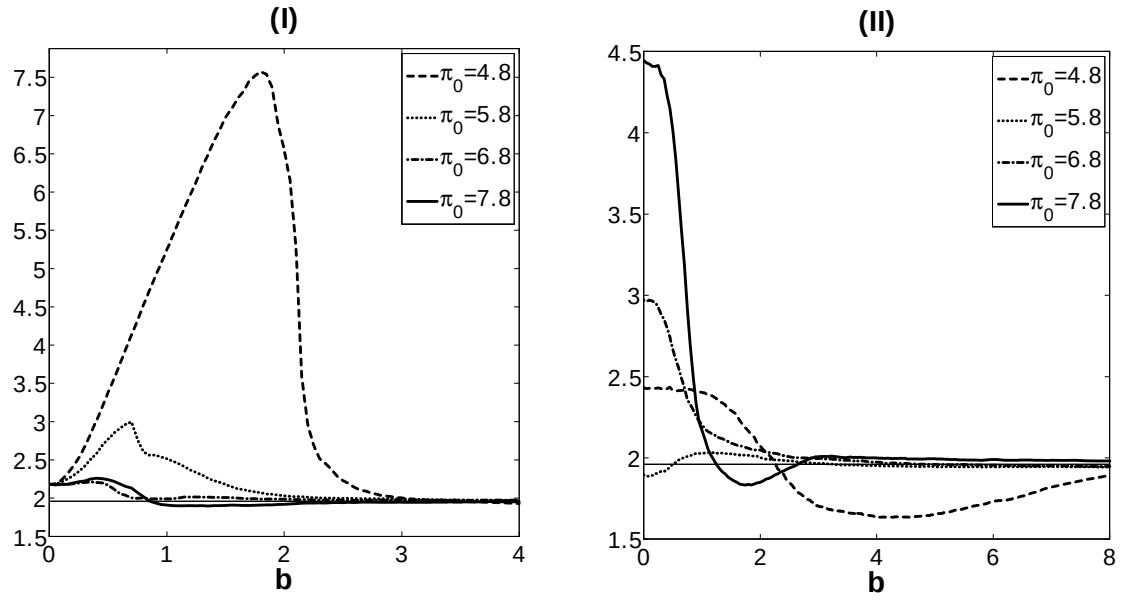

Figure 4.2: .95 Asymptotic Quantiles of the Test Statistics for Symmetric Two-sided CIs in Example 2 with Logistic Function: (I) CI for β and (II) CI for π.

when the model is strongly identified.

The robust CI improves upon the least favorable CI by using a model-selection procedure to choose the critical value. The idea is to use the data to determine whether b is finite. If b is finite, i.e. π is weakly identified (or not identified), the least favorable critical value should be employed to achieve correct asymptotic size. Otherwise, the standard critical value is used. This model-selection procedure used to choose the critical value is analogous to the generalized moment selection method in Andrews and Soares (2007).

The model-selection procedure is designed to choose between $M_0 : b \in R$ and $M_1 : |b| = \infty$. The statistic used for the model selection takes the form

$$t_n = \left| n^{1/2} \widehat{\beta}_n / \widehat{\sigma}_\beta \right| \tag{5.1}$$

and t_n is used to measure the degree of identification. Let $\{\kappa_n : n \geq 1\}$ be a sequence of constants that diverges to infinity as $n \to \infty$. We call κ_n the tuning parameter. We select M_0 if $t_n \leq \kappa_n$ and select M_1 otherwise. Under M_0, $t_n = |n^{1/2}(\widehat{\beta}_n - \beta_n)/\widehat{\sigma}_\beta + n^{1/2}\beta_n/\widehat{\sigma}_\beta| = O_p(1)$. Hence, we can consistently select M_0 provided the tuning parameter κ_n diverges to infinity. Suitable choices of κ_n include $(\ln n)^{1/2}$ and $(2 \ln \ln n)^{1/2}$ analogous to BIC and Hannan-Quinn information criteria, respectively. The finite-sample behavior of these choices are compared by simulation.

Following the model-selection procedure, the critical value for the robust CI is defined as

$$\widehat{c}_n(1-\alpha) = \begin{cases} \sup\limits_{h \in H} c_h(1-\alpha), & \text{if } t_n \leq \kappa_n \\ c_\infty(1-\alpha), & \text{if } t_n > \kappa_n \end{cases}. \tag{5.2}$$

Note that if the supremum of $c_h(1-\alpha)$ is attained at $|b| = \infty$, the robust CI is equivalent to the standard CI.

5.2 Construction Algorithm and Asymptotic Size Results

The algorithm to construct a robust CI in the simple model with one nonlinear regressor has four steps: (1) Estimate the model by the standard LS estimator, yielding $\widehat{\beta}$ and its standard error $\widehat{\sigma}_\beta$. (2) Use the model-selection procedure to determine whether the model is weakly identified. The model-selection statistic is defined in (5.1). If $t_n > \kappa_n$, the model is considered to be strongly identified and the standard critical value is used, i.e. $z_{1-\alpha/2}$ for a symmetric two-sided CI. If $t_n \leq \kappa_n$, we conclude that β is in an $n^{-1/2}$ neighborhood of 0 and π is weakly identified. In this case, continue to step 3. (3) Simulate the $1-\alpha$ quantile of the non-standard asymptotic distribution derived in Theorem 3.1(a) for a given h. (4) Take the supremum of the critical value obtained in step 3 over the parameter space of h. This is the critical value for the robust CI defined in (5.2).

Assumption R. $\kappa_n \to \infty$.

Theorem 5.1 *Suppose Assumptions* 1, 2, 3a, 4a, *and* R *hold. The nominal level* $1-\alpha$ *robust CI satisfies* $AsyCS = 1-\alpha$.

Comment: Theorem 5.1 states that the robust CI has correct asymptotic size provided the tuning parameter for model selection diverges to infinity with the sample size.

5.3 Simulations for the Robust CI

In this subsection, we first use simulation to demonstrate the finite-sample performance of the robust CI and to compare different choices of the tuning parameter κ_n. Second, we explain the good finite-sample performance of the robust CI by comparing the finite-sample quantiles of the test statistics with the asymptotic quantiles simulated from the local limit theory. Finally, different types of CIs are compared under various identification scenarios.

5.3.1 Finite-Sample Coverage Probability

The finite-sample sizes of the nominal 95% symmetric two-sided robust CIs in Example 1 and Example 2 are reported in Table 5.1 on the following page. We report the robust CIs with the tuning parameter κ_n being $(\ln n)^{1/2}$ and $(2\ln\ln n)^{1/2}$ and label them with Rob1 and Rob2, respectively.

Table 5.1: Finite-sample Sizes of Nominal 95% Symmetric Two-sided Robust CIs

	Std	$n = 100$			$n = 250$			$n = 500$		
Example 1. Box-Cox Function										
		Rob1	Rob2	LF	Rob1	Rob2	LF	Rob1	Rob2	LF
ζ_1	72.6	94.2	94.2	95.9	94.4	94.4	95.7	94.7	94.3	94.9
β	52.9	93.4	92.7	96.4	93.5	92.4	96.0	93.2	92.0	95.1
π	94.5	94.3	94.3	94.6	94.4	94.4	94.7	94.8	94.8	95.1
Example 2. Logistic Smooth Transition Function										
ζ_1	73.9	91.4	90.8	94.5	92.4	92.0	94.8	92.0	91.8	94.5
β	73.3	91.9	89.6	94.5	92.8	90.9	94.9	92.5	90.6	94.6
π	63.6	88.7	86.4	92.5	91.4	89.0	93.9	92.5	89.9	94.2

Note: Std is short for standard CI. Its asymptotic size is listed here for comparison.
Rob1 and Rob2 are the robust CIs with κ_n being $(\ln n)^{1/2}$ and $(2 \ln \ln n)^{1/2}$, respectively.
LF is the "least favorable" CI obtained without model selection.

We also report finite-sample results for a CI whose critical value is given by the least favorable (LF) asymptotic distribution. The LF CI does not employ the model-selection procedure and always takes the critical value that the robust CI would choose when the model is weakly identified. With a critical value large enough for all identification scenarios, the LF CI has slightly larger finite-sample sizes than those of the robust CIs. However, the LF CI can be significantly longer than the robust CI when the model is strongly identified.

In contrast to the standard CI and the subsampling CI, the robust CI has finite-sample size close to the nominal level. In Example 1, the robust CI improves the finite-sample sizes for ζ_1 and β when $n = 100$ from 75.1% and 53.7% to 94.2% and 93.4%, respectively. In Example 2, the robust CI improves the finite-sample sizes when $n = 250$ from 73.6%, 73.1%, and 62.6% to 92.4%, 92.8%, and 91.4%, respectively, for ζ_1, β, and π.[10]

The choice of tuning parameter between $(\ln n)^{1/2}$ and $(2 \ln \ln n)^{1/2}$ is a trade-off between the size and length of the CI. As evident in Table 5.1, Rob1 has a larger finite-sample size than that of Rob2. Hence, Rob1 is recommended if a good finite-sample size is the main focus, as is typically the case. However, Rob2 has a shorter average length than that of Rob1 because of its smaller tuning parameter. The length comparison is in Table 5.2 on page 20. Therefore, Rob2 might be chosen if a short length is valued more than a large coverage probability.

5.3.2 Comparison of Asymptotic and Finite-sample Quantile Graphs

The robust CI out-performs the standard CI by a large margin because under weak identification the local limit theory provides a good uniform approximation to the finite-sample distribution, while the standard normal distribution does not. To illustrate this point, finite-sample and as-

[10] The comparison is between the standard CI and the robust CI with $(\ln n)^{1/2}$ as the tuning parameter.

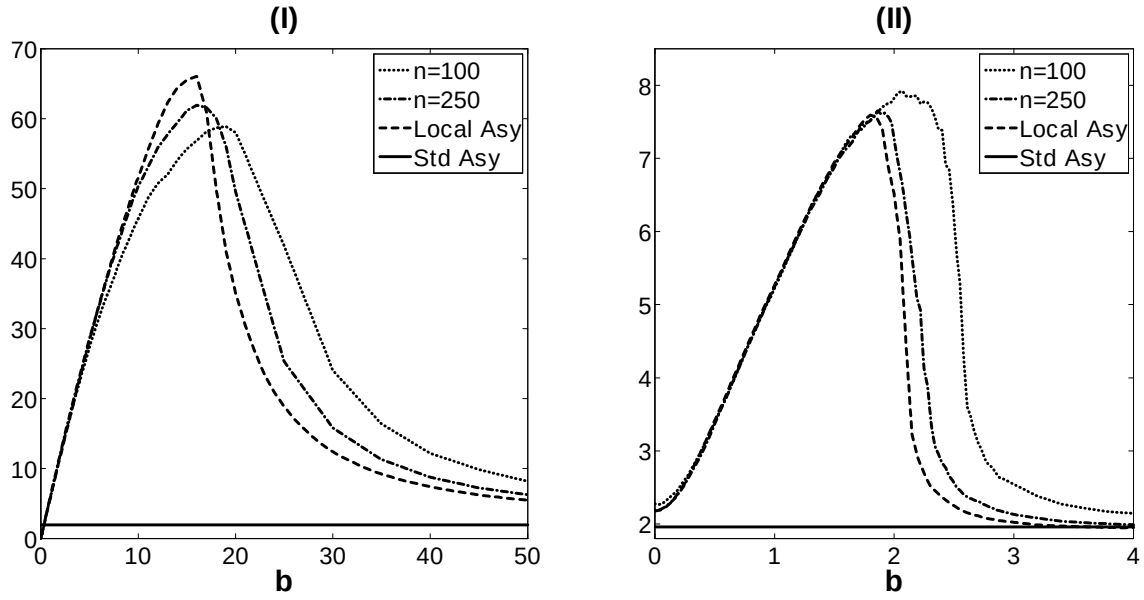

Figure 5.1: .95 Finite-sample and Asymptotic Quantiles of the Test Statistics for the Symmetric Two-sided CIs for β : (I) Example 1 with the Box-Cox Function and (II) Example 2 with the Logistic Function.

ymptotic quantiles simulated from the local limit theory in Theorem 3.1 are presented in the same figure for comparison. Figure 5.1 on the next page presents quantiles of the test statistics for the symmetric two-sided CIs for β. Because the true value of π affects the finite-sample and asymptotic distributions, we fix π_0 at 1.8 in Example 1 and at 4.8 in Example 2.[11] The true value of β is $b/n^{1/2}$ for the finite-sample simulations under different b values.

The finite-sample quantile, if used as the critical value, will lead to a CI with correct finite-sample size. It is clear in Figure 5.1 that quantiles from the local limit theory are closer to the finite-sample quantiles than quantiles from the standard asymptotic distribution are. This comparison explains the improvement of the local-limit-theory-based robust CI upon the standard CI.

5.3.3 Comparison of Different CIs

Next we compare the lengths and coverage probabilities of the standard and robust CIs under different true values of β that correspond to various strengths of identification. The purpose of this comparison is twofold. First, under weak identification, i.e. small β, the standard CI under-covers and the robust CI has good finite-sample coverage probabilities. This conclusion is based on comparing the finite-sample sizes of the standard CIs in Table 4.1 and those of the robust CIs in Table 5.1. In Table 5.2 on the following page, we explicitly provide the values of β to demonstrate the effect of identification strength. Second, we show that under strong identification the robust

[11] As shown in Figures 4.1 and 4.2, these are the π values that lead to the largest quantiles in most cases.

Table 5.2: Finite-sample Length and Coverage Probabilities of Nominal 95% Symmetric Two-sided CIs for β in Example 2 with the Logistic Function with Sample $n = 250$

	Std		Rob1		Rob2		LF	
β	L	CP	L	CP	L	CP	L	CP
0	0.15	91.0	0.58	96.6	0.54	91.0	0.59	100
0.05	0.17	73.3	0.61	96.9	0.53	94.3	0.65	100
0.10	0.20	82.9	0.58	94.4	0.45	93.6	0.77	96.5
0.20	0.23	93.7	0.29	95.9	0.25	95.0	0.87	98.7
0.40	0.23	95.1	0.23	95.1	0.23	95.1	0.89	100

Note: The explanations of Std, Rob1, Rob2, and LF are the same as in Table 5.1.
L denotes the length of the CI.
CP denotes the minimal coverage probability over the parameter space of π.

CI has similar length to that of the standard CI, both of which are much shorter than the LF CI obtained without model selection.

The CIs reported in Table 5.1 are symmetric two-sided CIs constructed in the smooth transition model in Example 2, with the true value of π being 4.8. The lengths and coverage probabilities are both averages over 10,000 simulation repetitions with sample size $n = 250$. As shown in Table 5.2, the standard CI under-covers severely when β is 0.05 and 0.1, while the BIC-type robust CI (Rob1) has coverage probabilities of 96.9% and 94.4%, respectively. Not surprisingly, the robust CI is three and a half times as long as the standard CI when β is 0.05 and almost three times as long as the latter when β is 0.1. As β gets larger, the robust CI and the standard CI get closer, both in terms of lengths and coverage probabilities. When β is 0.2 and 0.4, the LF CI is significantly longer than the robust CI and over-covers severely, manifesting the advantage of the robust CI.

6 General Model and Asymptotic Results

The general model defined in (2.1) allows for linear regressors as well as multiple nonlinear regressors to characterize different nonlinear relationships. Including more nonlinear regressors increases flexibility of the parametric nonlinear regression model. However, the asymptotic distributions of the LS estimators and test statistics are complicated in the general model. The reason is that the nonlinear regressors may have different strengths of identification. Assuming all of them are weakly identified leads to a very large CI that may not be informative. Thus, we aim to develop a general local limit theory that allows for different strengths of identification for different nonlinear regressors. A model selection procedure is applied to determine which nonlinear regressors involve weak identification. The difficulty raised by multiple strengths of identification is solved by a sequential procedure introduced below.

6.1 Asymptotic Distributions in a Model with Two Nonlinear Regressors

To understand the manner in which the results generalize from a model with one nonlinear regressor to a model with multiple nonlinear regressors, it helps to introduce the sequential procedure in a two-regressor model first. We do this in the present section. The model we consider here is

$$Y_i = g_1(X_i, \pi_1)\beta_1 + g_2(X_i, \pi_2)\beta_2 + U_i, \tag{6.1}$$

where $\beta_1,\ \beta_2 \in R$, $\beta = (\beta_1, \beta_2)'$ and $\pi = (\pi_1, \pi_2)'$. By introducing a second nonlinear regressor, we can examine the effect of β_1 on the asymptotic properties of $\widehat{\pi}_2$ and $\widehat{\beta}_2$. As in the simple one-regressor model, we consider drifting sequences of true parameters $\left(\beta_n', \pi_n'\right)'$, where $\beta_n = (\beta_{1n}, \beta_{2n})'$ and $\pi_n = (\pi_{1n}, \pi_{2n})'$. The drifting sequences of true parameters are characterized by the localization parameter

$$\begin{aligned} h \ &= \ (b, \pi_0)', \text{ where } b = (b_1, b_2)', \ \pi_0 = (\pi_{10}, \pi_{20})', \text{ and} \\ n^{1/2}\beta_{1n} \to b_1,\ n^{1/2}\beta_{2n} \to b_2,\ \pi_{1n} \to \pi_{10}\ , &\pi_{2n} \to \pi_{20}. \end{aligned} \tag{6.2}$$

The parameter space for h is $H = R_{[\pm\infty]} \times R_{[\pm\infty]} \times \Pi_1 \times \Pi_2$. Without loss of generality, we assume that β_{1n} converges to 0 slower than β_{2n} or at the same rate. We further assume that the limit of β_n/β_{1n} exists and call it $\Delta \in R^2$.

Given results obtained in the last section, we expect that the asymptotic distributions of the LS estimators depend on the magnitudes of both β_{1n} and β_{2n} relative to $n^{-1/2}$. Thus, we need to consider three cases: (I) Both β_{1n} and β_{2n} are $O(n^{-1/2})$. (II) Both β_{1n} and β_{2n} are of larger orders than $O(n^{-1/2})$. (III) β_{1n} is of a larger order than $O(n^{-1/2})$ and β_{2n} is $O(n^{-1/2})$. Intuitively, the consistency of $\widehat{\pi}_{jn}$ depends on the strength of the signal from $g_j(X_i, \pi_{jn})\beta_{jn}$, which is proportional to β_{jn}, for $j = 1$ and 2. Hence, $\widehat{\pi}_{1n}$ should be consistent as long as β_{1n} is big relative to $n^{-1/2}$, even if β_{2n} is small or even 0.

To develop this idea, we start with the concentrated sample criterion function $Q_n(\pi_1, \pi_2)$ obtained by concentrating out β_1 and β_2. In case (I), we show that after proper re-centering and re-scaling by n, the criterion function, indexed by (π_1, π_2), converges to a two-dimensional stochastic process analogous to the rhs of Lemma 3.1(a). Such a generalization from the one-dimensional case to the multiple-dimensional case can be carried out because the re-scaling parameter n does not depend on β_{1n} or β_{2n}. However, in cases (II) and (III), we need to derive results similar to Lemma 3.1(b), where either β_{1n} or β_{2n} is used as the re-scaling parameter.

To derive asymptotic results in cases (II) and (III), we need to view the minimization of the sample criterion function $Q_n(\pi_1, \pi_2)$ in a sequential way. First, for any given π_2, $Q_n(\pi_1, \pi_2)$ is a function of π_1. We write the criterion function when π_2 is fixed as $Q_n(\pi_1|\pi_2)$ and minimize $Q_n(\pi_1|\pi_2)$ over Π_1 to obtain $\widehat{\pi}_1(\pi_2)$, which depends on π_2. Then we plug $\widehat{\pi}_1(\pi_2)$ back into the

criterion function $Q_n(\pi_1, \pi_2)$ and estimate π_2 by minimizing $Q_n(\widehat{\pi}_1(\pi_2), \pi_2)$ over Π_2, yielding $\widehat{\pi}_2$. The LS estimator of (π_1, π_2) is $(\widehat{\pi}_1(\widehat{\pi}_2), \widehat{\pi}_2)$. Note that this sequential procedure is equivalent to minimizing $Q_n(\pi)$ over Π.

Before analyzing the criterion function $Q_n(\pi_1, \pi_2)$ sequentially, we first define some notation that is useful for partitioned regression. Let $g_2(\pi_2) = (g_2(X_1, \pi_2), ..., g_2(X_n, \pi_2))'$. Define an orthogonal projection and a population projection as $M_2(\pi_2) = I_n - g_2(\pi_2)\left(g_2(\pi_2)' g_2(\pi_2)\right)^{-1} g_2(\pi_2)$ and $\rho_2(\pi_1, \pi_2) = \left(Eg_2^2(X_i, \pi_2)\right)^{-1} Eg_2(X_i, \pi_2) g_1(X_i, \pi_1)$, respectively. When focusing on π_1, we first project out $g_2(\pi_2)$. The residuals after a population projection are written as $\widetilde{g}_{1,i}(\pi_1|\pi_2) = g_1(X_{i,}\pi_1) - g_2(X_i, \pi_2)\rho_2(\pi_1, \pi_2)$.

To characterize the random limits under weak identification, we let

$$\overline{S}(\pi_1, \pi_2, \pi_1) = (S_1(\pi_1), S_2(\pi_2), S_{\pi 1}(\pi_1)) \tag{6.3}$$

be a mean zero three-dimensional Gaussian process indexed by (π_1, π_2, π_1) with covariance kernel $\Omega(\pi; \overline{\pi}) = EU_i^2 s(X_i, \pi) s(X_i, \overline{\pi})'$, where $s(X_i, \pi) = (g_1(X_i, \pi_1), g_2(X_i, \pi_2), g_{\pi 1}(X_i, \pi_1))'$. Define the functions

$$\begin{aligned}
\Phi(\pi, \overline{\pi}) &= Eg(X_i, \pi) g(X_i, \overline{\pi})', \ \Phi_1(\pi_1, \overline{\pi}_1|\pi_2) = E\widetilde{g}_{1,i}(\pi_1|\pi_2)\widetilde{g}_{1,i}(\overline{\pi}_1|\pi_2), \\
\Phi_2(\pi_2) &= Es(X_i, \pi_{10}, \pi_2) s(X_i, \pi_{10}, \pi_2)', \ \Phi_{2s}(\pi_2, \overline{\pi}_2) = Es(X_i, \pi_{10}, \pi_2) g_2(X_i, \overline{\pi}_2).
\end{aligned} \tag{6.4}$$

Assumption 3b. $\lambda_{\min}(\Phi(\pi, \pi)) \geq \varepsilon$, $\Phi_1(\pi_1, \pi_1|\pi_2) \geq \varepsilon$, $\Phi_2(\pi_2) \geq \varepsilon$ $\forall \pi \in \Pi$ for some $\varepsilon > 0$.

Lemma 6.1 below establishes asymptotic properties of the concentrated sample criterion function $Q_n(\pi_1, \pi_2)$. It is a generalization of Lemma 3.1. When β_{2n} is of a smaller order than β_{1n} and they are not both $O(n^{-1/2})$, we need to analyze $Q_n(\pi_1, \pi_2)$ sequentially. Specifically, Lemma 6.1(b) provides the asymptotic distribution of $Q_n(\pi_1|\pi_2)$, where π_2 is fixed, and Lemma 6.1(c) and (d) provide the asymptotic distribution of $Q_n(\widehat{\pi}_1(\pi_2), \pi_2)$, where $\widehat{\pi}_1(\pi_2)$ is the optimal value of $\widehat{\pi}_1$ for a given value of π_2. Note that Lemma 6.1(c) and (d) are both sub-cases of Lemma 6.1(b). Lemma 6.1(a) and (e) are the two cases where we can analyze $Q_n(\pi)$ in one step.

Lemma 6.1 *Suppose Assumptions* 1, 2, *and* 3b *hold.*
(**a**) *When* $n^{1/2}\beta_n \to b \in R^2$,
$n\left(Q_n(\pi) - n^{-1}Y'Y\right) \Rightarrow -(S(\pi) + \Phi(\pi, \pi_0) b)' \Phi^{-1}(\pi, \pi)(S(\pi) + \Phi(\pi, \pi_0) b)$,
where $S(\pi) = (S_1(\pi_1), S_2(\pi_2))'$.

(**b**) *When* $\left|n^{1/2}\beta_{1n}\right| \to \infty$ *and* $\beta_{2n} = o(\beta_{1n})$,
$\beta_{1n}^{-2}\left(Q_n(\pi_1|\pi_2) - n^{-1}Y'M_2(\pi_2)Y\right) \to_p -\Phi_1^{-1}(\pi_1, \pi_1|\pi_2)\Phi_1^2(\pi_1, \pi_{10}|\pi_2)$ *uniformly over* $\Pi_1 \times \Pi_2$.
$\sup_{\pi_2 \in \Pi_2} |\widehat{\pi}_1(\pi_2) - \pi_{1n}| \to_p 0$.

(**c**) *When* $\left|n^{1/2}\beta_{1n}\right| \to \infty$ *and* $n^{1/2}\beta_{2n} \to b_2 \in R$,

$n\left(Q_n\left(\widehat{\pi}_1\left(\pi_2\right),\pi_2\right)-n^{-1}\left(Y-g_1\left(\pi_{1n}\right)\beta_{1n}\right)'\left(Y-g_1\left(\pi_{1n}\right)\beta_{1n}\right)\right)\Rightarrow$
$-\left(\overline{S}\left(\pi_2\right)+\Phi_{2s}\left(\pi_2,\pi_{20}\right)b_2\right)'\Phi_2^{-1}\left(\pi_2\right)\left(\overline{S}(\pi_2)+\Phi_{2s}\left(\pi_2,\pi_{20}\right)b_2\right)$, *where* $\overline{S}\left(\pi_2\right)=\overline{S}\left(\pi_{10},\pi_2,\pi_{10}\right)$.

(**d**) *When* $\left|n^{1/2}\beta_{1n}\right|\to\infty$, $\left|n^{1/2}\beta_{2n}\right|\to\infty$, *and* $\beta_{2n}=o\left(\beta_{1n}\right)$,
$\beta_{2n}^{-2}\left(Q_n\left(\widehat{\pi}_1\left(\pi_2\right),\pi_2\right)-n^{-1}\left(Y-g_1\left(\pi_{1n}\right)\beta_{1n}\right)'\left(Y-g_1\left(\pi_{1n}\right)\beta_{1n}\right)\right)\to_p$
$-\Phi_{2s}\left(\pi_2,\pi_{20}\right)'\Phi_2^{-1}\left(\pi_2\right)\Phi_{2s}\left(\pi_2,\pi_{20}\right)$ *uniformly over* Π_2.

(**e**) *When* $\left|n^{1/2}\beta_{1n}\right|\to\infty$, $\left|n^{1/2}\beta_{2n}\right|\to\infty$, $\beta_{1n}=O\left(\beta_{2n}\right)$ *and* $\beta_{2n}=O\left(\beta_{1n}\right)$,
$\beta_{1n}^{-2}\left(Q_n(\pi)-n^{-1}Y'Y\right)\Rightarrow-\Delta'\Phi\left(\pi,\pi_0\right)'\Phi^{-1}\left(\pi,\pi\right)\Phi\left(\pi,\pi_0\right)\Delta$, *uniformly over* Π,
where $\Delta=\lim_{n\to\infty}\left(\beta_n/\beta_{1n}\right)$.

Comments: 1. Lemma 6.1(a) shows that when both β_{1n} and β_{2n} are $O\left(n^{-1/2}\right)$, we can analyze $Q_n\left(\pi_1,\pi_2\right)$ in one step because its rate of convergence is n^{-1}, independent of β_{1n} or β_{2n}. However, when either of β_{1n} and β_{2n} is of a larger order than $O(n^{-1/2})$, we need to analyze $Q_n\left(\pi_1,\pi_2\right)$ sequentially in order to get a non-degenerate limit. The non-degenerate limit is necessary for the purpose of deriving consistency properties of $\widehat{\pi}_{1n}$ and $\widehat{\pi}_{2n}$.

2. In Lemma 6.1(b), β_{1n} is of a larger order than β_{2n}. As π_2 is fixed, the asymptotic properties of $Q_n\left(\pi_1|\pi_2\right)$ depend on the signal strength from $g_1\left(X_i,\pi_{1n}\right)\beta_{1n}$, which is determined by the magnitude of β_{1n} relative to $n^{-1/2}$. Because β_{1n} is larger than $O(n^{-1/2})$, $Q_n\left(\pi_1|\pi_2\right)$ has a non-random limit after re-scaling by β_{1n}^{-2}, as in Lemma 3.1(b). For any fixed π_2, this non-random limit is uniquely minimized at π_{10} by the Cauchy-Schwarz inequality. As a result, $\widehat{\pi}_1\left(\pi_2\right)$ is consistent uniformly over Π_2. The uniform consistency is obtained under the assumption that $\beta_{2n}=o(\beta_{1n})$.

3. Lemma 6.1(c) and (d) are both sub-cases of Lemma 6.1(b). When plugged into $Q_n\left(\pi_1,\pi_2\right)$, $\widehat{\pi}_1\left(\pi_2\right)$ becomes the second channel through which π_2 enters the criterion function. This second effect is taken into account by including $g_1(X_i,\pi_1)$ and $g_{\pi 1}\left(X_i,\pi_1\right)$ in the vector $s\left(X_i,\pi\right)$. This vector is a key element in defining the Gaussian processes $\overline{S}\left(\pi_1,\pi_2,\pi_1\right)$ and the function $\Phi_{2s}\left(\pi_2,\overline{\pi}_2\right)$. Because of the uniform consistency of $\widehat{\pi}_1\left(\pi_2\right)$, minimization of $Q_n\left(\widehat{\pi}_1\left(\pi_2\right),\pi_2\right)$ over Π_2 is analogous to a problem with the nonlinear regressor $g_2\left(X_i,\pi_2\right)$ alone. Lemma 6.1(c) and (d) are comparable to Lemma 3.1(a) and (b), respectively.

4. In Lemma 6.1(e), we analyze the criterion function in one step as in Lemma 6.1(a). The same rate of convergence guarantees that all elements of Δ are finite and different from 0.

The next lemma provides asymptotic limits of the LS estimator $\widehat{\pi}_n$. They are obtained by minimizing the concentrated sample criterion functions and their limits in Lemma 6.1.

Lemma 6.2 *Suppose Assumptions* 1, 2, *and* 3b *hold.*
(**a**) *When* $n^{1/2}\beta_n\to b\in R^2$, $\widehat{\pi}_n\Rightarrow\pi^*\left(h\right)$, *where*
$\pi^*\left(h\right)=\underset{\pi\in\Pi}{\arg\min}\left(-\{S\left(\pi\right)+\Phi\left(\pi,\pi_0\right)b'\Phi^{-1}\left(\pi,\pi\right)S\left(\pi\right)+\Phi\left(\pi,\pi_0\right)b\}\right)$.
(**b**) *When* $\left|n^{1/2}\beta_{1n}\right|\to\infty$, $\widehat{\pi}_{1n}-\pi_{1n}\to_p 0$.

(**c**) *When* $\left|n^{1/2}\beta_{1n}\right| \to \infty$ *and* $n^{1/2}\beta_{2n} \to b_2 \in R$, $\widehat{\pi}_{2n} \Rightarrow \pi_2^*(h)$, *where*
$\pi_2^*(h) = \underset{\pi_2 \in \Pi_2}{\arg\min}\ (-\{\left(\overline{S}(\pi_2) + \Phi_{2s}(\pi_2, \pi_{20})\, b_2\right)' \Phi_2^{-1}(\pi_2) \left(\overline{S}(\pi_2) + \Phi_{2s}(\pi_2, \pi_{20})\, b_2\right)\}).$

(**d**) *When* $\left|n^{1/2}\beta_{1n}\right| \to \infty$ *and* $\left|n^{1/2}\beta_{2n}\right| \to \infty$, $\widehat{\pi}_{2n} - \pi_{2n} \to_p 0$.

Comments: 1. Assumptions on uniqueness of $\pi^*(h)$ and $\pi_2^*(h)$ are presented in the next subsection in a general set-up.

2. When both β_{1n} and β_{2n} are $O(n^{-1/2})$, neither π_{1n} nor π_{2n} can be estimated consistently. Both of their LS estimators converge to random variables, whose asymptotic limits jointly minimize the sample paths of a non-central chi-square process.

3. The consistency of $\widehat{\pi}_{1n}$ only depends on the magnitude of β_{1n}. It is not affected by the magnitude of β_{2n} or whether $\widehat{\pi}_{2n}$ is consistent. This is consistent with the intuition obtained from the simple model, in which we show that $\widehat{\pi}_{1n}$ is consistent provided the the signal from $g_1(X_i, \pi_{1n})\beta_{1n}$ is stronger than the noise from the errors.

4. Lemma 6.2 (c) and (d) are sub-cases of Lemma 6.2 (b). Lemma 6.2(d) corresponds to both Lemma 6.1(d) and (e). Although the criterion function has different asymptotic distributions in these two cases, the LS estimator $\widehat{\pi}_{2n}$ is consistent as long as β_{2n} is larger than $O(n^{-1/2})$.

Lemma 6.2 provides conditions for consistency of $\widehat{\pi}_n$ and its random limits in the absence of consistency. Lemma 6.3 below derives the asymptotic distributions of the LS estimators $\widehat{\beta}_n$ and $\widehat{\pi}_n$ when they are consistent. The asymptotic covariance matrices $G(\pi), V(\pi)$, and $\Sigma(\pi)$ are the same as in the simple case, with the adjustment that $m_i(\pi) = (g(X_i, \pi)', g_\pi(X_i, \pi)')'$.

Assumption 4b. $\lambda_{\min}(G(\pi)) \geq \varepsilon\ \forall\ \pi \in \Pi$ for some $\varepsilon > 0$.

Lemma 6.3 *Suppose Assumptions* 1, 2, 3b, *and* 4b *hold.*
(**a**) *When* $n^{1/2}\beta_n \to b \in R^2$,

$$n^{1/2}\left(\widehat{\beta}_n - \beta_n\right) \Rightarrow \tau(\pi^*(h), h), \text{ where } \tau(\pi, h) = \Phi^{-1}(\pi, \pi)\left(S(\pi) + \Phi(\pi, \pi_0)\, b\right) - b.$$

(**b**) *When* $\left|n^{1/2}\beta_{1n}\right| \to \infty$ *and* $n^{1/2}\beta_{2n} \to b_2 \in R$,

$$\begin{pmatrix} n^{1/2}\left(\widehat{\beta}_{1n} - \beta_{1n}\right) \\ n^{1/2}\left(\widehat{\beta}_{2n} - \beta_{2n}\right) \\ n^{1/2}\beta_{1n}\left(\widehat{\pi}_{1n} - \pi_{1n}\right) \end{pmatrix} \Rightarrow \tau_2(\pi_2^*(h), h), \text{ where}$$

$$\tau_2(\pi_2) = \Phi_2^{-1}(\pi_2)\left(\overline{S}(\pi_2) + \Phi_{2s}(\pi_2, \pi_{20})\, b_2\right) - \iota_2 b_2 \text{ and } \iota_2 = (0, 1, 0)'.$$

(**c**) *When* $\left|n^{1/2}\beta_{1n}\right| \to \infty$ *and* $\left|n^{1/2}\beta_{2n}\right| \to \infty$,

$$\begin{pmatrix} n^{1/2}\left(\widehat{\beta}_n - \beta_n\right) \\ n^{1/2}\beta_{1n}\left(\widehat{\pi}_{1n} - \pi_{1n}\right) \\ n^{1/2}\beta_{2n}\left(\widehat{\pi}_{2n} - \pi_{2n}\right) \end{pmatrix} \Rightarrow N\left(0, \Sigma\left(\pi_0\right)\right).$$

Comments: 1. When b is finite, the limit of $n^{1/2}(\widehat{\beta}_n - \beta_n)$ is non-standard and is characterized by the Gaussian process $\tau\left(\pi, h\right)$ and the random variable $\pi^*\left(h\right)$. This is the same as Lemma 3.4 generalized to a vector case.

2. In Lemma 6.3(b), for fixed π_2, $(\widehat{\beta}_1\left(\pi_2\right), \widehat{\beta}_2\left(\pi_2\right), \widehat{\pi}_1\left(\pi_2\right))'$ has an asymptotic normal distribution whose variance depends on π_2. Due to the inconsistency of $\widehat{\pi}_{2n}$, the asymptotic distribution of $(\widehat{\beta}_{1n}, \widehat{\beta}_{2n}, \widehat{\pi}_{1n})'$ is characterized by the three-dimensional Gaussian process $\tau_2\left(\pi_2\right)$ and the random variable $\pi_2^*\left(h\right)$, both of which are continuous functions of the Gaussian process $\overline{S}\left(\pi_2\right)$. A special case is when $g_2\left(X_i, \pi_2\right)$ is uncorrelated with $g_1\left(X_i, \pi_1\right)$ and $g_{\pi 1}\left(X_i, \pi_1\right)$ for any π_1 and π_2. In this situation, asymptotic distributions of $\widehat{\beta}_{1n}\left(\pi_2\right)$ and $\widehat{\pi}_{1n}\left(\pi_2\right)$ do not depend on π_2. Therefore, $\widehat{\beta}_{1n}$ and $\widehat{\pi}_{1n}$ have asymptotic normal distributions despite the non-standard behaviors of $\widehat{\beta}_{2n}$ and $\widehat{\pi}_{2n}$.

3. Finally, if both b_1 and b_2 are infinite, as in Lemma 6.3(c), all parameters can be estimated consistently and have asymptotic normal distributions. As shown in the simple model, the rate of convergence of $\widehat{\pi}_{jn}$ depends on β_{jn}, for $j = 1$ and 2. The general rule is that the faster β_{jn} converges to 0 the slower is the convergence rate of $\widehat{\pi}_{jn}$.

With two nonlinear regressors in the model, we find that $\widehat{\beta}_{1n}$ is always consistent but $\widehat{\pi}_{1n}$ is consistent if and only if β_{1n} is larger than $O(n^{-1/2})$. The asymptotic distribution of $(\widehat{\beta}_{1n}, \widehat{\pi}_{1n})$ depends on the convergence rates of both β_{1n} and β_{2n}. Analogous results apply to $\widehat{\beta}_{2n}$ and $\widehat{\pi}_{2n}$.

Lemmas 6.1 and 6.3 indicate whether we can analyze $\widehat{\pi}_{1n}$ and $\widehat{\pi}_{2n}$ together when deriving their asymptotic properties. First, when both β_{1n} and β_{2n} are $O\left(n^{-1/2}\right)$, we can always put π_1 and π_2 together and generalize the asymptotic results obtained in the single-regressor model to the multi-regressor model. Second, when only β_{2n} is $O\left(n^{-1/2}\right)$, the sequential procedure is needed because $\widehat{\pi}_{1n}$ is consistent but $\widehat{\pi}_{2n}$ is inconsistent. Finally, if both β_{1n} and β_{2n} are larger than $O\left(n^{-1/2}\right)$, we need the sequential procedure for consistency results if β_{1n} and β_{2n} have different orders. However, we can analyze their asymptotic distributions together because both $\widehat{\pi}_{1n}$ and $\widehat{\pi}_{2n}$ have asymptotic normal distributions. The convergence rate of $\widehat{\pi}_{jn}$ is proportional to β_{jn} for $j = 1$ and 2. These general rules will guide us in determining asymptotic results in a general model with an arbitrary number of nonlinear regressors in addition to some linear regressors. For simplicity, we leave the asymptotic distributions of the test statistics as special cases of the general results given in the next subsection.

6.2 Asymptotic Distributions in the General Model

We consider the general model including multiple nonlinear regressors as well as linear regressors in the present section. The general model, discussed in (2.1) already, takes the form

$$Y_i = g(X_i, \pi)'\beta + Z_i'\zeta + U_i \text{ for } i = 1, ..., n, \tag{6.5}$$

where $g(X_i, \pi)$ and β are both $p \times 1$ dimensional vectors. As in the simple model, we consider the asymptotic distributions of the LS estimators and test statistics along drifting sequences $(\beta_n', \pi_n', \zeta_n)'$ characterized by the localization parameter

$$h = \left(b', \pi_0'\right)', \text{ where } n^{1/2}\beta_n \to b \text{ and } \pi_n \to \pi_0. \tag{6.6}$$

The parameter space for h is $H = R^p_{[\pm\infty]} \times \Pi$, where $R^p_{[\pm\infty]} = \{(x_1, ..., x_p) : x_j \in R \cup \{\pm\infty\} \text{ for } j = 1, ..., p\}$. The finite-sample and asymptotic distributions of the test statistics are invariant to ζ_n. Let $\beta_{j,n}$ denote the j^{th} element of β_n. Without loss of generality, we also assume that the order of $\beta_{j,n}$ is larger than or equal to that of $\beta_{j',n}$ $\forall j < j'$. In other words, $\beta_{j',n} = O(\beta_{j,n})$ $\forall j < j'$.

Results obtained in the two-regressor model indicate that asymptotic properties of the LS estimators depend on the magnitude of β_n. Thus, we first group the coefficients $\beta_{j,n}$, for $j = 1, ..., p$, based on their orders. The grouping rule is summarized as follows. (1) All $\beta_{j,n}$ that are $O(n^{-1/2})$ are put in the last group. (2) If $\beta_{j,n} \gg O(n^{-1/2})$, the following rule applies: $\forall k < k'$, parameters in group k converge to 0 slower than those in group k' and parameters in the same group converge to 0 at the same rate. Hence, if $\beta_{j,n}$ is bounded away from 0, it is put in the first group.

Let $\xi_{k,n}$ be the p_k dimensional sub-vector of β_n that represents the k^{th} group. Suppose $\beta_{j,n} \in \xi_{k,n}$, $\beta_{j',n} \in \xi_{k',n}$ and $k < k'$. According to the grouping rule, $\beta_{j',n} = o(\beta_{j,n})$. Suppose there are K groups in total and p_k elements in each group for $k = 1, ..., K$, then $\sum_{k=1}^K p_k = p$ and $\beta_n = \left(\xi_{1,n}', ..., \xi_{K,n}'\right)'$. Here is an example to illustrate the grouping rule. Suppose $\beta_n = (1, n^{-1/4}, n^{-1/3}, 2n^{-1/3}, n^{-1/2}, n^{-1})$. The above grouping rule gives $\xi_{1,n} = 1$, $\xi_{2,n} = n^{-1/4}$, $\xi_{3,n} = (n^{-1/3}, 2n^{-1/3})$, and $\xi_{4,n} = (n^{-1/2}, n^{-1})$.

Note that the group index k for $\beta_{j,n}$ is a property associated with the drifting sequence $\{\beta_{j,n} : n = 1, 2, ...\}$ and therefore does not change with sample size n. Hence, we suppress the subscript n unless it is used to denote the true values with sample size n. Let $\xi_k = (\beta_{k_1}, ..., \beta_{k_{p_k}})'$, where k_1 to k_{p_k} are the indexes of the elements of β that belong to the k^{th} group. The parameters, parameter spaces, regressors, and their derivatives associated with group k are written as

$$\begin{aligned} \psi_k &= (\pi_{k_1}, ..., \pi_{k_{p_k}})', \ \Psi_k = \Pi_{k_1} \times \cdots \times \Pi_{k_{p_k}}, \\ f_k(\psi_k) &= [g_{k_1}(\pi_{k_1}), \cdots, g_{k_{p_k}}(\pi_{k_{p_k}})], \ f_{\psi k}(\psi_k) = [g_{\pi k_1}(\pi_{k_1}), \cdots, g_{\pi k_{p_k}}(\pi_{k_{p_k}})], \text{ where} \\ g_j(\pi_j) &= (g_j(X_1, \pi_j), ..., g_j(X_n, \pi_j))' \text{ and } g_{\pi j}(\pi_j) = (g_{\pi j}(X_1, \pi_j), ..., g_{\pi j}(X_n, \pi_j))'. \end{aligned} \tag{6.7}$$

Employing the group set-up, we use ξ, ψ, and $f(\cdot)$ to replace β, π, and $g(\cdot)$, respectively. Let $\xi_{k,n}$ and $\psi_{k,n}$ be the true values of ξ_k and ψ_k when the sample size is n, and $\psi_{k,0}$ be the limit of $\psi_{k,n}$ as n goes to infinity. We assume that the limit of $\xi_{k,n}/\beta_{k_1,n}$ exists and call it $\Delta_k \in R^{p_k}$, where $\beta_{k_1,n}$ is the first element of $\xi_{k,n}$. The localization parameter h defined in (6.6) is equivalent to

$$\begin{aligned} h &= \left(b', \pi_0'\right)', \text{ where } b = \left(b_1', ..., b_K'\right)', \ \pi_0 = \left(\psi_{1,0}', ..., \psi_{K,0}'\right)', \\ n^{1/2}\xi_{k,n} &\to b_k \text{ and } \psi_{k,n} \to \psi_{k,0}, \text{ for } k = 1, ..., K. \end{aligned} \tag{6.8}$$

Note that according to the grouping rule, $b_k \in \{\pm\infty\}$ for $k < K$ and $b_K \in R^{p_K} \cup \{\pm\infty\}$. After grouping, the model is written in matrix notation as

$$Y = Z\zeta + f_1(\psi_1)\xi_1 + ... + f_K(\psi_K)\xi_K + U. \tag{6.9}$$

As in the two-regressor model, minimization of the concentrated sample criterion function $Q_n(\pi)$ defined in (2.2) can be viewed in a sequential way. Define $\psi_{k^-} = (\psi_1', ..., \psi_{(k-1)}')'$ and $\psi_{k^+} = (\psi_{(k+1)}', ..., \psi_K')'$. Then $\pi = (\psi_{k^-}', \psi_k', \psi_{k^+}')'$. Let $\psi_{k^-,n}$ and $\psi_{k^+,n}$ be the true values of ψ_{k^-} and ψ_{k^+} when the sample size is n, and $\psi_{k^-,0}$ and $\psi_{k^+,0}$ be the limits of $\psi_{k^-,n}$ and $\psi_{k^+,n}$. For fixed ψ_{k^+}, the sample criterion function $Q_n(\pi)$ is indexed by $(\psi_{k^-}', \psi_k')'$ and now is written as $Q_n(\psi_{k^-}, \psi_k|\psi_{k^+})$. Whenever ψ_{1^-} and ψ_{K^+} are involved, they are omitted.

For $k = 1$, let $\widehat{\psi}_1\left(\psi_{1+}\right) = \arg\min_{\psi_1 \in \Psi_1} Q_n(\psi_1|\psi_{1^+})$. For $k = 2$, we plug $\widehat{\psi}_1(\psi_{1^+})$ into $Q_n(\psi_1, \psi_2|\psi_{2^+})$ and get $\widehat{\psi}_2(\psi_{2^+}) = \arg\min_{\psi_2 \in \Psi_2} Q_n(\widehat{\psi}_1(\psi_2, \psi_{2^+}), \psi_2|\psi_{2^+})$, where $\widehat{\psi}_1(\psi_2, \psi_{2^+})$ is a second channel for ψ_2 to enter the criterion function. Now the LS estimator of ψ_1 given ψ_{2^+} becomes $\widehat{\psi}_1(\psi_{2^+}) = \widehat{\psi}_1(\widehat{\psi}_2(\psi_{2^+}), \psi_{2^+})$. Continuing the above procedure sequentially, we get $\widehat{\psi}_k(\psi_{k^+}) = \arg\min_{\psi_k \in \Psi_k} Q_n(\widehat{\psi}_{k^-}(\psi_k, \psi_{k^+}), \psi_k, |\psi_{k^+})$, for $k = 2$ to $k-1$, and finally $\widehat{\psi}_K = \arg\min_{\psi_K \in \Psi_K} Q_n(\widehat{\psi}_{K^-}(\psi_K), \psi_K)$. The last step is to plug $\widehat{\psi}_K$ back into $\widehat{\psi}_k(\psi_{k^+})$ for $k = K-1$ and sequentially obtain $\widehat{\psi}_{(K-1)} = \widehat{\psi}_{(K-1)}(\widehat{\psi}_K), ..., \widehat{\psi}_1 = \widehat{\psi}_1(\widehat{\psi}_2, ..., \widehat{\psi}_K)$. Note that this sequential procedure is equivalent to minimizing $Q_n(\pi)$ over Π. For notational simplicity, $Q_n(\psi_K)$ stands for $Q_n(\widehat{\psi}_{K^-}(\psi_K), \psi_K)$ hereafter.

Analogous to ψ_{k^-} and ψ_{k^+}, we define $\xi_{k^-} = (\zeta', \xi_1', ..., \xi_{(k-1)}')'$ and $\xi_{k^+} = (\xi_{(k+1)}', ..., \xi_K')'$. Note that we put ζ in ξ_{k^-} in order to analyze ζ in the first step during the above sequential procedure. To analyze the criterion function sequentially, we also define

$$\begin{aligned} &f_{k^-}(\psi_{k^-}) = [Z, f_1(\psi_1), \cdots, f_{(k-1)}(\psi_{(k-1)})], \ f_{k^+}(\psi_{k^+}) = [f_{(k+1)}(\psi_{(k+1)}), \cdots, f_K(\psi_K)], \\ &f_{\psi k^-}(\psi_{k^-}) = [f_{\psi 1}(\psi_1), \cdots, f_{\psi(k-1)}(\psi_{(k-1)})], \ s_k(\psi_{k^-}, \psi_k) = [f_{k^-}(\psi_{k^-}), f_k(\psi_k), f_{\psi k^-}(\psi_{k^-})]. \end{aligned} \tag{6.10}$$

The linear regressor Z is put in $f_{k^-}(\psi_{k^-})$, corresponding to ζ in ξ_{k^-}.

When focusing on ψ_k, we fix ψ_{k^+} and project out $f_{k^+}(\psi_{k^+})$ using $M_{k^+}(\psi_{k^+})$, where $M_{k^+}(\psi_{k^+}) = M(f_{k^+}(\psi_{k^+}))$ and the function $M(X) = I_n - X(X'X)^{-1}X$ is an orthogonal projection matrix

for any X. Let the subscript i stands for the ith row of a matrix written as a column vector. The corresponding population projection coefficient and the projection residual are

$$\begin{aligned}\rho_{k^+}\left(\psi_{k^-},\psi_k,\psi_{k^+}\right) &= \left(Ef_{k^+,i}\left(\psi_{k^+}\right)f_{k^+,i}\left(\psi_{k^+}\right)'\right)^{-1}Ef_{k^+,i}\left(\psi_{k^+}\right)s_{k,i}\left(\psi_{k^-},\psi_k|\psi_{k^+}\right) \text{ and}\\ \widetilde{s}_k\left(\psi_{k^-},\psi_k|\psi_{k^+}\right) &= s_k\left(\psi_{k^-},\psi_k\right)-f_{k^+}\left(\psi_{k^+}\right)\rho_{k^+}\left(\psi_{k^-},\psi_k,\psi_{k^+}\right). \end{aligned} \tag{6.11}$$

Define the functions

$$\begin{aligned}\Phi_k\left(\psi_k|\psi_{k^+}\right) &= E\widetilde{s}_{k,i}\left(\psi_{k^-,0},\psi_k|\psi_{k^+}\right)\widetilde{s}_{k,i}\left(\psi_{k^-,0},\psi_k|\psi_{k^+}\right)' \text{ and}\\ \Phi_{ks}\left(\psi_k,\overline{\psi}_k|\psi_{k^+}\right) &= E\widetilde{s}_{k,i}\left(\psi_{k^-,0},\psi_k|\psi_{k^+}\right)f_{k,i}\left(\overline{\psi}_k\right)'. \end{aligned} \tag{6.12}$$

Let $S\left(\psi_{K^-},\psi_K,\psi_{K^-}\right)$ be a $(q+2p-p_K)$ dimensional Gaussian process indexed by $(\psi_{K^-},\psi_K,\psi_{K^-})$ with covariance kernel

$$\Omega\left(\psi_{K^-},\psi_K,\psi_{K^-};\overline{\psi}_{K^-},\overline{\psi}_K,\overline{\psi}_{K^-}\right)=EU_i^2s_{K,i}\left(\psi_{K^-},\psi_K\right)s_{K,i}\left(\overline{\psi}_{K^-},\overline{\psi}_K\right)'. \tag{6.13}$$

For notational simplicity, we write $S\left(\psi_K\right)=S\left(\psi_{K^-,0},\psi_K,\psi_{K^-,0}\right)$.

Assumption 3c. $\lambda_{\min}(\Phi_k\left(\psi_k|\psi_{k^+}\right))\geq\varepsilon\ \forall\ \psi_k\in\Psi_k$ and $\psi_{k^+}\in\Psi_{k^+}$ for $k=1,...,K$, where Ψ_{k^+} is the parameter space for ψ_{k^+}.

Lemma 6.4 *Suppose Assumptions* 1, 2, *and* 3c *hold.*
(**a**) *For* $k=1,...,K-1$,

$$\begin{aligned}&\beta_{k_1,n}^{-2}(Q_n(\widehat{\psi}_{k^-}\left(\psi_k,\psi_{k^+}\right),\psi_k|\psi_{k^+})-n^{-1}(Y-f_{k^-}\left(\psi_{k^-,n}\right)\xi_{k^-,n})'M_{k^+}\left(\psi_{k^+}\right)(Y-f_{k^-}\left(\psi_{k^-,n}\right)\xi_{k^-,n}))\\ &\to_p-\Delta_k'\Phi_{ks}\left(\psi_k,\psi_{k,0}|\psi_{k^+}\right)'\Phi_k^{-1}\left(\psi_k|\psi_{k^+}\right)\Phi_{ks}\left(\psi_k,\psi_{k,0}|\psi_{k^+}\right)\Delta_k,\end{aligned}$$

uniformly over $\Psi_k\times\Psi_{k^+}$, *where* $\beta_{k_1,n}$ *is the first element of* $\xi_{k,n}$, $\widehat{\psi}_{k^-}\left(\psi_k,\psi_{k^+}\right)$ *is the LS estimator of* ψ_{k^-} *given* (ψ_k,ψ_{k^+}), $\Delta_k=\lim_{n\to\infty}\left(\xi_{k,n}/\beta_{k_1,n}\right)\in R^{p_k}$.
(**b**) *For* $k=1,...,K-1$,

$$\sup_{\psi_{k^+}\in\Psi_{k^+}}\left|\widehat{\psi}_k\left(\psi_{k^+}\right)-\psi_{k,n}\right|\to_p 0.$$

(**c**) *When* $\left\|n^{1/2}\xi_{K,n}\right\|\to\infty$, *the results in* (a) *and* (b) *also apply to* $k=K$, *with* ψ_{K^+} *omitted.*
(**d**) *When* $n^{1/2}\xi_{K,n}\to b_K\in R^{p_K}$,

$$\begin{aligned}&n\left(Q_n(\psi_K)-n^{-1}\left(Y-f_{K^-}\left(\psi_{K^-,n}\right)\xi_K\right)'\left(Y-f_{K^-}\left(\psi_{K^-,n}\right)\xi_K\right)\right)\\ \Rightarrow\ &-\left(S\left(\psi_K\right)+\Phi_{Ks}\left(\psi_K,\psi_{K,0}\right)b_K\right)'\Phi_K^{-1}\left(\psi_K\right)\left(S\left(\psi_K\right)+\Phi_{Ks}\left(\psi_K,\psi_{K,0}\right)b_K\right).\end{aligned}$$

Comments: 1. This is a generalization of Lemmas 3.1 and 6.1. The concentrated sample criterion function $Q_n(\widehat{\psi}_{k^-}\left(\psi_k,\psi_{k^+},\right),\psi_k|\psi_{k^+})$ is defined by fixing ψ_{k^+} and taking $\widehat{\psi}_{k^-}$ at the optimal value

given ψ_k and ψ_{k^+}. When ξ_k is larger than $O(n^{-1/2})$, this sample criterion function has a non-random limit after suitable re-centering and re-scaling, as in Lemma 3.1(b). The convergence rates are different across groups, but they are all slower than n^{-1}. Because Δ_k is finite and bounded away from 0, the rhs of Lemma 6.4(a) is uniquely minimized at $\psi_{k,0}$ according to a vector Cauchy-Schwarz inequality. This leads to uniform consistency of $\widehat{\psi}_k(\psi_{k^+})$ over Ψ_{k^+}.

2. If $\xi_{K,n}$ is $O(n^{-1/2})$, the signal from $f_K(\psi_{K,n})\xi_{K,n}$ is not stronger than the noise from the errors. In consequence, $Q_n(\psi_K)$ converges to a non-central chi-square process after suitable re-centering and re-scaling, as in Lemma 3.1(a), Lemma 6.1(a), and Lemma 6.1(c).

Let $Q(\psi_K)$ denote the chi-square process on the rhs of Lemma 6.4(d). Assumption U below ensures that the argmin function of $Q(\psi_K)$ is continuous. This assumption is verified in the simple model by Lemma 3.2.

Assumption U. The sample paths of the non-central chi-square process $Q(\psi_K)$ have unique minima over $\psi_K \in \Psi_K$ with probability one.

Lemma 6.5 *Suppose Assumptions* 1, 2, 3c, *and* U *hold.*
(**a**) *When* $n^{1/2}\xi_{K,n} \to b_K \in R^{p_K}$,

$$\begin{aligned}
&\widehat{\psi}_{k,n} - \psi_{k,n} \to_p 0 \ \forall k \leq K-1,\ \widehat{\psi}_{K,n} \Rightarrow \psi_K^*(h), \\
&\widehat{\pi}_n = (\widehat{\psi}'_{1,n}, ..., \widehat{\psi}'_{K,n})' \Rightarrow \pi^*(h), \ \textit{where} \\
&\psi_K^*(h) = \underset{\psi_K \in \Psi_K}{\arg\min}\, Q(\psi_K) \ \textit{and}\ \pi^*(h) = (\psi'_{1,0}, ..., \psi'_{(K-1),0}, \psi_K^*(h)')'.
\end{aligned}$$

(**b**) *When* $\left\|n^{1/2}\xi_{Kn}\right\| \to \infty$, $\widehat{\psi}_{k,n} - \psi_{k,n} \to_p 0\ \forall\ k \leq K$ *and* $\widehat{\pi}_n - \pi_n \to_p 0$.

The LS estimator $\widehat{\psi}_{k,n}$, for $k < K$, is always consistent because ξ_k is larger than $O(n^{-1/2})$ $\forall\ k < K$. The consistency of $\widehat{\psi}_{K,n}$ depends on whether there exist any $O(n^{-1/2})$ elements in β_n. If there is no $O(n^{-1/2})$ element in β_n, then $\widehat{\psi}_{K,n}$ is consistent. Otherwise, $\widehat{\psi}_{K,n}$ converges in distribution to a random variable that minimizes the sample paths of a non-central chi-square process, as in Lemma 3.3(a) and Lemma 6.2(a).

In the presence of the linear regressors, the covariance matrices $G(\pi)$, $V(\pi)$, and $\Sigma(\pi)$ are the same as in the simple model, with the adjustment that $m_i(\pi) = \left(Z_i', g(X_i,\pi)', g_\pi(X_i,\pi)'\right)'$.

Assumption 4c. $\lambda_{\min}(G(\pi)) \geq \varepsilon\ \forall \pi \in \Pi$ for some $\varepsilon > 0$.

Lemma 6.6 *Suppose Assumptions* 1, 2, 3c, 4c, *and* U *hold.*

(**a**) *When* $n^{1/2}\xi_{K,n} \to b_K \in R^{p_K}$,

$$\left(\begin{array}{c} n^{1/2}\left(\widehat{\xi}_{K^-,n} - \xi_{K^-,n}\right) \\ n^{1/2}\left(\widehat{\xi}_{K,n} - \xi_{K,n}\right) \\ n^{1/2} D\left(\xi_{K^-,n}\right)\left(\widehat{\psi}_{K^-,n} - \psi_{K,n}\right) \end{array}\right) \Rightarrow \tau\left(\psi_K^*(h), h\right), \text{ where}$$

$$\tau(\psi_K, h) = \Phi_K^{-1}(\psi_K)\left(S(\psi_K) + \Phi_{Ks}\left(\psi_K, \psi_{K,0}\right) b_K\right) - \iota_K b_K,$$

$$D\left(\xi_{K^-,n}\right) = diag\left\{\xi'_{1,n}, ..., \xi'_{(K-1),n}\right\} \text{ and } \iota_K = \left(0_{p_K \times (q+p-p_K)}, I_{p_K}, 0_{p_K \times (p-p_K)}\right)'.$$

(**b**) *When* $\left\|n^{1/2}\xi_{K,n}\right\| \to \infty$,

$$\left(\begin{array}{c} n^{1/2}\left(\widehat{\zeta}_n - \zeta_n\right) \\ n^{1/2}\left(\widehat{\beta}_n - \beta_n\right) \\ n^{1/2} D(\beta_n)\left(\widehat{\pi}_n - \pi_n\right) \end{array}\right) \Rightarrow N\left(0, \Sigma(\pi_0)\right), \text{ where } D(\beta_n) = diag\left\{\beta'_n\right\}.$$

Comments: 1. This is a generalization of Lemma 6.3. In Lemma 6.3(b), where β_{2n} is $O(n^{-1/2})$ but β_{1n} is not, β_1 and β_2 play the roles of ξ_{K^-} and ξ_K, respectively. When ψ_K cannot be consistently estimated, as in Lemma 6.6(a), the asymptotic distribution involves the Gaussian process $\tau(\psi_K, h)$ and the random variable $\psi_K^*(h)$. Note that ζ is included in ξ_{K^-} by definition.

2. In Lemma 6.6(b), all parameters can be consistently estimated and have asymptotic normal distributions. The convergence rate of $\widehat{\pi}_{j,n}$ depends on $\beta_{j,n}$, as in Lemma 6.3(b). Note that $\Sigma(\pi_0)$ is not the standard asymptotic covariance matrix of the LS estimator. It does not contain β_n to avoid the problem of singularity when some components of β_n converge to 0. When $D(\beta_n)$ is nonsingular as n goes to infinity, we can move $D(\beta_n)$ to the rhs of Lemma 6.6(b) and obtain the standard asymptotic covariance matrix for the LS estimator.

6.3 Test Statistics in the General Model

To construct valid CSs in the general model, we first derive the asymptotic distributions of various LS-based test statistics. We start with the loading coefficients $\mu = \left(\zeta', \beta'\right)'$, where ζ is the coefficient of the linear regressors and β is the coefficient of the nonlinear regressors. We are interested in CSs for some vector-valued linear combination of μ given by $R\mu$, where R is a $r \times (p+q)$ selector matrix of rank r. Let $W_n(\mu_R)$ be a test statistic for $H_0 : R\mu = \mu_R$ with sample size n. The nominal level $1-\alpha$ confidence set (CS) for $R\mu$ is $\{\mu_R : W_n(\mu_R) < c_{n,1-\alpha}(\mu_R)\}$, where $c_{n,1-\alpha}(\mu_R)$ is the critical value.

We use the Wald statistic of the form

$$W_n\left(\mu_{R,n}\right) = n\left(R\widehat{\mu}_n - \mu_{R,n}\right)'\left(R\widehat{\Sigma}_\mu\left(\widehat{\pi}_n\right)R'\right)^{-1}\left(R\widehat{\mu}_n - \mu_{R,n}\right), \tag{6.14}$$

where $\mu_{R,n}$ is the true value with sample size n, $\widehat{\Sigma}_\mu(\pi) = R_\mu \widehat{\Sigma}_n(\pi) R_\mu'$ and $R_\mu = \left(I_{p+q}, 0_{(p+q)\times p}\right)$. The selector matrix R_μ is used to select the variance of $\widehat{\mu}_n$ out of $\widehat{\Sigma}_n(\widehat{\pi}_n)$, which contains the variance of $\widehat{\pi}_n$ as well. The matrix $\widehat{\Sigma}_\mu(\widehat{\pi}_n)$ used in $W_n(\mu_R)$ is equivalent to the standard covariance matrix even though $\widehat{\Sigma}_n(\widehat{\pi}_n)$ is not. The reason is that R_μ does not select the non-standard part of $\widehat{\Sigma}_n(\widehat{\pi}_n)$. When the rank of R is 1, we can also use the t statistic for $R\mu$. The t statistic takes the form

$$T_{\mu,n}\left(\mu_{R,n}\right) = \frac{n^{1/2}\left(R\widehat{\mu}_n - \mu_{R,n}\right)}{\left(R\widehat{\Sigma}_\mu(\widehat{\pi}_n) R'\right)^{1/2}}. \tag{6.15}$$

To characterize the asymptotic distributions of $W_n\left(\mu_{R,n}\right)$ and $T_{\mu,n}\left(\mu_{R,n}\right)$, we let $\tau_\mu(\psi_K, h)$ be a $(p+q)$ dimensional sub-vector of $\tau(\psi_K, h)$ that corresponds to μ. It is defined as

$$\tau_\mu(\psi_K, h) = \left(I_{p+q}, 0_{(p+q)\times(p-p_K)}\right)\tau(\psi_K, h), \tag{6.16}$$

where $\tau(\psi_K, h)$ is the $(q+2p-p_K)$ dimensional Gaussian process defined in Lemma 6.6(a). Analogous to $\widehat{\Sigma}_\mu(\pi)$, we define $\Sigma_\mu(\psi_K) = R_\mu \Sigma(\psi_K) R_\mu'$, where $\Sigma(\psi_K) = \Sigma\left(\psi_{K^-,0}, \psi_K\right)$.

Theorem 6.1 *Suppose Assumptions* 1, 2, 3c, 4c, *and* U *hold.*
(**a**) *When* $n^{1/2}\xi_{K,n} \to b_K \in R^{p_K}$,

$$\begin{aligned}
&W_n\left(\mu_{R,n}\right) \Rightarrow W\left(\psi_K^*(h), h\right),\ T_{\mu,n}\left(\mu_{R,n}\right) \Rightarrow T_\mu\left(\psi_K^*(h), h\right),\ \textit{where}\\
&W(\psi_K, h) = \tau_\mu(\psi_K, h)' R'\left(R\Sigma_\mu(\psi_K) R'\right)^{-1} R\tau_\mu(\psi_K, h)\ \textit{and}\\
&T_\mu(\psi_K, h) = R\tau_\mu(\psi_K, h)\left(R\Sigma_\mu(\psi_K) R'\right)^{-1/2}.
\end{aligned}$$

(**b**) *When* $\left\|n^{1/2}\xi_{K,n}\right\| \to \infty$,

$$W_n\left(\mu_{R,n}\right) \Rightarrow \chi_r \textit{ and } T_{\mu,n}\left(\mu_{R,n}\right) \Rightarrow N(0,1).$$

Comment: The Wald statistic $W_n\left(\mu_{R,n}\right)$ has a chi-square distribution in the standard case. In a non-standard situation, i.e. $b_K \in R^{P_K}$, its asymptotic distribution involves a chi-square process and a random variable $\pi^*(h)$. The asymptotic distribution of $T_{\mu,n}(\mu_{R,n})$ is a generalization of that of $T_{\beta,n}(\beta_n)$ in Theorem 3.1. The non-standard asymptotic distributions all depend on the finite localization parameter h.

The next step is to derive asymptotic distributions of the test statistics for π. Different from β and ζ, π cannot always be estimated consistently. Even in the presence of consistency, the rate of convergence is not always $n^{-1/2}$. Due to these non-standard features, we consider CSs for $\pi_{j,n}$ for $j = 1, ..., p$, individually. In this case, CSs become CIs. Let R_j be a row vector used to select the j^{th} element out of a column vector. For notational simplicity, we adjust R_j to adapt to the

dimension of the vector to be selected. The t statistic for π_j takes the form

$$T_{\pi_j,n}(\pi_{j,n}) = n^{1/2}(\widehat{\pi}_{j,n} - \pi_{j,n})/\widehat{\sigma}_{\pi,j}, \tag{6.17}$$

where $\widehat{\sigma}_{\pi,j} = (R_j\widehat{\Sigma}_\pi(\widehat{\pi}_n)R_j')\widehat{\beta}_{j,n}^{-1}$ and $\widehat{\Sigma}_\pi(\pi) = R_\pi\widehat{\Sigma}_n(\pi)R_\pi'$. The selector matrix $R_\pi = \left(0_{p\times(p+q)}, I_p\right)$ is used to select the variance of $\widehat{\pi}$ out of $\widehat{\Sigma}_n(\widehat{\pi}_n)$ and R_j is used to select π_j from the π vector. Analogous to $\widehat{\Sigma}_\pi(\pi)$, we define $\Sigma_\pi(\psi_K) = R_\pi\Sigma(\psi_K)R_\pi'$. Note that $T_{\pi_j,n}(\pi_{j,n})$ is equivalent to the standard LS-based t statistic, although $\widehat{\beta}_{j,n}$ is written separately from the rest of the covariance matrix.

Let $\tau_\pi(\psi_K, h)$ and $\tau_\beta(\psi_K, h)$ be sub-vectors of $\tau(\psi_K, h)$ that correspond to π and β, respectively. They are defined as

$$\begin{aligned}\tau_\pi(\psi_K, h) &= \left(0_{(p-p_K)\times(p+q)}, I_{p-p_K}\right)\tau(\psi_K, h) \text{ and}\\ \tau_\beta(\psi_K, h) &= \left(0_{p\times q}, I_p, 0_{p\times(p-p_K)}\right)\tau(\psi_K, h) + b_K.\end{aligned} \tag{6.18}$$

The Gaussian process $\tau_\beta(\psi_K, h)$ is needed to determine the asymptotic distribution of $\widehat{\sigma}_{\pi,j}$, which depends on $\widehat{\beta}_{j,n}$. Note that we add b_K to $\tau_\beta(\psi_K, h)$ because $\widehat{\beta}_{j,n}$ is not centered in the definition of $\widehat{\sigma}_{\pi,j}$. Let $\psi_{K,0}$ be the limit of $\psi_{K,n}$ as $n \to \infty$.

Theorem 6.2 *Suppose Assumptions* 1, 2, 3c, 4c, *and* U *hold.*
(**a**) *When* $n^{1/2}\xi_{K,n} \to b_K \in R^{p_K}$,

$$\begin{aligned}&T_{\pi_j,n}(\pi_{j,n}) \Rightarrow T_{\pi_j}(\psi_K^*(h), h), \text{ where}\\ &T_{\pi_j}(\psi_K, h) = R_j\tau_\pi(\psi_K, h)\left(R_j\Sigma_\pi(\psi_K)R_j'\right)^{-1/2} \text{ for } j \le p - p_K, \text{ and}\\ &T_{\pi_j}(\psi_K, h) = R_j\tau_\beta(\psi_K, h)R_{j-(p-p_K)}\left(\psi_K - \psi_{K,0}\right)\left(R_j\Sigma_\pi(\psi_K)R_j'\right)^{-1/2} \text{ for } j > p - p_K.\end{aligned}$$

(**b**) *When* $\left\|n^{1/2}\xi_{K,n}\right\| \to \infty$,

$$T_{\pi_j,n}(\pi_{j,n}) \Rightarrow N(0,1) \ \forall j = 1, ..., p.$$

Comment: When $n^{1/2}\xi_{K,n} \to b_K \in R^{p_K}$, the asymptotic distributions are non-standard due to the inconsistency of $\widehat{\psi}_{K,n}$. Depending on whether π_j belongs to ψ_K, we have two different types of non-standard asymptotic distributions for $T_{\pi_j,n}(\pi_{j,n})$. When π_j does not belong to ψ_K, which means $\left|n^{1/2}\beta_{j,n}\right| \to \infty$ and $\widehat{\pi}_{j,n}$ is consistent, $T_{\pi_j,n}(\pi_{j,n})$ has an asymptotic distribution analogous to that of $T_{\mu,n}\left(\mu_{R,n}\right)$ in Theorem 6.1. On the other hand, when π_j belongs to ψ_K, $\widehat{\pi}_{j,n}$ is inconsistent. In this case, $T_{\pi_j,n}(\pi_{j,n})$ has an asymptotic distribution analogous to that of $T_{\pi,n}(\pi_n)$ in Theorem 3.1(a). In the definition of $T_{\pi_j}(\psi_K, h)$ for $j > p - p_K$, the first R_j selects the limit of $\widehat{\beta}_{j,n}$ from $\tau_\beta(\psi_K, h)$, the second R_j corresponds to that in the definition of $\widehat{\sigma}_{\pi,j}$, and $R_{j-(p-p_K)}$ selects $\widehat{\pi}_{j,n} - \pi_{j,0}$ from $\psi_K - \psi_{K,0}$.

The non-standard asymptotic distributions derived in Theorems 6.1 and 6.2 provide good approximations to the finite-sample distributions of the test statistics under weak identification. Allowing the number of elements of b_K to vary from 0 to p, the local limit theory covers all identification scenarios. For a given localization parameter h, quantiles can be simulated from the analytical formulae in these theorems.

7 General Confidence Set

Applying the local limit theory derived in the last section, this section generalizes results on the standard CI, subsampling CI, and robust CI in Sections 4 and 5 to the general model with multiple nonlinear regressors and linear regressors.

We are interested in constructing a CS for some sub-vector of $\mu = (\beta', \zeta')'$ denoted by $R\mu$ and for π_j. The asymptotic size of the CS is defined in (3.3) by replacing CI with CS. We first analyze the asymptotic sizes of the standard CS and the subsampling CS. Then we construct a robust CS with correct asymptotic size in the general model.

7.1 Standard Confidence Set and Subsampling Confidence Set

The critical value for a standard CS is obtained by assuming all parameters are strongly identified. The test statistic used for $R\mu$ is the Wald statistic $W_n\left(\mu_{R,n}\right)$. The standard critical value for a nominal level $1-\alpha$ CS is $\chi_r\left(1-\alpha\right)$, which is the $1-\alpha$ quantile of the chi-square distribution with r degree of freedom. The t statistics $T_{\mu,n}\left(\mu_{R,n}\right)$ and $T_{\pi_j,n}\left(\pi_{j,n}\right)$ are used to construct CSs for $R\mu$ with $r=1$ and for π_j, respectively. When a t statistic is used, the standard critical value is a quantile from the standard normal distribution.[12] Using the general notation defined in Section 4, the standard critical value is $c_\infty\left(1-\alpha\right)$.

The subsampling critical values are the same as discussed in Section 4. With multiple nonlinear regressors, the set LH defined in (4.1) is modified to

$$\begin{aligned} LH = \{&(l,h) \in H \times H : l = (l_b, \pi_0),\ h = (b, \pi_0),\ \text{and for } j = 1, ..., p, \\ &\text{(i) } l_{b,j} = 0 \text{ if } |b_j| < \infty,\ \text{(ii) } l_{b,j} \in R_{+,\infty} \text{ if } b_j = +\infty,\ \text{and (iii) } l_{b,j} \in R_{-,\infty} \text{ if } b_j = -\infty\}, \end{aligned} \quad (7.1)$$

where $b = (b_1, ..., b_p)'$ and $l_b = (l_{b,1}, ..., l_{b,p})'$. The idea of this adjustment is that the subsampling localization parameter $l_{b,j}$ is closer to 0 than the full-sample localization parameter b_j, but is not related to $b_{j'}$ for any $j' \neq j$.

Theorem 7.1 *Suppose Assumptions* 1, 2, 3c, 4c, *and* U *hold. Then,*
(**a**) $AsyCS = \inf_{h \in H} J_h(c_\infty(1-\alpha))$ *for the standard CS and*

[12] When a symmetric two-sided CI is constructed, the asymptotic distribution of the test statistic is the absolute value of a random variable with standard normal distribution.

(**b**) $AsyCS = \inf_{(l,h)\in LH} J_h(c_l(1-\alpha))$ *for the subsampling CS.*

As in the simple model, the asymptotic sizes of the standard CS and subsampling CS can be simulated with these explicit formulae. The non-standard asymptotic distributions J_h are provided in Theorems 6.1 and 6.2 for $R\mu$ and π_j, respectively.

7.2 General Robust Confidence Set

Following the idea of the robust CI in Section 5, we choose the critical value for the robust CS in a general model by a model-selection procedure. The model-selection procedure provides two advantages to the general robust CS. First, it reduces the volume of the CS relative to the least favorable CS when some nonlinear regressors are strongly identified. Second, it reduces the optimization dimensions in the critical value simulation. The dimension reduction is made clear below in the description of the critical value for the robust CS.

The model-selection procedure is used to determine whether β_j is close to 0, modelled as $\beta_j = O(n^{-1/2})$. Specifically, we choose between $M_{0,j} : b_j$ is finite and $M_{1,j} : |b| = \infty$, for $j = 1, ..., p$, where b_j is the limit of $n^{1/2}\beta_{j,n}$ as n goes to infinity.[13] The statistic used for selecting between $M_{0,j}$ and $M_{1,j}$ takes the form

$$t_{n,j} = \left| n^{1/2}\widehat{\beta}_{n,j}/\widehat{\sigma}_{\beta,j} \right|. \tag{7.2}$$

Model $M_{0,j}$ is selected if $t_{n,j} \leq \kappa_n$ and $M_{1,j}$ is selected otherwise. The statistic $t_{n,j}$ is $O_p(1)$ if and only if $b_j \in R$. Hence, we can consistently select $M_{0,j}$ provided that the tuning parameter κ_n diverges to infinity.

We now define the critical value for the robust CS. Let $b^* = \left(b_1^*, ..., b_p^*\right)'$, where $b_j^* = \infty$ if $M_{1,j}$ is chosen, i.e. $t_{n,j} > \kappa_n$.[14] The critical value for the robust CS is

$$\widehat{c}_n(1-\alpha) = \sup_{h^*\in H} c_{h^*}(1-\alpha), \text{ where } h^* = (b^{*\prime}, \pi')'. \tag{7.3}$$

Note that if the model-selection procedure suggests $|b_j| = \infty \ \forall \ j = 1, ..., p$, $\widehat{c}_n(1-\alpha)$ is equivalent to the standard critical value $c_\infty(1-\alpha)$. When $p = 1$, the definition of $\widehat{c}_n(1-\alpha)$ is equivalent to that in (5.2). By applying the model-selection procedure, we use the data to determine which nonlinear regressors are weakly identified. On the one hand, the critical value we choose is large enough to cover the weak identification suggested by the data. On the other hand, the critical value is not unnecessarily large by setting b_j^* equal to ∞ when the data suggests that β_j is large and π_j is strongly identified.

[13] In this section, we use the scalar b_j to denote the limit of $n^{1/2}\beta_{j,n}$ for $j = 1, ..., p$. This is different from, but closely related to, the vector b_k, for $k = 1, ..., K$, defined in (3.4). In (3.4), b_k is the limit of $n^{1/2}\xi_k$, which is a group of β_j with the same rate of convergence. Here we recycle the notation b_j to reflect that it is the localization parameter corresponding to β_j.

[14] Setting $b_j^* = \infty$ and $-\infty$ leads to the same result because both correspond to strong identification. Hence, we do not distinguish them here.

The algorithm to construct a robust CS in the general model has four steps: (1) Estimate the general model (2.1) by the LS estimator, yielding $\widehat{\beta}_j$ and its standard error $\widehat{\sigma}_{\beta,j}$, for $j = 1, ..., p$. (2) Use the model-selection procedure to determine the weak-identification group. The statistics for model selection are constructed according to (7.2) . If $t_{n,j} \leq \kappa_n$, β_j is in a $n^{-1/2}$ neighborhood of 0 and π_j is taken to be weakly identified. Hence, $\beta_j \in \xi_K$ and $\pi_j \in \psi_K$ according to the grouping rule described in Section 6.2. Otherwise, β_j and ψ_j are not in the weak-identification group. (3) Simulate the $1 - \alpha$ quantile of the non-standard asymptotic distributions derived in Theorems 6.1 and 6.2 for given b_K and π_0. In this step, one uses the knowledge of ξ_K and ψ_K obtained from the model-selection procedure. (4) Take the supremum of the critical value $c_{h^*}(1 - \alpha)$ obtained in step 3 over $h^* \in H$, as defined in (7.3) . This is the critical value for the robust CS.

The following theorem states that the robust CS constructed above has corrected asymptotic size provided the tuning parameter κ_n for the model-selection procedure diverges to infinity with the sample size.

Theorem 7.2 *Suppose Assumptions* 1, 2, 3c, 4c, U*, and* R *hold. The nominal level* $1 - \alpha$ *robust CS satisfies* $AsyCS = 1 - \alpha$.

8 Conclusion

This paper develops a robust inference method under weak identification, focussing on constructing CIs in a nonlinear regression model. Under general conditions, we show that the new robust CI has correct asymptotic size while the standard CI and the subsampling CI are prone to severe size distortions. We develop a local limit theory under sequences of parameters that drift to the non-identification point(s). We start with the asymptotic distribution of the sample criterion function and develop consistency, rates of convergence, and inference results for LS estimators and LS-based test statistics. Under weak identification, non-standard asymptotic distributions based on the local limit theory provide good uniform approximations to the finite-sample distributions of the test statistics. Thus, the robust CI based on the local limit theory has finite-sample coverage probability close to the nominal level, as demonstrated by simulation results. We use a model selection procedure to shorten the robust CI under strong identification and to simplify critical value simulation in a general model with multiple nonlinear regressors. A sequential procedure is used to deal with multiple strengths of identification in the general model. Although the paper focuses on LS-based CIs in a nonlinear regression model, the empirical process method used to analyze the sample criterion function and test statistics can be applied to other criterion-based estimators in a general weak identification set-up. This is work in progress by Andrews and Cheng.

There are several directions for further work. First, we are interested in a Bonferonni CI that will reduce the non-similarity of the robust CI while keeping the correct asymptotic size. Second, instead of making a sharp switch between the standard critical value and the least favorable critical

value based on the model selection result, we can use a weighted average of them. Some practical averaging methods were proposed in Andrews and Soares (2007) and Andrews and Jia (2008) in a model with moment inequalities. Optimal weights can be developed based on Hansen (2007) using model averaging methods. Third, we can construct the generalized method of moments (GMM) CIs to deal with endogeneity. We can further relax the smoothness assumption to allow kinks in the nonlinear functions. Fourth, the methods developed here can be generalized to a time series set-up and applied to nonlinear time-series models, such as the widely used smooth transition autoregressive model and the threshold model.

Appendix

A Proofs for the Simple Model with One Nonlinear Regressor

The following Lemma provides uniform convergence results that are used in the sequel.

Lemma A.1 *Suppose Assumptions* 1 *and* 2 *hold.*
(**a**) $n^{-1}\sum_{i=1}^{n} m_i(\pi)\, m_i(\overline{\pi})' \to_p Em_i(\pi)\, m_i(\overline{\pi})'$ *uniformly over* $\Pi \times \Pi$.
(**b**) *When* $n^{1/2}\beta_n \to b \in R$, $n^{-1}\sum_{i=1}^{n} \widehat{u}_i(\pi)^2 m_i(\pi)\, m_i(\pi)' \to_p V(\pi)$ *uniformly over* Π *and* $\widehat{\Sigma}_n(\pi) \to_p \Sigma(\pi)$ *uniformly over* Π.
In part (a), (b), *and* (c), *the rhs are all uniformly continuous in the parameters.*

Let $\overline{S}_n(\pi) = n^{-1/2}\sum_{i=1}^{n} U_i m_i(\pi)$. Next lemma shows weak convergence of the empirical process $\overline{S}_n(\pi)$. The weak convergence result is used in derivation of asymptotic theory on the sample criterion function. Note that $S_n(\pi) = n^{-1/2}\sum_{i=1}^{n} U_i g(X_i, \pi)$ is a sub-vector of $\overline{S}_n(\pi)$. Hence, weak convergence of $\overline{S}_n(\pi)$ implies that of $S_n(\pi)$.

Lemma A.2 *Suppose Assumptions* 1 *and* 2 *hold. Then,* $\overline{S}_n(\pi) \Rightarrow \overline{S}(\pi)$, *where* $\overline{S}(\pi)$ *is a mean zero Gaussian process with covariance kernel* $\overline{\Omega}(\pi, \overline{\pi}) = E\sigma^2(X_i)\, m_i(\pi)\, m_i(\overline{\pi})'$.

Lemma A.3 below provides asymptotic results on the LS estimator of β when π is fixed. This result is used in the Proof of Lemma 3.1.

Lemma A.3 *Suppose Assumptions* 1, 2, *and* 3c *hold.*
(**a**) *When* $n^{1/2}\beta_n \to b \in R$, $n^{1/2}\widehat{\beta}(\pi) \Rightarrow \Phi^{-1}(\pi, \pi)\left(S(\pi) + \Phi(\pi, \pi_0)\, b\right)$.
(**b**) *When* $\left|n^{1/2}\beta_n\right| \to \infty$, $\widehat{\beta}(\pi)/\beta_n = \Phi^{-1}(\pi, \pi)\,\Phi(\pi, \pi_0)$.

Proof of Lemma A.3.
(a) Given π, the LS estimator of β is

$$\begin{aligned}\widehat{\beta}(\pi) &= \left(g(X, \pi)'\, g(X, \pi)\right)^{-1}\left(g(X, \pi)'\, Y\right) \\ &= \left(g(X, \pi)'\, g(X, \pi)\right)^{-1}\left(g(X, \pi)'\, U + g(X, \pi)'\, g(X, \pi_n)\,\beta_n\right). \end{aligned} \tag{A.1}$$

Applying Lemma A.1 and Lemma A.2, we have

$$\begin{aligned} n^{1/2}\widehat{\beta}(\pi) &= \left(n^{-1} g(X, \pi)'\, g(X, \pi)\right)^{-1}\left(n^{-1/2} g(X, \pi)'\, U + g(X, \pi)'\, g(X, \pi_n)\, n^{1/2}\beta_n\right) \\ &\Rightarrow \Phi^{-1}(\pi, \pi)\left(S(\pi) + \Phi(\pi, \pi_0)\, b\right). \end{aligned} \tag{A.2}$$

(b) When β_n converges to 0 slower than $n^{-1/2}$ (or is bounded away from 0),

$$\begin{aligned}\widehat{\beta}(\pi)/\beta_n &= \left(n^{-1} g(X, \pi)'\, g(X, \pi)\right)^{-1}\left(\left(n^{-1/2} g(X, \pi)'\, U\right) / \left(n^{1/2}\beta_n\right) + n^{-1} g(X, \pi)'\, g(X, \pi_n)\right) \\ &\to_p \Phi^{-1}(\pi, \pi)\,\Phi(\pi, \pi_0). \end{aligned} \tag{A.3}$$

uniformly over Π. The convergence in probability holds by Lemma A.1, Lemma A.2, and $\left|n^{1/2}\beta_n\right| \to \infty$.

Proof of Lemma 3.1.

(a) Given π, the LS residual is $\widehat{U}(\pi) = Y - g(X,\pi)\widehat{\beta}(\pi)$. The concentrated sample criterion function is

$$\begin{aligned} Q_n(\pi) &= n^{-1}\widehat{U}(\pi)'\widehat{U}(\pi) = n^{-1}\left(Y - g(X,\pi)\widehat{\beta}(\pi)\right)'\left(Y - g(X,\pi)\widehat{\beta}(\pi)\right) \\ &= n^{-1}Y'Y - \left(n^{-1}g(X,\pi)'g(X,\pi)\right)\widehat{\beta}^2(\pi). \end{aligned} \tag{A.4}$$

When $n^{1/2}\beta_n \to b \in R$, we have

$$\begin{aligned} & n\left(Q_n(\pi) - n^{-1}Y'Y\right) = -\left(n^{-1}g(X,\pi)'g(X,\pi)\right)\left(n^{1/2}\widehat{\beta}(\pi)\right)^2 \\ & \Rightarrow -\Phi^{-1}(\pi,\pi)\left(S(\pi) + \Phi(\pi,\pi_0)b\right)^2, \end{aligned} \tag{A.5}$$

where the weak convergence holds by Lemma A.3(a).

(b) When $\left|n^{1/2}\beta_n\right| \to \infty$, $(A.4)$ gives

$$\beta_n^{-2}\left(Q_n(\pi) - n^{-1}Y'Y\right) = \left(n^{-1}g(X,\pi)'g(X,\pi)\right)\left(\widehat{\beta}(\pi)/\beta_n\right)^2 \to_p -\Phi^{-1}(\pi,\pi)\Phi^2(\pi,\pi_0) \tag{A.6}$$

uniformly over Π. The uniform convergence in probability holds by Lemma A.3(b). □

Proof of Lemma 3.2.

Define the Gaussian process $Z(\pi) = \Phi^{-1/2}(\pi)\left(S(\pi) + \Phi(\pi,\pi_0)b\right)$. A sample path of the chi-square process $Z(\pi)^2$ can only achieve its supremum where $Z(\pi)$ achieves its supremum or infimum. Hence, we just need to show that with probability one, no sample path of $Z(\pi)$ achieves its supremum or infimum at two distinct points, and no sample has supremum and infimum with the same absolute value. According to Kim and Pollard (1990) (hereafter KP), if

$$Var\left(\Phi^{-1/2}(\pi_1)S(\pi_1) - \Phi^{-1/2}(\pi_2)S(\pi_2)\right) \neq 0,\ \forall \pi_1 \neq \pi_2, \tag{A.7}$$

no sample path of $Z(\pi)$ can achieve its supremum at two distinct points of Π with probability one. Applying this result, we know that under the same condition, no sample path of $Z(\pi)$ achieves its infimum at two distinct points of Π with probability one. It only remains to show that with probability one, no sample path of $Z(\pi)$ has supremum equal to the opposite value of its infimum. To this end, we also need

$$Var\left(\Phi^{-1/2}(\pi_1)S(\pi_1) + \Phi^{-1/2}(\pi_2)S(\pi_2)\right) \neq 0,\ \forall \pi_1 \neq \pi_2. \tag{A.8}$$

We first show that $(A.7)$ and $(A.8)$ are implied by Assumption 2(c). Suppose $(A.7)$ is not satisfied, so that $Var\left(\Phi^{-1/2}(\pi_1)S(\pi_1) - \Phi^{-1/2}(\pi_2)S(\pi_2)\right) = 0$, for some $\pi_1 \neq \pi_2$.Then $Corr(S(\pi_1), S(\pi_2)) = 1$, which implies that $g(X_i,\pi_1) = kg(X_i,\pi_2)$ a.s. for some $k > 0$. This contradicts Assumption 2(c). We can show $(A.8)$ analogously.

Next we show that under $(A.8)$, no sample path of $Z(\pi)$ has supremum equal to the opposite value of its infimum. The argument is analogous to that in Lemma 2.6 of KP. For each pair of distinct points π_0 and π_1, instead of taking supremum of $Z(\pi)$ over neighborhoods N_0 of π_0 and N_1 of π_1 as in KP, we take supremum of $Z(\pi)$ over N_0 and supremum of $-Z(\pi)$ over N_1. Using the notations in KP, the covariance of $Z(\pi_0)$ and $-Z(\pi_1)$ is $-H(\pi_0,\pi_1)$. Under $(A.8)$, $-H(\pi_0,\pi_1)$ cannot be equal to both $H(\pi_0,\pi_0)$ and $H(\pi_1,\pi_1)$. Suppose $H(\pi_0,\pi_0) > -H(\pi_0,\pi_1)$, we get $h(\pi_0) = 1 > -h(\pi_1)$, where $h(\pi)$ is defined to be $H(\pi_1,\pi_0)/H(\pi_0,\pi_0)$ as in KP. The rest

of the proof is the same as that in KP, but we change β_1 and $\Gamma_1(s)$ to $\beta_1 = \sup_{\pi \in N_0}(-h(\pi))$ and $\Gamma_1(s) = \sup_{\pi \in N_1}(-Y(\pi) - h(\pi)s)$. This leads to the desired result $P\{\sup_{\pi \in N_0} Z(\pi) = \sup_{\pi \in N_1}(-Z(\pi))\} = 0$. □

Proof of Lemma 3.3.

(a) When $n^{1/2}\beta_n \to b \in R$, we have shown the weak convergence of $Q_n(\pi)$ in Lemma 3.1(a). Let $\widehat{\pi}_n$ minimize $n(Q_n(\pi) - n^{-1}\sum_{i=1}^n Y'Y)$ up to $o_p(1)$. By Lemma 3.2, $\pi^*(h)$ uniquely minimizes the rhs of Lemma 3.1(a), whose sample paths are continuous with probability one and the parameter space Π is compact. We appeal to Theorem 3.2.2 (the argmax CMT) of van de Vaart and Wellner (1996, p.286) and get $\widehat{\pi}_n \Rightarrow \pi^*(h)$.

(b) By Cauchy-Schwarz inequality, $\pi_{10} = \arg\min_{\pi \in \Pi}(-\Phi^{-1}(\pi, \pi)\Phi^2(\pi, \pi_0))$ under Assumption 2. Hence, $\widehat{\pi}_n - \pi_n = (\widehat{\pi}_n - \pi_0) - (\pi_n - \pi_0) = o_p(1)$. □

Proof of Lemma 3.4.

(a) Define empirical process $\tau_n(\pi) = n^{1/2}(\widehat{\beta}(\pi) - \beta_n)$ and use $Q(\pi)$ to denote the rhs of Lemma 3.1(a) without the negative sign. By Lemma A.3 and Lemma 3.1,

$$\begin{pmatrix} \tau_n(\pi) \\ n(Q_n(\pi) - n^{-1}Y'Y) \end{pmatrix} \Rightarrow \begin{pmatrix} \tau(\pi) \\ -Q(\pi) \end{pmatrix}, \tag{A.9}$$

where the joint weak convergence holds because both $\tau(\pi)$ and $Q(\pi)$ are continuous functions of $S(\pi)$. Let $\pi^* = \arg\max_{\pi \in \Pi} Q(\pi)$. Since $\tau(\pi^*)$ is continuous wrt the process $(\tau(\pi), -Q(\pi))'$, we have $\tau_n(\widehat{\pi}_n) \Rightarrow \tau(\pi^*)$ by the CMT. Hence, $n^{1/2}(\widehat{\beta}(\pi) - \beta_n) \Rightarrow \tau(\pi^*)$. Note that all the stochastic processes here depend on the localization parameter h. We omit h for notational simplicity.

(b) Next, we derive the asymptotic distribution in Lemma 3.4(b). Consistency of $\widehat{\pi}_n$ is obtained in Lemma 3.3(b). The next is to show that $\widehat{\beta}_n$ is also consistent. For any fixed π, $|\widehat{\beta}(\pi) - \beta_n| = |\beta_n||\widehat{\beta}(\pi)/\beta_n - 1|$. Plugging in the consistent estimator $\widehat{\pi}_n$ and applying $(A.3)$, we have $|\widehat{\beta}_n - \beta_n| = |\widehat{\beta}(\widehat{\pi}_n) - \beta_n| \to_p 0$.

Let $\theta = (\beta, \pi)'$.[15] Assume the LS estimator $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{\pi}_n)'$ satisfies $\partial Q_n(\widehat{\theta}_n)/\partial\theta = o_p\left(n^{-1/2}\right)$. Mean value expansions of $\partial Q_n(\widehat{\theta}_n)/\partial\theta$ about θ_n yields

$$o_p(n^{-1/2}) = \frac{\partial Q_n(\widehat{\theta}_n)}{\partial\theta} = \frac{\partial Q_n(\theta_n)}{\partial\theta} + \frac{\partial^2 Q_n(\theta_n^*)}{\partial\theta\partial\theta'}\left(\widehat{\theta}_n - \theta_n\right), \tag{A.10}$$

where θ_n^* lies between $\widehat{\theta}_n$ and θ_n (and, hence, satisfies $\theta_n^* - \theta_n \to_p 0$). Define a weighting matrix $D(\beta) = Diag\{1, \beta\}$ and $D_n = Diag\{1, \beta_n\}$. Equation $(A.10)$ implies

$$n^{1/2}D_n\left(\widehat{\theta}_n - \theta_n\right) = \left(\frac{1}{2}D_n^{-1}\frac{\partial^2 Q_n(\theta_n^*)}{\partial\theta\partial\theta'}D_n^{-1}\right)^{-1}\left(-\frac{1}{2}D_n^{-1}n^{1/2}\frac{\partial Q_n(\theta_n)}{\partial\theta}\right) + o_p(1). \tag{A.11}$$

[15] Note that θ is used for a different purpose in the defnition of $T_n(\theta)$ in Section 4.1 and Section 7. In $T_n(\theta)$, θ represents the parameter of interest in general.

The first and second order derivatives take the form

$$\frac{\partial Q_n(\theta)}{\partial \theta} = -2D(\beta)\left(n^{-1}\sum_{i=1}^{n} U_i(\theta)\, m_i(\theta)\right) \text{ and}$$

$$\frac{\partial Q_n(\theta)}{\partial \theta \partial \theta'} = -2\left(n^{-1}\sum_{i=1}^{n} U_i(\theta)\, m_{\theta\theta}(X_i,\theta)\right) + 2D(\beta)\left(n^{-1}\sum_{i=1}^{n} m(X_i,\pi)\, m(X_i,\pi)'\right) D(\beta), \text{ where}$$

$$U_i(\theta) = Y_i - g(X_i,\pi)\beta \text{ and } m_{\theta\theta}(X_i,\theta) = \begin{pmatrix} 0 & g_\pi(X_i,\pi) \\ g_\pi(X_i,\pi) & g_{\pi\pi}(X_i,\pi)\beta \end{pmatrix}. \tag{A.12}$$

By Lemma A.1, the first order derivative in $(A.11)$ is

$$-\frac{1}{2}D_n^{-1} n^{1/2} \frac{\partial Q_n(\theta_n)}{\partial \theta} = n^{-1/2}\sum_{i=1}^{n} U_i m_i(\pi_n) + o_p(1) \Rightarrow N(0, V(\pi_0)). \tag{A.13}$$

The second order derivative in $(A.11)$ is

$$\frac{1}{2}D_n^{-1}\frac{\partial^2 Q_n(\theta_n^*)}{\partial \theta \partial \theta'} D_n^{-1} = n^{-1}\sum_{i=1}^{n} m_i(\pi_n^*)\, m_i(\pi_n^*)' -$$

$$D_n^{-1}\left(n^{-1}\sum_{i=1}^{n} U_i(\theta_n^*)\, m_{\theta\theta}(X_i,\theta_n^*)\right) D_n^{-1} + o_p(1) \to_p G(\pi_0), \tag{A.14}$$

where the convergence in probability holds by Lemma A.1. It remains to show that

$$D_n^{-1}\left(n^{-1}\sum_{i=1}^{n} U_i(\theta_n^*)\, m_{\theta\theta}(X_i,\theta_n^*)\right) D_n^{-1} = o_p(1). \tag{A.15}$$

Note that

$$n^{-1}\sum_{i=1}^{n} U_i(\theta_n^*)\, g_\pi(X_i,\pi_n^*)/\beta_n = n^{-1}\sum_{i=1}^{n}(U_i + g(X_i,\pi_n)\beta_n - g(X_i,\pi_n^*)\beta_n^*) g_\pi(X_i,\pi_n^*)/\beta_n$$

$$= \left(n^{-1/2}\sum_{i=1}^{n} U_i g_\pi(X_i,\pi_n^*)\right)/\left(n^{1/2}\beta_n\right) + n^{-1}\sum_{i=1}^{n} g(X_i,\pi_n)\, g_\pi(X_i,\pi_n^*)$$

$$-n^{-1}\sum_{i=1}^{n} g(X_i,\pi_n^*)\, g_\pi(X_i,\pi_n^*)(\beta_n^*/\beta_n) \to_p 0, \tag{A.16}$$

where the convergence in probability holds by weak convergence of $\overline{S}_n(\pi)$, uniform convergence of $n^{-1}\sum_{i=1}^{n} g(X_i,\pi)\, g_\pi(X_i,\pi)$, and $\beta_n^*/\beta_n = 1 + o_p(1)$ obtained from $\widehat{\beta}_n/\beta_n = 1 + o_p(1)$ based on Lemma A.3(b). Finally, we can show that $n^{-1}\sum_{i=1}^{n} U_i(\theta_n^*)\, g_\pi(X_i,\pi_n^*)\beta_n^*/\beta_n^2 = o_p(1)$ in the same way as in $(A.16)$.

Plugging $(A.13)$ and $(A.14)$ into $(A.11)$, we get the desired result. □

Proof of Theorem 3.1.

(a) When $n^{1/2}\beta_n \to b \in R$, we have

$$\frac{n^{1/2}\left(\widehat{\beta}_n(\pi)-\beta_n\right)}{\left(\widehat{\Sigma}_n(\pi)_{11}\right)^{1/2}} \Rightarrow T_\beta(\pi,h) \text{ and } \frac{n^{1/2}\widehat{\beta}(\pi)(\pi-\pi_n)}{\left(\widehat{\Sigma}_n(\pi)_{22}\right)^{1/2}} \Rightarrow T_\pi(\pi,h), \tag{A.17}$$

using uniform convergence of $\widehat{\Sigma}_n(\pi)$ to $\Sigma(\pi)$ together with Lemma A.3(a). Applying the same argument as in the proof of Lemma 3.4, we plug $\widehat{\pi}_n$ into the empirical processes on the lhs of $(A.17)$ and plug π^* into the Gaussian processes on the rhs. This lead to $T_{\beta,n} \Rightarrow T_\beta(\pi^*(h),h)$ and $T_{\pi,n} \Rightarrow T_\pi(\pi^*(h),h)$.

(b) When $\left|n^{1/2}\beta_n\right| \to \infty$, the results in part (b) are directly implied by Theorem 3.4(b). □

Next we prove Lemma A.1 and Lemma A.2 stated at the beginning of this subsection.

Proof of Lemma A.1.

To prove the uniform convergence and uniform continuity results in part (a) and (b), we invoke Theorem 4 and Lemma 4(a) of Andrews (1992), which state that when $\{X_i : i \geq 1\}$ are identically distributed, $\sup_{\tau\in\varsigma}|n^{-1}\sum_{i=1}^n q(X_i,\tau) - Eq(X_i,\tau)| \to_p 0$ and $Eq(X_i,\tau)$ is continuous in τ uniformly over ς if (i) ς is compact, (ii) $q(X_i,\tau)$ is continuous in τ $\forall\tau\in\varsigma$, $\forall X_i \in \mathcal{X}$, (iii) $E\sup_{\tau\in\varsigma}\|q(X_i,\tau)\| < \infty$, and (iv) $n^{-1}\sum_{i=1}^n q(X_i,\tau) \to_p Eq(X_i,\tau)$ pointwise for all $\tau\in\varsigma$. Condition (i) is always satisfied because the parameter space is a compact Euclidean space. Now we shall verify conditions (ii), (iii), and (iv) in different contexts with various functional forms of $q(X_i,\tau)$.

In part (a), the uniform convergence is applied to $q(X_i,\pi,\overline{\pi}) = m_i(\pi)\,m_i(\overline{\pi})'$. Conditions (ii) and (iii) are satisfied by Assumptions 2. Condition (iv) is satisfied by the WLLN and condition (iii) with *i.i.d.* data.

To prove part (b), note that $\widehat{U}_i(\pi) = U_i + g(X_i,\pi_n)\beta_n - g(X_i,\pi)\widehat{\beta}(\pi)$, where β_n and $\widehat{\beta}(\pi)$ are both $o_p(1)$ uniformly over Π. Then

$$\begin{aligned}
&n^{-1}\sum_{i=1}^n \widehat{U}_i(\pi)^2 m_i(\pi)\,m_i(\pi)' = n^{-1}\sum_{i=1}^n U_i^2 m_i(\pi)\,m_i(\pi)' + \\
&2n^{-1}\sum_{i=1}^n U_i\left(g(X_i,\pi_n)\beta_n - g(X_i,\pi)\widehat{\beta}(\pi)\right)m_i(\pi)\,m_i(\pi)' + \\
&n^{-1}\sum_{i=1}^n \left(g(X_i,\pi_n)\beta_n - g(X_i,\pi)\widehat{\beta}(\pi)\right)^2 m_i(\pi)\,m_i(\pi)' \to_p V(\pi)
\end{aligned} \tag{A.18}$$

uniformly over Π, because the second term and the third term on the rhs of the equality are both $o_p(1)$ uniformly over Π. In the last step of $(A.18)$, we use the same argument as in parts (a) by setting $q(X_i,\pi) = U_i^2 m_i(\pi)\,m_i(\pi)'$. The uniform convergence and continuity of $G(\pi)$ and $V(\pi)$ together with Assumption 4a lead to uniform convergence and continuity of $\Sigma(\pi)$ over Π. □

Proof of Lemma A.2. To show weak convergence of the empirical process $\overline{S}_n(\pi)$, we use the proposition in Andrews (1994, p.2251). Because the parameter space Π is compact, it is remaining to show that $\overline{S}_n(\pi)$ is stochastic equicontinuous and the finite dimensional convergence holds. First, we show the stochastic equicontinuity of $\overline{S}_n(\pi)$. To this end, we appeal to example 3 of Andrews (1994), which states that a sufficient condition for the stochastic equicontinuity of $\overline{S}_n(\pi)$ is that $g(X,\pi)$ is differentiable wrt π and

$$E\sup_{\pi\in\Pi}\|U_i g_\pi(X_i,\pi)\| < \infty \text{ and } E\sup_{\pi\in\Pi}\|U_i g_{\pi\pi}(X_i,\pi)\| < \infty. \tag{A.19}$$

This sufficient condition is satisfied because $EU_i^2 < \infty$, $E \sup_{\pi \in \Pi} g_{\pi j}^2 (X_i, \pi_j) < \infty$, and $E \sup_{\pi \in \Pi} g_{\pi \pi j}^2 (X_i, \pi_j) < \infty$ for $j = 1$ and 2 by Assumption 2. Because $U_i m_i (\pi)$ is *i.i.d.* for any given π, the multivariate CLT establishes the finite dimensional convergence with the covariance kernel $\overline{\Omega}(\pi, \overline{\pi}) = E\left(m_i(\pi) m_i(\pi) \sigma^2(X_i)\right)$. □

B Proofs for the General Model with Multiple Nonlinear Regressors

Results in a two-regressor model discussed in Section 6.1 are special cases of those in Section 6.2. In this subsection, we first prove the general results in Section 6.2 and then discuss the corresponding results in Section 6.1 as special cases. Proofs of Lemmas 6.1, 6.2, and 6.3 follow those of Lemmas 6.4, 6.5, and B.1, respectively.

Lemma B.1 *Suppose Assumptions* 1 *and* 2 *hold.*

(**a**) $n^{-1} \sum_{i=1}^n m_i(\pi) m_i(\overline{\pi})' \to_p E m_i(\pi) m_i(\overline{\pi})$ *uniformly over* $\Pi \times \Pi$.

(**b**) $n^{-1} s_k(\psi_{k^-}, \psi_k)' M_{k^+}(\psi_{k^+}) s_k(\psi_{k^-}, \overline{\psi}_k) \to_p \overline{\Phi}_k(\psi_{k^-}, \psi_k, \overline{\psi}_k | \psi_{k^+})$ *uniformly over* $(\psi_{k^-}, \psi_k, \overline{\psi}_k, \psi_{k^+}) \in \Psi_{k^-} \times \Psi_k \times \Psi_k \times \Psi_{k^+}$, *where* $\overline{\Phi}_k(\psi_{k^-}, \psi_k, \overline{\psi}_k | \psi_{k^+}) = E \widetilde{s}_{k,i}(\psi_{k^-}, \psi_k | \psi_{k^+}) \widetilde{s}_{k,i}(\psi_{k^-}, \overline{\psi}_k | \psi_{k^+})'$.

(**c**) *When* $n^{1/2} \xi_{K,n} \to b_K \in R^{P_K}$, $n^{-1} \sum_{i=1}^n \widehat{U}_i(\psi_K)^2 m_i(\psi_K) m_i(\psi_K)' \to_p V(\psi_{K^-,0}, \psi_K)$ *uniformly over* Ψ_K *and* $\widehat{\Sigma}_n(\widehat{\psi}_K(\psi_K), \psi_K) \to_p \Sigma(\psi_K)$ *uniformly over* Ψ_K, *where* $\widehat{U}_i(\psi_K)$ *is the ith row of* $\widehat{U}(\psi_K) = Y - Z\widehat{\zeta}(\psi_K) - f_{K^-}(\widehat{\psi}_{K^-}(\psi_K))\widehat{\xi}_{K^-}(\psi_K) - f_K(\psi_K)\widehat{\xi}_K(\psi_K)$ *and* $m_i(\psi_K) = m_i(\widehat{\psi}_{K^-}(\psi_K), \psi_K)$.

In part (a), (b), *and* (c), *the rhs are all uniformly continuous in the parameters.*

Comment: Note that $\overline{\Phi}_k(\psi_{k^-,0}, \psi_k, \overline{\psi}_k | \psi_{k^+}) = \Phi_k(\psi_k, \overline{\psi}_k | \psi_{k^+})$. Hence, when $\widehat{\psi}_{k^-,n}(\psi_k | \psi_{k^+})$ is plugged into the lhs of part (b) and converges to $\psi_{k^-,0}$ uniformly over $\Psi_k \times \Psi_{k^+}$, the asymptotic limit in part (b) becomes $\Phi_k(\psi_k, \overline{\psi}_k | \psi_{k^+})$.

The next Lemma is a generalization of Lemma A.2. Let $\overline{S}_n(\pi) = n^{-1/2} \sum_{i=1}^n U_i m_i(\pi)$.

Lemma B.2 *Suppose Assumption 1 and 2 holds. Then* $\overline{S}_n(\pi) \Rightarrow \overline{S}(\pi)$, *where* $\overline{S}(\pi)$ *is a mean zero Gaussian process with covariance kernel* $\overline{\Omega}(\pi, \overline{\pi}) = E U_i^2 m_i(\pi) m_i(\overline{\pi})'$.

Comment: Note that $S_n(\pi) = n^{-1/2} s_K(\psi_{K^-}, \psi_K)' U$ is a sub-vector of $\overline{S}_n(\pi)$. Thus, weak convergence of $\overline{S}_n(\pi)$ implies weak convergence of $S_n(\pi)$.

Lemma B.3 *Suppose Assumptions* 1, 2, 3c, *and* 4c *hold.*

(**a**) *For* $k = 1, ..., K$, *when* $\left| n^{1/2} \xi_k \right| \to \infty$,

$$\beta_{k_1,n}^{-1} \begin{pmatrix} \widehat{\xi}_{k^-}(\psi_k | \psi_{k^+}) - \xi_{k^-,n} \\ \widehat{\xi}_k(\psi_k | \psi_{k^+}) \\ D\left(\xi_{k^-,n}\right)\left(\widehat{\psi}_{k^-}(\psi_k | \psi_{k^+}) - \psi_{k^-,n}\right) \end{pmatrix} \to_p \Phi_k^{-1}(\psi_k | \psi_{k^+}) \Phi_{ks}\left(\psi_k, \psi_{k,0} | \psi_{k^+}\right) \Delta_k$$

uniformly over $\Psi_k \times \Psi_{k^+}$, *where* $\Delta_k = \lim_{n \to \infty}\left(\xi_{k,n} / \beta_{k_1,n}\right)$.

(**b**) *When* $\left\| n^{1/2} \xi_{Kn} \right\| \to b_K \in R^{p_K}$,

$$n^{1/2} \begin{pmatrix} \widehat{\xi}_{K^-}(\psi_K) - \xi_{K^-,n} \\ \widehat{\xi}_K(\psi_K) \\ D\left(\xi_{K^-,n}\right)\left(\widehat{\psi}_{K^-}(\psi_K) - \psi_{K,n}\right) \end{pmatrix} \Rightarrow \Phi_K^{-1}(\psi_K)\left(S(\psi_K) + \Phi_{Ks}\left(\psi_K, \psi_{K,0}\right) b_K\right).$$

Comment: 1. A sequential procedure from $k = 1$ to K is employed in the proofs of Lemma B.3 and Lemma 6.4. Specifically, we first prove Lemma B.3(a) for $k = 1$, which is then used to show Lemma 6.4(a) for $k = 1$. Lemma 6.4(a) for $k = 1$ can in turn be used in the proof of B.3(a) for $k = 2$, and so on.

2. Lemma A.3 is a special case of Lemma B.3 when $k = 1$, $p_1 = 1$. The special case of Lemma B.3 with two nonlinear regressors is presented in Corollary B.1 below.

Corollary B.1 *Suppose Assumptions* 1, 2, 3c, *and* 4c *hold.*

(**a**) *When* $n^{1/2}\beta_n \to b \in R^2$, $n^{1/2}\widehat{\beta}(\pi) \Rightarrow \Phi^{-1}(\pi,\pi)\left(S(\pi) + \Phi(\pi,\pi_0)\, b\right)$.

(**b**) *When* $\left|n^{1/2}\beta_{1n}\right| \to \infty$ *and* $\beta_{2n} = o(\beta_{1n})$, $\beta_{1n}^{-1}\widehat{\beta}_1(\pi_1|\pi_2) \to_p \Phi_1^{-1}(\pi_1|\pi_2)\,\Phi_1(\pi_1,\pi_{10}|\pi_2)$, *uniformly over* $\Pi_1 \times \Pi_2$.

(**c**) *When* $\left|n^{1/2}\beta_{1n}\right| \to \infty$ *and* $n^{1/2}\beta_{2n} \to b_2 \in R$,

$$n^{1/2}\begin{pmatrix} \widehat{\beta}_1(\pi_2) - \beta_{1n} \\ \widehat{\beta}_2(\pi_2) \\ \beta_{1n}\left(\widehat{\pi}_1(\pi_2) - \pi_{1n}\right) \end{pmatrix} \Rightarrow \Phi_2^{-1}(\pi_2)\left(\overline{S}(\pi_2) + \Phi_{2s}(\pi_2,\pi_{20})b_2\right).$$

(**d**) *When* $\left|n^{1/2}\beta_{1n}\right| \to \infty$, $\left|n^{1/2}\beta_{2n}\right| \to \infty$, *and* $\beta_{2n} = o(\beta_{1n})$,

$$\beta_{2n}^{-1}\begin{pmatrix} \widehat{\beta}_1(\pi_2) - \beta_{1n} \\ \widehat{\beta}_2(\pi_2) \\ \beta_{1n}\left(\widehat{\pi}_1(\pi_2) - \pi_{1n}\right) \end{pmatrix} \to_p \Phi_2^{-1}(\pi_2)\,\Phi_{2s}(\pi_2,\pi_{20})$$

uniformly over Π_2.

(**e**) *When* $\left|n^{1/2}\beta_{1n}\right| \to \infty$, $\left|n^{1/2}\beta_{2n}\right| \to \infty$, $\beta_{2n} = O(\beta_{1n})$, *and* $\beta_{1n} = O(\beta_{2n})$, $\beta_{1n}^{-1}\widehat{\beta}(\pi) \Rightarrow \Phi^{-1}(\pi,\pi)\,\Phi(\pi,\pi_0)\,\Delta$.

Comment: Part (a) is an application of Lemma B.3(b) with $K = 1$, where K is the number of groups we defined in Section 6.2. Part (b) is an application of Lemma B.3(a) on β_1 when $K = 2$. Both (c) and (d) are sub-cases of (b). Part (c) is an application of Lemma B.3(b) on β_2 and part (d) is an application of Lemma B.3(a) on β_2. Finally, part (e) corresponds to the case $K = 1$ but β is larger than $O(n^{-1/2})$. It is an application of Lemma B.3(a).

Proof of Lemma B.3.

(a) In the proof of Lemma B.3, we start with $k = 1$ and establish the results sequentially for $k = 2, ..., K$.

When $k = 1$, ψ_{1^+} is fixed. We estimate $\varrho_1 = \left(\zeta', \xi_1'\right)'$ by partitioned regression for given ψ_1, yielding

$$\begin{aligned}
&\widehat{\varrho}_1(\psi_1|\psi_{1^+}) = \left(s_1(\psi_1)' M_{1^+}(\psi_{1^+}) s_1(\psi_1)\right)^{-1}\left(s_1(\psi_1)' M_{1^+}(\psi_{1^+}) Y\right) \\
&= \left(n^{-1}s_1(\psi_1)' M_{1^+}(\psi_{1^+}) s_1(\psi_1)\right)^{-1}\left(n^{-1}s_1(\psi_1)' M_{1^+}(\psi_{1^+}) U + \right. \\
&\left. n^{-1}s_1(\psi_1)' M_{1^+}(\psi_{1^+})\left(s_1(\psi_{1n})\varrho_{1n} + f_{1^+}\left(\psi_{1^+,n}\right)\xi_{1^+,n}\right)\right), \qquad \text{(B.1)}
\end{aligned}$$

where $s_1(\psi_1) = [Z, f_1(\psi_1)]$ by definition in (6.10). Let $\widetilde{\varrho}_{1n} = (\zeta_n', 0_{p_1}')'$. Subtracting both sides of $(B.1)$ by ϱ_{1n} and dividing both sides by $\beta_{1_1,n}$, which is the first element of $\xi_{1,n}$, we get

$$\beta_{1_1,n}^{-1}\left(\widehat{\varrho}_1(\psi_1|\psi_{1^+}) - \widetilde{\varrho}_{1n}\right) = \Phi_1^{-1}(\psi_1|\psi_{1^+})\,\Phi_{1s}(\psi_1,\psi_{10}|\psi_{1^+})\,\Delta_1, \qquad \text{(B.2)}$$

uniformly over $\Psi_1 \times \Psi_{1^+}$, where $\Delta_1 = \lim_{n\to\infty} \beta_{1_1,n}^{-1} \xi_{1n} \in R^{p_1}$. In $(B.2)$, the convergence in probability holds by Lemma B.1, Lemma B.2, $|n^{1/2}\beta_{1_1,n}| \to \infty$, and $\xi_{1^+,n} = o\left(\beta_{1_1,n}\right)$. With $(B.2)$, Lemma 6.4(a), (b) for $k = 1$ can be established. For details, see proof of Lemma 6.4. Lemma 6.4(b) provides uniform consistency of $\widehat{\psi}_1(\psi_{1^+})$ over Ψ_{1^+}.

For notational simplicity, we write $\widehat{\varrho}_k(\widehat{\psi}_{k^-}(\psi_{k^+}), \widehat{\psi}_k(\psi_{k^+})|\psi_{k^+})$ as $\widehat{\varrho}_k(\psi_{k^+})$. Plugging $\widehat{\psi}_1(\psi_{1^+})$ into the lhs of $(B.2)$ and subtracting both sides by $(0_q', \xi_{1n}')'$, we have

$$\beta_{1_1,n}^{-1}\left(\widehat{\varrho}_1(\psi_1|\psi_{1^+}) - \varrho_{1n}\right) \to_p 0. \tag{B.3}$$

Hence, $\widehat{\varrho}_1(\psi_{1^+})$ is uniformly consistent over Ψ_{1^+}.

We have established Lemma B.3(a) for $k = 1$ already. The proof will be complete if the argument can be extended from $k-1$ to k, $\forall\ k = 2, ..., K$. Next, we prove Lemma B.3(a) for k, assuming that we have got the result for $k-1$. Let $\varrho_k = \left(\zeta', \xi_1', ..., \xi_k'\right)'$. Suppose the step for $k-1$ provides uniform consistency of $\widehat{\varrho}_{(k-1)}(\psi_{(k-1)^+})$ and $\widehat{\psi}_{(k-1)}(\psi_{(k-1)^+})$ over $\Psi_{(k-1)^+}$. Note that we have done so for $k = 2$ by showing uniform consistency of $\widehat{\varrho}_1(\psi_{1^+})$ and $\widehat{\psi}_1(\psi_{1^+})$.

We first need to show uniform consistency of $\widehat{\varrho}_k(\psi_k|\psi_{k^+})$ uniformly over $\Psi_k \times \Psi_{k^+}$ before analyzing the concentrated sample criterion function. Note that $\widehat{\varrho}_k(\psi_k|\psi_{k^+}) = (\widehat{\varrho}_{(k-1)}(\psi_k, \psi_{k^+})', \widehat{\xi}_k(\psi_k|\psi_{k^+})')'$. To show the uniform consistency of $\widehat{\varrho}_k(\psi_k|\psi_{k^+})$, we only need to show it for $\widehat{\xi}_k(\psi_k|\psi_{k^+})$, as the consistency of $\widehat{\varrho}_{(k-1)}(\psi_k, \psi_{k^+})$ is already established in the step for $k-1$. Note that $\widehat{\xi}_k(\psi_k|\psi_{k^+})$ is the LS estimator of ξ_k when ψ_k and ψ_{k^+} are fixed and ψ_{k^-} is $\widehat{\psi}_{k^-}(\psi_k, \psi_{k^+})$. To use partitioned regression for $\widehat{\xi}_k(\psi_k|\psi_{k^+})$, define

$$\begin{aligned}
&M_{-k}(\psi_k, \psi_{k^+}) = M\left(f_{-k}\left(\widehat{\psi}_{k^-}(\psi_k, \psi_{k^+}), \psi_{k^+}\right)\right), \text{ where} \\
&M(X) = I_n - X\left(X'X\right)^{-1}X' \text{ and } f_{-k}(\psi_{k^-}, \psi_{k^+}) = [f_{k^-}(\psi_{k^-}), f_{k^+}(\psi_{k^+})].
\end{aligned} \tag{B.4}$$

By partitioned regression,

$$\begin{aligned}
&\widehat{\xi}_k(\psi_k|\psi_{k^+}) = \left(f_k(\psi_k)' M_{-k}(\psi_k, \psi_{k^+}) f_k(\psi_k)\right)^{-1}\left(f_k(\psi_k)' M_{-k}(\psi_k, \psi_{k^+}) Y\right) \\
= &\left(n^{-1} f_k(\psi_k)' M_{-k}(\psi_k, \psi_{k^+}) f_k(\psi_k)\right)^{-1} \times [n^{-1} f_k(\psi_k)' M_{-k}(\psi_k, \psi_{k^+}) f_k\left(\psi_{k,n}\right)\xi_k + \\
&n^{-1} f_k(\psi_k)' M_{-k}(\psi_k, \psi_{k^+})\left(f_{k^-}\left(\psi_{k^-,n}\right)\xi_{k^-,n} + f_{k^+}\left(\psi_{k^+,n}\right)\xi_{k^+,n}\right) + \\
&n^{-1}(f_k(\psi_k)' M_{-k}(\psi_k, \psi_{k^+}) U]
\end{aligned} \tag{B.5}$$

Now we analyze the four terms in the rhs of $(B.5)$ using Lemma B.1 and Lemma B.2. The first term is $O_p(1)$ by expanding $M_{-k}(\psi_k, \psi_{k^+})$ and applying Lemma B.1(a). The second term is $o_p(1)$ because ξ_k is $o_p(1)$. The third term is $o_p(1)$ by uniform convergence of $\widehat{\psi}_{k^-}(\psi_k, \psi_{k^+})$, differentiability of $f_{k^-}(\psi_k)$ wrt to ψ_k, the moment condition the first order derivative, and the fact that $M_{-k}(\psi_k, \psi_{k^+})$ is orthogonal to $f_{k^-}(\widehat{\psi}_{k^-}(\psi_k, \psi_{k^+}))$ and $f_{k^+}(\psi_{k^+,n})$. Finally, the last term is $o_p(1)$ by expanding $M_{-k}(\psi_k, \psi_{k^+})$ and applying Lemma B.2. Because $\xi_{k,n}$ is $o_p(1)$ for $k > 1$, we have $|\widehat{\xi}_k(\psi_k|\psi_{k^+}) - \xi_{k,n}| = o_p(1)$ uniformly over $\Pi_k \times \Pi_{k^+}$.This leads to consistency of $\widehat{\varrho}_k(\psi_k|\psi_{k^+})$ uniformly over $\Psi_k \times \Psi_{k^+}$. The result also implies uniform consistency of $\widehat{\varrho}_k(\psi_{k^+})$ over Ψ_{k^+}, which can be used for the step of $k+1$.

The next is to derive the asymptotic distributions. By definition,

$$\theta_k = (\varrho_k, \psi_{k^-}) = \left(\xi_{k^-}', \xi_k', \psi_{k^-}'\right)'. \tag{B.6}$$

Note that ζ is included in ξ_{k^-} already. For fixed ψ_{k^+}, let $\widehat{\theta}_k(\psi_k|\psi_{k^+})$ be the LS estimator of θ_k as a function of ψ_k. We have shown the uniform consistency of $\widehat{\theta}_k(\psi_k|\psi_{k^+})$ over $\Psi_k \times \Psi_{k^+}$ by proving this property for both $\widehat{\varrho}_k(\psi_k|\psi_{k^+})$ and $\widehat{\psi}_{k^-}(\psi_k|\psi_{k^+})$. Suppose the LS estimator satisfies

$$\partial Q_n(\widehat{\theta}_k(\psi_k|\psi_{k^+}), \psi_k|\psi_{k^+})/\partial\theta_k = o_p(n^{-1/2}) \tag{B.7}$$

uniformly over $\Psi_k \times \Psi_{k^+}$, where

$$\begin{aligned} Q_n(\theta_k, \psi_k|\psi_{k^+}) &= n^{-1}U(\theta_k, \psi_k|\psi_{k^+})' U(\theta_k, \psi_k|\psi_{k^+}) \text{ and} \\ U(\theta_k, \psi_k|\psi_{k^+}) &= M_{k^+}(\psi_{k^+})(Y - f_{k^-}(\psi_{k^-})\xi_{k^-} - f_k(\psi_k)\xi_k). \end{aligned} \tag{B.8}$$

Mean expansion of $\partial Q_n(\widehat{\theta}_k(\psi_k|\psi_{k^+}), \psi_k|\psi_{k^+})/\partial\theta_k$ about $\theta_{k,n} = (\xi'_{k^-,n}, 0, \psi'_{k^-,n})'$ yields

$$\begin{aligned} o_p\left(n^{-1/2}\right) &= \frac{\partial Q_n(\widehat{\theta}_k(\psi_k|\psi_{k^+}), \psi_k|\psi_{k^+})}{\partial\theta_k} \\ &= \frac{\partial Q_n(\theta_{k,n}, \psi_k|\psi_{k^+})}{\partial\theta_k} + \frac{\partial^2 Q_n\left(\theta^*_{k,n}(\psi_k|\psi_{k^+}), \psi_k|\psi_{k^+}\right)}{\partial\theta_k\partial\theta'_k}\left(\widehat{\theta}_k(\psi_k|\psi_{k^+}) - \theta_{k,n}\right), \end{aligned} \tag{B.9}$$

where $\theta^*_{k,n}(\psi_k|\psi_{k^+})$ lies between $\widehat{\theta}_k(\psi_k|\psi_{k^+})$ and $\theta_{k,n}$ (and, hence, satisfies $\widehat{\theta}_k(\psi_k|\psi_{k^+}) - \theta_{k,n} \to_p 0$ uniformly over $\Pi_k \times \Pi_{k^+}$). Note that in the definition of $\theta_{k,n}$, ξ_k is fixed at 0 instead of $\xi_{k,n}$. This is just for simplicity of the result. Using $\xi_{k.n}$ as the center will lead to a more complicated, but equivalent, expression of the result.

Define weighting matrices $D(\xi_{k^-}) = Diag\{\iota_{q+n_k}, \xi'_{k^-}\}$ and $D_k = Diag\{\iota_{q+n_k}, \xi'_{k^-,n}\}$, where ι_{q+n_k} is a $q + n_k$ dimensional row vector of 1 and $n_k = \sum_{i=1}^k p_i$. The first and second order derivatives take the form

$$\begin{aligned} \frac{\partial Q_n(\theta_k, \psi_k|\psi_{k^+})}{\partial\theta_k} &= -2D(\xi_{k^-})n^{-1}s_k(\psi_{k^-}, \psi_k)' U(\theta_k, \psi_k|\psi_{k^+}) \text{ and} \\ \frac{\partial Q_n(\theta_k, , \psi_k|\psi_{k^+})}{\partial\theta_k\partial\theta'_k} &= 2D(\xi_{k^-})\left(n^{-1}s_k(\psi_{k^-}, \psi_k)' M_{k^+}(\psi_{k^+}) s_k(\psi_{k^-}, \psi_k)\right) D(\xi_{k^-}) - \\ &\qquad 2\left(n^{-1}\sum_{i=1}^n U(\theta_k, \psi_k|\psi_{k^+}) m_{\theta\theta,ki}(\psi_{k^-})\right), \text{ where} \\ m_{\theta\theta,ki}(\psi_{k^-}) &= \begin{pmatrix} 0 & f_{\psi_{k^-},i}(\psi_{k^-})' \\ f_{\psi_{k^-},i}(\psi_{k^-}) & D(\xi_{k^-}) diag\left\{f_{\psi\psi_{k^-},i}(\psi_{k^-})\right\} \end{pmatrix}. \end{aligned} \tag{B.10}$$

Note that

$$U(\theta_{k,n}, \psi_k|\psi_{k^+}) = M_{k^+}(\xi_{k^+})\left(U + f_k\left(\psi_{k,n}\right)\xi_{k,n} + f_{k^+}\left(\psi_{k^+,n}\right)\xi_{k^+,n}\right). \tag{B.11}$$

When $\left|n^{1/2}\xi_{k,n}\right| \to \infty$, by $(B.9)$, we have

$$\begin{aligned} D_k\left(\widehat{\theta}_k(\psi_k|\psi_{k^+}) - \theta_{k,n}\right)\beta^{-1}_{k_1,n} &= \left(\frac{1}{2}D_k^{-1}\frac{\partial^2 Q_n\left(\theta^*_{k,n}(\psi_k|\psi_{k^+}), \psi_k|\psi_{k^+}\right)}{\partial\theta_k\partial\theta'_k}D_k^{-1}\right)^{-1} \times \\ \left(-\frac{1}{2}D_k^{-1}\frac{\partial Q_n(\theta_{k,n}, \psi_k|\psi_{k^+})}{\partial\theta_k}\beta^{-1}_{k_1,n}\right) + o_p(1). \end{aligned} \tag{B.12}$$

In $(B.12)$, the part with first order derivative is

$$-\frac{1}{2}D_k^{-1}\frac{\partial Q_n\left(\theta_{k,n},\psi_k|\psi_{k^+}\right)}{\partial\theta_k}\beta_{k_1,n}^{-1}=n^{-1}\sum_{i=1}^{n}U_i\left(\theta_{k,n},\psi_k|\psi_{k^+}\right)s_{k,i}\left(\psi_{k^-,n},\psi_k\right)\beta_{k_1,n}^{-1}$$
$$=n^{-1/2}s_k\left(\psi_{k^-,n},\psi_k\right)'M_{k^+}\left(\xi_{k^+}\right)U/\left(n^{1/2}\beta_{k_1,n}\right)+$$
$$\left(n^{-1}s_k\left(\psi_{k^-,n},\psi_k\right)'M_{k^+}\left(\xi_{k^+}\right)f_k\left(\psi_{k,n}\right)\right)\xi_{k,n}\beta_{k_1,n}^{-1}+$$
$$\left(n^{-1}s_k\left(\psi_{k^-,n},\psi_k\right)'M_{k^+}\left(\xi_{k^+}\right)f_k\left(\psi_{k,n}\right)\right)\xi_{k^+,n}\beta_{k_1,n}^{-1}+o_p\left(1\right)\to_p\Phi_{ks}\left(\psi_k,\psi_{k,0}|\psi_{k^+}\right)\Delta_k, \quad (B.13)$$

where the first and third terms in the second step are both $o_p\left(1\right)$ uniformly over $\Pi_k\times\Pi_{k^+}$. The second order derivative is

$$\frac{1}{2}D_k^{-1}\frac{\partial Q_n\left(\theta_{k,n}^*\left(\psi_k|\psi_{k^+}\right),\psi_k|\psi_{k^+}\right)}{\partial\theta_k\partial\theta_k'}D_k^{-1}\to_p\ \Phi_k\left(\psi_k|\psi_{k^+}\right), \quad (B.14)$$

where the convergence holds by Lemma B.1 and

$$D_k^{-1}\left(n^{-1}\sum_{i=1}^{n}U\left(\theta_k,\psi_k|\psi_{k^+}\right)m_{\theta\theta,ki}\left(\psi_{k^-}\right)\right)D_k^{-1}=o_p\left(1\right) \quad (B.15)$$

uniformly over Π_{k^+}. The result of $(B.15)$ can be shown term by term using the same argument as in $(A.15)$. As such,

$$\beta_{k_1,n}^{-1}D_k\left(\widehat{\theta}_k\left(\psi_k|\psi_{k^+}\right)-\theta_{k,n}\right)\to_p\Phi_k^{-1}\left(\psi_k|\psi_{k^+}\right)\Phi_{ks}\left(\psi_k,\psi_{k,0}|\psi_{k^+}\right)\Delta_k \quad (B.16)$$

uniformly over $\Pi_k\times\Pi_{k^+}$.

It remains to show the uniform consistency of $\widehat{\psi}_k\left(\psi_{k^+}\right)$ uniformly over Π_{k^+}. Using Lemma B.3(a) for k, we get Lemma 6.4(a) for k, as shown in the proof of Lemma 6.4 below. This lemma gives uniform consistency of $\widehat{\psi}_k\left(\psi_{k^+}\right)$.

Therefore, we have proved Lemma 6.4(a) for $k=1$ and shown that the results hold for k as long as they hold for $k-1$. This completes the proof of part(a).

(b) The lhs of Lemma B.3(b) is an empirical process indexed by ψ_K. We denote this empirical process by $\tau_n\left(\psi_K\right)$. Equation $(B.9)$ implies

$$\tau_n\left(\psi_K\right)=n^{1/2}D_K\left(\widehat{\theta}_K\left(\psi_K\right)-\theta_{K,n}\right)=\left(\frac{1}{2}D_K^{-1}\frac{\partial^2Q_n\left(\theta_{K,n}^*\left(\psi_K\right),\psi_K\right)}{\partial\theta_K\partial\theta_K'}D_K^{-1}\right)^{-1}\times$$
$$\left(-\frac{1}{2}D_K^{-1}n^{1/2}\frac{\partial Q_n\left(\theta_{K,n},\psi_K\right)}{\partial\theta_K}\right)+o_p\left(1\right). \quad (B.17)$$

By Lemma B.1 and Lemma B.2, we have

$$-\frac{1}{2}D_K^{-1}n^{1/2}\frac{\partial Q_n\left(\theta_{K,n},\psi_K\right)}{\partial\theta_K}=n^{-1/2}s_K\left(\theta_{Kn},\psi_K\right)'U_i\left(\theta_{Kn},\psi_K\right)+o_p\left(1\right)$$
$$=n^{-1/2}s_K\left(\theta_{Kn},\psi_K\right)'U_i+n^{-1}s_K\left(\theta_{Kn},\psi_K\right)'f_K\left(\psi_{K,n}\right)\left(n^{1/2}\xi_{K,n}\right)+o_p\left(1\right)$$
$$\Rightarrow S\left(\psi_K\right)+\Phi_{Ks}\left(\psi_K,\psi_{K,0}\right)b_K. \quad (B.18)$$

Now we turn to the second order derivative

$$\begin{aligned}&\frac{1}{2}D_K^{-1}\frac{\partial^2 Q_n\left(\theta_{K,n}^*\left(\psi_K\right),\psi_K\right)}{\partial\theta_K\partial\theta_K'}D_K^{-1}\\&=n^{-1}s_K\left(\theta_{K,n}^*\left(\psi_K\right),\psi_K\right)'s_K\left(\theta_{K,n}^*\left(\psi_K\right),\psi_K\right)+o_p\left(1\right)\to_p\Phi_K\left(\psi_K\right),\end{aligned}\tag{B.19}$$

where the convergence in probability holds by Lemma B.1 and the $o_p\left(1\right)$ term is obtained in the same way as in $(A.15)$.

Plugging $(B.18)$ and $(B.19)$ into $(B.17)$, we get

$$\tau_n\left(\psi_K\right)\Rightarrow\Phi_K^{-1}\left(\psi_K\right)\left(S\left(\psi_K\right)+\Phi_{Ks}\left(\psi_K,\psi_{K,0}\right)b_K\right).\tag{B.20}$$

□

Proof of Lemma 6.4.

For fixed ψ_{k^+}, the LS residual as a function of ψ_k takes the form

$$\widehat{U}\left(\psi_k|\psi_{k^+}\right)=U\left(\widehat{\theta}_k\left(\psi_k,\psi_{k^+}\right),\psi_k|\psi_{k^+}\right),\text{ where}\tag{B.21}$$

$U\left(\theta_k,\psi_k|\psi_{k^+}\right)$ is defined in $(B.8)$. Plugging in the general model with true value and use row by row Taylor expansion of $f_{k^-}(\widehat{\psi}_{k^-}\left(\psi_k,\psi_{k^+}\right))\widehat{\xi}_{k^-}\left(\psi_k|\psi_{k^+}\right)$ around $(\xi_{k^-,n}',\psi_{k^-,n}')'$, we have

$$\begin{aligned}&\widehat{U}\left(\psi_k|\psi_{k^+}\right)=M_{k^+}\left(\psi_{k^+}\right)\left(Y-f_{k^-}\left(\psi_{k^-,n}\right)\xi_{k^-,n}\right)-M_{k^+}\left(\psi_{k^+}\right)s_k\left(\psi_{k^-,n},\psi_k\right)\times\\&D_k\left(\widehat{\theta}_{k,n}\left(\psi_k|\psi_{k^+}\right)-\theta_{k,n}\right)+o\left(\left\|W_k\left(\widehat{\theta}_{k,n}\left(\psi_k|\psi_{k^+}\right)-\theta_{k,n}\right)\right\|\right),\end{aligned}\tag{B.22}$$

where the smaller order term is uniformly over $X_i\in\mathcal{X}$. The concentrated sample criterion is

$$\begin{aligned}&Q_n\left(\psi_k|\psi_{k^+}\right)=n^{-1}\widehat{U}\left(\psi_k|\psi_{k^+}\right)'\widehat{U}\left(\psi_k|\psi_{k^+}\right)\\&=n^{-1}\left(Y-f_{k^-}\left(\psi_{k^-,n}\right)\xi_{k^-,n}\right)'M_{k^+}\left(\psi_{k^+}\right)\left(Y-f_{k^-}\left(\psi_{k^-,n}\right)\xi_{k^-,n}\right)\qquad\text{(B.23)}\\&-2n^{-1}\left(Y-f_{k^-}\left(\psi_{k^-,n}\right)\xi_{k^-,n}\right)'M_{k^+}\left(\psi_{k^+}\right)F_k\left(\psi_{k^-,n},\psi_k\right)\left(D_k\left(\widehat{\theta}_{k,n}\left(\psi_k|\psi_{k^+}\right)-\theta_{k,n}\right)\right)\\&+\left(D_k\left(\widehat{\theta}_{k,n}\left(\psi_k|\psi_{k^+}\right)-\theta_{k,n}\right)\right)'\left(n^{-1}s_k\left(\psi_{k^-,n},\psi_k\right)'M_{k^+}\left(\psi_{k^+}\right)s_k\left(\psi_{k^-,n},\psi_k\right)\right)\times\\&\left(D_n\left(\widehat{\theta}_{k,n}\left(\psi_k|\psi_{k^+}\right)-\theta_{k,n}\right)\right)+o\left(\left\|D_k\left(\widehat{\theta}_{k,n}\left(\psi_k|\psi_{k^+}\right)-\theta_{k,n}\right)\right\|\right).\qquad\text{(B.24)}\end{aligned}$$

(a) When $\left\|n^{1/2}\xi_{k,n}\right\|\to\infty$,

$$\begin{aligned}&\beta_{k_1,n}^{-1}D_k\left(\widehat{\theta}_{k,n}\left(\psi_k|\psi_{k^+}\right)-\theta_{k,n}\right)\to_p\Phi_k^{-1}\left(\psi_k|\psi_{k^+}\right)\Phi_{ks}\left(\psi_k,\psi_{k,0}|\psi_{k^+}\right)\Delta_k,\text{ and}\\&n^{-1}\left(Y-f_{k^-}\left(\psi_{k^-,n}\right)\xi_{k^-,n}\right)'M_{k^+}\left(\psi_{k^+}\right)s_k\left(\psi_{k^-,n},\psi_k\right)\beta_{k_1,n}^{-1}\\&=n^{-1}\left(U+f_k\left(\psi_{k,n}\right)\xi_{k,n}+f_{k^+}\left(\psi_{k^+,n}\right)\xi_{k^+,n}\right)'\times\\&M_{k^+}\left(\psi_{k^+}\right)s_k\left(\psi_{k^-,n},\psi_k\right)\beta_{k_1,n}^{-1}\to_p\left(\Phi_{ks}\left(\psi_k,\psi_{k0}|\psi_{k^+}\right)\Delta_k\right)'.\end{aligned}\tag{B.25}$$

Re-centering and rescaling $Q_n\left(\psi_k|\psi_{k^+}\right)$ according to $(B.24)$ we get

$$\begin{aligned}&\beta_{k_1,n}^{-2}\left(Q_n\left(\psi_k|\psi_{k^+}\right)-n^{-1}\left(Y-f_{k^-}\left(\psi_{k^-,n}\right)\xi_{k^-,n}\right)'M_{k^+}\left(\psi_{k^+}\right)\left(Y-f_{k^-}\left(\psi_{k^-,n}\right)\xi_{k^-,n}\right)\right)\\&\to_p\quad-\Delta_k'\Phi_{ks}\left(\psi_k,\psi_{k,0}|\psi_{k^+}\right)'\Phi_k^{-1}\left(\psi_k|\psi_{k^+}\right)\Phi_{ks}\left(\psi_k,\psi_{k,0}|\psi_{k^+}\right)\Delta_k,\end{aligned}\tag{B.26}$$

where the weak convergence is obtained by $(B.25)$ and the smaller order term is controlled by Lemma B.3(a).

The rhs of $(B.26)$ is uniquely minimized by $\psi_{k,0}$ by a vector Cauchy-Schwarz inequality under Assumption 2. Hence

$$\left\|\widehat{\psi}_k\left(\psi_{k^+}\right)-\psi_{k,n}\right\| \leq \left\|\widehat{\psi}_k\left(\psi_{k^+}\right)-\psi_{k,0}\right\| + \left\|\psi_{k,n}-\psi_{k,0}\right\| = o_p\left(1\right) \tag{B.27}$$

uniformly over Π_{k^+}, where the first term after the inequality is $o_p\left(1\right)$ by minimizing both sides of $(B.26)$ and invoking the CMT.

(b) When $n^{1/2}\xi_{K,n} \to b_K \in R^{p_K}$,

$$\begin{aligned}
&n^{-1/2}\left(Y-f_{K^-}\left(\psi_{K^-,n}\right)\xi_{K^-,n}\right)' s_K\left(\psi_{K^-,n},\psi_K\right) = n^{-1/2}\left(U+f_K\left(\psi_{K,n}\right)\xi_{K,n}\right)' s_K\left(\psi_{K^-,n},\psi_K\right) \\
&= n^{-1/2}U' s_K\left(\psi_{K^-,n},\psi_K\right) + \left(n^{1/2}\xi_{K,n}\right)' n^{-1} f_K\left(\psi_{k,n}\right)' s_K\left(\psi_{K^-,n},\psi_K\right) \\
&\Rightarrow \left(S\left(\psi_K\right)+\Phi_{Ks}\left(\psi_K,\psi_{K,0}\right)b_K\right)'.
\end{aligned} \tag{B.28}$$

Using $(B.24)$, $(B.28)$, and Lemma B.3(b), we have

$$\begin{aligned}
&n\left(Q_n\left(\psi_K\right)-n^{-1}\left(Y-f_{K^-}\left(\psi_{K^-,n}\right)\xi_{K^-,n}\right)'\left(Y-f_{K^-}\left(\psi_{K^-,n}\right)\xi_{K^-,n}\right)\right) \\
&= -2n^{-1/2}\left(Y-f_{K^-}\left(\psi_{K^-,n}\right)\xi_{K^-,n}\right)'\left(n^{1/2}D_K\left(\widehat{\theta}_{K,n}\left(\psi_K\right)-\theta_{K,n}\right)\right) \\
&+\left(n^{1/2}D_K\left(\widehat{\theta}_{K,n}\left(\psi_K\right)-\theta_{K,n}\right)\right)'\times \\
&\left(n^{-1}s_K\left(\psi_{K^-,n},\psi_K\right)' s_K\left(\psi_{K^-,n},\psi_K\right)\right)\times\left(n^{1/2}D_K\left(\widehat{\theta}_{K,n}\left(\psi_K\right)-\theta_{K,n}\right)\right) \\
&\Rightarrow -\left(S\left(\psi_K\right)+\Phi_{Ks}\left(\psi_K,\psi_{K,0}\right)b_K\right)'\Phi_K^{-1}\left(\psi_K\right)\left(S\left(\psi_K\right)+\Phi_{Ks}\left(\psi_K,\psi_{K,0}\right)b_K\right).
\end{aligned} \tag{B.29}$$

□

Proof of Lemma 6.1.

Lemma 6.1 is a special case of Lemma 6.4 with $p=2$. Lemma 6.1(a) is a special case of Lemma 6.4(b) with $K=1$ and uses Corollary B.1(a) in its proof. Lemma 6.1(b) is a special case of Lemma 6.4(a) on π_1 when $K=2$. It uses Corollary B.1(b) in its proof. Both Lemma 6.1(c) and (d) are results on $Q_n\left(\pi_2\right)$ when $K=2$, with (c) being a special case of Lemma 6.4(b) and (d) corresponding to Lemma 6.4(a). They use Corollary B.1 (c) and (d), respectively, in their proofs. Finally, Lemma 6.1(e) is a special case of Lemma 6.4(a) with $K=1$ and it uses Lemma 6.1(e) in the proof. □

Proof of Lemma 6.5.

The consistency of $\widehat{\psi}_{k,n}$ for $k=1,...,K-1$, and the consistency of $\widehat{\psi}_{K,n}$ in Lemma 6.5 are directly obtained by

$$\left\|\widehat{\psi}_k\left(\psi_{k^+}\right)-\psi_{k,n}\right\| = o_p\left(1\right) \text{ uniformly over } \Pi_{k^+}, \tag{B.30}$$

from Lemma 6.4 and a vector Cauchy-Schwarz inequality as in Tripathi (1999). Applying CMT, we have Lemma 6.5(a) by minimizing both sides of Lemma 6.4(c) and applying the CMT. □

Proof of Lemma 6.2.

The proof follows the same idea as that of Lemma 6.5 by applying the CMT. □

Proof of Lemma 6.6.

The proof is analogous to that of Theorem 3.4. Part(a) is based on Lemma B.3(b). We re-center $\widehat{\xi}_{K,n}$ at the true value $\xi_{K,n}$ and plug $\widehat{\psi}_{K,n}$ and ψ_K^* into each side of Lemma B.3(b). We get the desired result by using the same argument as in Theorem 3.4. In part (b), the consistency of $\widehat{\zeta}_n$ and $\widehat{\beta}_n$ are established using $(B.5)$ at the step $k = K$ and the consistency of $\widehat{\pi}_n$ has been shown in Lemma 6.5. With consistency at hand, we can derive the asymptotic normal distribution as we did in Theorem 3.4(b) with direct vector generalization of each term. □

Proof of Lemma 6.3.

Both Theorem 6.3(a) and (b) are special cases of Theorem 6.6(a), corresponding to the cases $K = 1$ and $K = 2$, respectively. Theorem 6.3(c) corresponds to Theorem 6.6(b), where all the loading coefficient converge to 0 slower than $n^{-1/2}$ or are bounded away from 0. □

Proof of Theorem 6.1.

(a) Let $W_n(\psi_K)$ and $T_{\mu,n}(\psi_K)$ be empirical processes indexed by ψ_K. They are defined as

$$W_n(\psi_K) = n^{1/2}\left(\widehat{\mu}_n(\psi_K) - \mu_n\right)' R' \left(R\widehat{\Sigma}_\mu(\psi_K) R'\right)^{-1} Rn^{1/2}\left(\widehat{\mu}_n(\psi_K) - \mu_n\right) \text{ and}$$
$$T_{\mu,n}(\psi_K) = \frac{n^{1/2} R\left(\widehat{\mu}_n(\psi_K) - \mu_n\right)}{\left(R\widehat{\Sigma}_\mu(\psi_K) R'\right)^{1/2}}. \tag{B.31}$$

Because $n^{1/2}(\widehat{\mu}_n(\psi_K) - \mu_n) \Rightarrow \tau_\mu(\psi_K, h)$ by Lemma B.3 and $\widehat{\Sigma}_\mu(\psi_K)$ uniformly converges to $\Sigma_\mu(\psi_K)$ by Lemma B.1, we have

$$W_n(\psi_K) \Rightarrow W(\psi_K, h) \text{ and } T_{\mu,n}(\psi_K) \Rightarrow T_\mu(\psi_K, h). \tag{B.32}$$

We get the desired results by replacing μ_{R_n} with $R\mu_n$, plugging $\widehat{\psi}_{K,n}$ to the empirical processes $W_n(\psi_K)$ and $T_{\mu,n}(\psi_K)$, and plugging $\psi_K^*(h)$ to their limits in $(B.32)$. The plug-in method is valid because $\psi_K^*(h)$, $W(\psi_K, h)$, and $T_\mu(\psi_K, h)$ are all continuous functions of $S(\psi_K)$. The justification of this method is discussed in Lemma 6.6(b).

(b) The chi-square and standard normal distributions are directly implied by Lemma 6.6(b). □

Proof of Theorem 6.2.

(a) Let $T_{\pi_j,n}(\psi_K)$ be an empirical process indexed by ψ_K. For $j \le p - p_K$,

$$T_{\pi_j,n}(\psi_K) = \frac{n^{1/2}\widehat{\beta}_{j,n}(\psi_K)\left(\widehat{\pi}_{j,n}(\psi_K) - \pi_{j,n}\right)}{\left(R_j\widehat{\Sigma}_\pi(\psi_K) R_j'\right)^{1/2}}$$
$$= \frac{n^{1/2}\left(\widehat{\beta}_{j,n}(\psi_K) - \beta_{j,n}\right)\left(\widehat{\pi}_{j,n}(\psi_K) - \pi_{j,n}\right)}{\left(R_j\widehat{\Sigma}_\pi(\psi_K) R_j'\right)^{1/2}} + \frac{n^{1/2}\beta_{j,n}\left(\widehat{\pi}_{j,n}(\psi_K) - \pi_{j,n}\right)}{\left(R_j\widehat{\Sigma}_\pi(\psi_K) R_j'\right)^{1/2}} \Rightarrow T_{\pi_j}(\psi_K, h), \tag{B.33}$$

where the first term in the second inequality is $o_p(1)$ by Lemma B.3.

For $j > n - p_K$,

$$T_{\pi_j,n}(\psi_K) = \frac{n^{1/2}\widehat{\beta}_{j,n}(\psi_K)\left(\pi_j - \pi_{j,n}\right)}{\left(R_j\widehat{\Sigma}_\pi(\psi_K) R_j'\right)^{1/2}} \Rightarrow T_{\pi_j}(\psi_K, h) \tag{B.34}$$

by Lemma B.3. Note that the definitions of $T_{\pi_j}(\psi_K, h)$ are different for $j \leq n - p_K$ and $j > n - p_K$. Finally, we plug $\widehat{\psi}_{K,n}$ into $T_{\pi_j,n}(\psi_K)$ and plug $\psi_K^*(h)$ into $T_{\pi_j}(\psi_K, h)$. The plug-in method is valid because both $\psi_K^*(h)$ and $T_{\pi_j}(\psi_K, h)$ are continuous functions of $S(\psi_K)$.

(b) Part (b) is directly implied by Lemma 6.6(b). □

Proof of Lemma B.1.

(a) Part (a) can be proved in the same way as Lemma A.1(a).

(b) Replacing the sample projection matrix with the population projection matrix, we have

$$M_{k^+}(\psi_{k^+}) F_k(\psi_{k^-}, \psi_k) = \widetilde{s}_k(\psi_{k^-}, \psi_k | \psi_{k^+}) - f_{k^+}(\psi_{k^+}) \delta(\psi_{k^-}, \psi_k, \psi_{k^+}), \text{ where}$$
$$\delta(\psi_{k^-}, \psi_k, \psi_{k^+}) = \left(f_{k^+}(\psi_{k^+})' f_{k^+}(\psi_{k^+})\right)^{-1} (f_{k^+}(\psi_{k^+}) s_k(\psi_{k^-}), \psi_k) - \rho_{k^+}(\psi_{k^-}, \psi_k, \psi_{k^+}) \quad \text{(B.35)}$$

Using $(B.35)$, we have

$$n^{-1} s_k(\psi_{k^-}, \psi_k)' M_{k^+}(\psi_{k^+}) s_k\left(\psi_{k^-}, \overline{\psi}_k\right) \to_p \overline{\Phi}_k\left(\psi_{k^-}, \psi_k, \overline{\psi}_k | \psi_{k^+}\right), \quad \text{(B.36)}$$

because the rows of $\widetilde{s}_k(\psi_{k^-}, \psi_k)$ are *iid* by construction and $\delta(\psi_{k^-}, \psi_k, \psi_{k^+})$ is $o_p(1)$ uniformly over $\Pi_{k^-} \times \Pi_k \times \Pi_{k^+}$.

(c) Part (c) can be proved in the same way as Lemma A.1(b) by replacing $g(X_i, \pi_n) \beta_n - g(X_i, \pi) \widehat{\beta}(\pi)$ with

$$\delta_i^* = Z_i' \left(\zeta_n - \widehat{\zeta}(\psi_K)\right) + \left(f_{K^-}\left(\psi_{K^-,n}\right) \xi_{K^-,n} - f_{K^-}(\widehat{\psi}_{K^-}(\psi_K)) \widehat{\xi}_{K^-}(\psi_K)\right) +$$
$$f_K(\psi_K) \left(\xi_{K,n} - \widehat{\xi}_K(\psi_K)\right), \quad \text{(B.37)}$$

where the second term can be further simplified by Taylor expansion as in $(B.24)$. The rest of the proof is the same as in Lemma A.1(b). □

Proof of Lemma B.2: The proof follows that of Lemma A.2. □

C Proofs for Asymptotic Sizes and Robust CIs

Theorem 4.1 and Theorem 5.1 are special cases of Theorem 7.1 and Theorem 7.2, respectively, when $p = 1$. Hence, we only prove the general case in this section.

We first prove that the asymptotic distributions of the test statistics are all continuous $\forall h \in H$. The continuity property is used in the derivation of the asymptotic sizes of the CIs, as indicated in Corollary 3 of AG.

Lemma C.1 *Suppose Assumptions* 1, 2, 3c, 4c, *and* U *hold. For any* $h \in H$, $W(\psi_K^*(h), h)$, $T_\mu(\psi_K^*(h), h)$, *and* $T_{\pi,j}(\psi_K^*(h), h)$ *are all continuous distributions.*

Proof of Theorem C.1.

For notational simplicity, we omit h in this proof. For any $x \in R^r$,

$$P(W(\psi_K^*) \leq x) = E(1(W(\psi_K^*) \leq x)) = E(E(W(\phi) \leq x) | \psi_K^* = \phi))$$
$$= \int_{\Pi_K} \int_{-\infty}^{x} w(\phi, s) \, ds d\mu_\phi = \int_{-\infty}^{x} \left(\int_{\Pi_K} w(\phi, s) \, d\mu_\phi\right) ds, \quad \text{(C.1)}$$

where $w(\phi, s)$ is the normal density at s when ψ_K is fixed at ϕ, and μ_ϕ is the measure of ψ_K^*. The first equality holds by definition, the second equality holds by law of iterated expectation, the third equality holds because $W(\psi_K)$ is a r dimensional continuously distributed random variable for any fix ψ_K, and the the last equality is from Fubini's Theorem. We conclude $W(\psi_K^*)$ is a continuous distribution because it is absolute continuous wrt the Lebesgue measure as shown in $(C.1)$. Analogously, we can show $T_\mu(\psi_K^*(h), h)$, and $T_{\pi,j}(\psi_K^*(h), h)$ are both continuous distributions for any given h. □

Proof of Theorem 7.1.

This is an application of Corollary 3 of AG. The key assumption, Assumption B in AG, is verified by Theorem 6.1 and Theorem 6.2. The other assumptions can be verified as in AG, p. 20-21. Theorem 4.1 is a special case of Theorem 7.1 when $p = 1$. □

Theorem 5.1 is a special case of Theorem 7.2 when $p = 1$. The asymptotic distribution of the test statistic $W_n(\mu_{R,n})$, $T_{\mu,n}(\mu_{R,n})$, and $T_{\pi_j,n}(\pi_{j,n})$ are given in Theorem 6.1 and Theorem 6.2. We denote the test statistic in general by $T_n(\theta_n)$ and call its asymptotic distribution J_h, with the $1-\alpha$ quantile $c_h(1-\alpha)$.

Parameter $\eta = (\eta_1, ..., \eta_p)'$ is defined below in Lemma C.2. Let $\eta^* = (\eta_1^*, ..., \eta_p^*)'$, where $\eta_j^* = \infty$ if $\eta_j \neq 0$. Given η, define

$$c_{\eta^*}(1-\alpha) = \sup_{h_\eta \in H} c_{h_\eta}(1-\alpha), \text{ where } h_\eta = (\eta^{*\prime}, \pi_0')'.^{16} \tag{C.2}$$

Lemma C.2 *Suppose Assumptions* 1, 2, 3c, 4c, U, *and* R *hold. Let* $h = (b', \pi_0')'$ *and* $\{\gamma_{n,h} = (\beta_n', \pi_n', \zeta_n')' : n \geq 1\}$ *be a sequence of points in its parameter space that satisfies*
(**i**) $n^{1/2}\beta_n \to b$ *for some* $b \in R^p_{[\pm\infty]}$,
(**ii**) $\kappa_n^{-1} n^{1/2} \beta_{j,n} \to \eta_j$ *for some* $\eta_j \in R_{[\pm\infty]}$, *and*
(**iii**) $\pi_n \to \pi_0$ *for some* $\pi_0 \in \Pi$. *Then,*
(**a**) $\widehat{c}_n(1-\alpha) \geq c_n^*$ *for all* n *for a sequence of random variables* $\{c_n^* : n \geq 1\}$ *that satisfies* $c_n^* \to_p c_{\eta^*}(1-\alpha)$ *under* $\{\gamma_{n,h} : n \geq 1\}$.
(**b**) $\liminf_{n\to\infty} P_{\gamma_{n,h}}(T_n(\theta_{n,h}) \leq \widehat{c}_n(1-\alpha)) \geq 1-\alpha$.

Proof of Lemma C.2.

Let $\widetilde{\eta} = (\widetilde{\eta}_1, ..., \widetilde{\eta}_p)$, where $\widetilde{\eta}_j = \infty$ if $\eta_j \neq 0$ **or** $\varphi_{j,n} > 1$, where $\varphi_{j,n} = \kappa_n^{-1} t_{j,n}$. Given η, define

$$c_n^* = \sup_{\widetilde{h} \in H} c_{\widetilde{h}}(1-\alpha), \text{ where } \widetilde{h} = (\widetilde{\eta}', \pi_0)'. \tag{C.3}$$

When comparing c_n^* and $c_{\eta^*}(1-\alpha)$, we only need to consider those dimensions with $\eta_j = 0$, because $\widetilde{\eta}_j = \eta_j^* = \infty$ when $\eta_j \neq 0$. Note that

$$c_n^* = c_{\eta^*}(1-\alpha), \text{ when } \varphi_{j,n} \leq 1 \text{ for all } j = 1, ..., p_\varphi \text{ such that } \eta_j = 0. \tag{C.4}$$

Then $\widehat{c}_n(1-\alpha) \geq c_n^*$ by construction because the probability that $\widetilde{\eta}_j = \infty$ is greater than that of $b_j^* = \infty$. By making restrictions such as $\widetilde{\eta}_j = \infty$, we take supremum over a smaller parameter space.

Next, we need to show $c_n^* \to_p c_{\eta^*}(1-\alpha)$. Under $\{\gamma_{n,h} : n \geq 1\}$, we have

[16] Parameter η and critical value $c_{\eta^*}(1-\alpha)$ here correspond to π and $c_{\pi^*}(1-\alpha)$ in Section 12.3 of AS, respectively.

$$\varphi_{j,n} = \kappa_n^{-1} \left| \frac{n^{1/2}\widehat{\beta}_{j,n}}{\widehat{\sigma}_{\beta_j,n}} \right| = \kappa_n^{-1} \left| \frac{n^{1/2}\left(\widehat{\beta}_{j,n} - \beta_{j,n}\right)}{\widehat{\sigma}_{\beta_j,n}} + \frac{n^{1/2}\beta_{j,n}}{\widehat{\sigma}_{\beta_j,n}} \right| \leq o_p(1) + \left| \frac{n^{1/2}\beta_{j,n}}{\kappa_n \widehat{\sigma}_{\beta_j,n}} \right| \rightarrow_p \left| \frac{\eta_j}{\sigma_{\beta_j}} \right|, \tag{C.5}$$

where σ_{β_j} is the limit of $\widehat{\sigma}_{\beta_j,n}$. Using results in Section 6.2,

$$\sigma_\beta = \left(R_{q+j}\Sigma_\mu(\widetilde{\pi})R'_{q+j}\right)^{1/2}, \text{ where } \widetilde{\pi} = \begin{cases} \pi^*(h), & \text{if } n^{1/2}\xi_K \to b_K \in R^{p_K} \\ \pi_0, & \text{if } \left\|n^{1/2}\xi_K\right\| \to \infty. \end{cases} \tag{C.6}$$

and R_{q+j} is a row vector with the $(q+j)th$ element being one and the rest being 0. Hence,

$$P\left(|c_n^* - c_{\eta^*}(1-\alpha)| \leq \varepsilon\right) > \prod_{i=1}^{p_\varphi} P\left(\varphi_{j,n} \leq 1\right) \to 1 \text{ for any } \varepsilon > 0, \tag{C.7}$$

because $\varphi_{j,n} \rightarrow_p 0$ when $\eta_j = 0$.

(b) We first compare $c_{\eta^*}(1-\alpha)$ and $c_h(1-\alpha)$. To this end, we need to check whether the restriction that $h = (\eta^{*\prime}, h_2')'$ imposed on $c_{\eta^*}(1-\alpha)$ is satisfied by the true localization parameter. If this is true, taking supremum over a space include the true parameter leads to $c_{\eta^*}(1-\alpha) \geq c_h(1-\alpha)$.

The only restriction on η^* is that when $\eta_j \neq 0$, $\eta_j^* = \infty$. By definition of η_j and the assumption that $\kappa_n \to \infty$, we know $h_{1,j} = \infty$. As such the condition on η_j^* is satisfied by the true parameter, which implied that $c_{\eta^*}(1-\alpha) \geq c_h(1-\alpha)$.

Because $T_n(\theta_n) \Rightarrow J_h$ under $\{\gamma_n\}$, we now have

$$\liminf_{n\to\infty} P_{\gamma_{n,h}}\left(T_n(\theta_n) \leq \widehat{c}_n(1-\alpha)\right) \geq \liminf_{n\to\infty} P_{\gamma_{n,h}}\left(T_n(\theta_n) \leq c_n^*\right) \geq J_h\left(c_{\eta^*}(1-\alpha)-\right) \tag{C.8}$$

where $J_h(x-)$ denotes the limit from the left of J_h at x, the first inequality holds because $\widehat{c}_n(1-\alpha) \geq c_n^*$ and the second inequality holds by part (a) of the Lemma and $T_n(\theta_n) \Rightarrow J_h$. Since J_h is continuous by Lemma C.1, we have

$$J_h\left(c_{\eta^*}(1-\alpha)-\right) = J_h\left(c_{\eta^*}(1-\alpha)\right) \geq 1-\alpha, \tag{C.9}$$

where the inequality holds by $c_{\eta^*}(1-\alpha) \geq c_h(1-\alpha)$. □

Lemma C.3 *Suppose Assumptions* 1, 2, 3c, 4c, U, *and* R *hold. The* $1-\alpha$ *quantile of* $W\left(\psi_K^*(h), h\right)$, $T_\mu\left(\psi_K^*(h), h\right)$, *and* $T_{\pi,j}\left(\psi_K^*(h), h\right)$ *are all continuous wrt* h, *where* $h = (b', \pi_0)'$.

The proof of Lemma C.3 is at the end of the section. The continuity property in Lemma C.3 is useful in the proof of Lemma C.4 below.

The next Lemma is analogous to Lemma 3 of Andrews and Soares (2007). Here we need to find the sequence of true parameters $\left\{\gamma_n^* = (\beta_n^{*\prime}, \pi_n^{*\prime}, \zeta_n^{*\prime})' : n \geq 1\right\}$ under which the asymptotic coverage probability of the robust CI is exactly $1-\alpha$, which shows that the robust CI is not asymptotically conservative.

Let

$$\overline{h} = \arg\max_{h\in H} c_h(1-\alpha), \tag{C.10}$$

where $\overline{h} = (\overline{b}', \overline{\pi}_0')'$ and $\overline{b} = (\overline{b}_1, ..., \overline{b}_p)'$. Note that we have shown the continuity of $c_h(1-\alpha)$ wrt h and the parameter space of b_j includes ∞. Hence, the maximum can be attained at $\overline{h}$.

Let $\beta_n^* = (\beta_{1,n}^*, ..., \beta_{p,n}^*)'$. For $j = 1, ..., p$, the sequence we need is

$$\beta_{j,n}^* = \begin{cases} n^{-1/2}\overline{b}_j, \text{ if } \overline{b}_j \in R \\ \beta_j^* \neq 0, \text{ if } \overline{b}_j = \infty \end{cases}, \ \pi_n^* = \overline{\pi}_0, \text{ and } \zeta_n^* = \zeta_0 \in R^q. \tag{C.11}$$

Lemma C.4 *When the true distribution is determined by γ_n^* for all n, we have* $\lim_{n\to\infty} P_{\gamma_n^*}(T_n(\theta_n^*) \leq \widehat{c}_n(1-\alpha)) = 1-\alpha$.

Proof of Lemma C.4.

By construction, $c_{\overline{h}}(1-\alpha)$ is the $1-\alpha$ quantile of the asymptotic distribution $J_{\overline{h}}$ under γ_n^*. Let $\widehat{c}_{\gamma^*}(1-\alpha)$ be the robust critical value under γ_n^*. To get the desired result, we need to show $\widehat{c}_{\gamma^*}(1-\alpha) \to c_{\overline{h}}(1-\alpha)$. When $\overline{b}_j = \infty$, the model-selection procedure on b_j does not affect the maximization of $c_h(1-\alpha)$ over b_j. With either testing result, $\widehat{c}_{\gamma^*}(1-\alpha)$ and $c_{\overline{h}}(1-\alpha)$ are both attained at $b_j = \infty$. Therefore, we only need to consider $\overline{b}_j \in R$ and check whether the model selection narrow the optimization space for $\widehat{c}_{\gamma^*}(1-\alpha)$. By definition,

$$\widehat{c}_{\gamma^*}(1-\alpha) = c_{\overline{h}}(1-\alpha) \text{ if } \varphi_{j,n} \leq 1 \text{ for all } j = 1, ..., p. \tag{C.12}$$

When $\overline{b}_j = 0$, $\eta_j = 0$ by the definition. In this case, $\varphi_{j,n} \to_p 0$, so that $\varphi_{j,n} \leq 1$ with probability 1. Thus,

$$P\left(\left|\widehat{c}_{\gamma^*}(1-\alpha) - c_{\overline{h}}(1-\alpha)\right| \leq \varepsilon\right) > \prod_{i=1}^{p} P\left(\varphi_{j,n} \leq 1\right) \to 1, \tag{C.13}$$

for any $\varepsilon > 0$. Then

$$\lim_{n\to\infty} P_{\gamma_n^*}(T_n(\theta_n^*) \leq \widehat{c}_{\gamma^*}(1-\alpha)) = J_{\overline{h}}\left(c_{\overline{h}}(1-\alpha)\right) = 1-\alpha, \tag{C.14}$$

where the first equality holds by Lemma 3 of AG1 and the second equality holds by Lemma C.1. □

Proof of Theorem 5.1.

The proof is the same as that of Theorem 1 in AS by replacing Lemma 2 and Lemma 3 in AS with Lemma C.2 and Lemma C.4, respectively. □

Proof of Lemma C.3.

We write $J_h = T(\psi_K^*(h), h)$, where $T(\psi_K, h)$ is a stochastic process indexed by ψ_K.

Let $S^*(\psi_K)$ be a continuous sample path of the Gaussian process $S(\psi_K)$. Let $Q(S(\psi_K), \psi_K, h)$ denote the non-central chi-square process on the rhs of Lemma 6.5(a). We use sup norm to measure the distance between two continuous functions. The sample path of $Q(S(\psi_K), \psi_K, h)$, denoted by $Q(S^*(\psi_K), \psi_K, h)$, is continuous in h. Because $Q(S^*(\psi_K), \psi_K, h)$ has unique minimizer with probability one, $\psi_K^*(h)$ is continuous wrt h, a.s. $[S^*(\psi_K)]$.

Next, let h_n be a sequence converges to h_0. Conditional on $S^*(\psi_K)$,

$$\begin{aligned} &|T(\psi_K^*(h_n), h_n) - T(\psi_K^*(h_0), h_0)| \\ \leq\ &|T(\psi_K^*(h_n), h_n) - T(\psi_K^*(h_n), h_0)| + |T(\psi_K^*(h_n), h_0) - T(\psi_K^*(h_0), h_0)|, \end{aligned} \tag{C.15}$$

by triangle inequality. In $(C.15)$, the first term on the rhs converges to 0 because $T(\psi_K, h)$ is continuous wrt h under sup norm, and the second term on the rhs converges to 0 because the sample path is continuous and $\psi_K^*(h_n) \to \psi_K^*(h_0)$.

Using $(C.15)$, we have

$$\begin{aligned} T\left(\psi_K^*\left(h_n\right), h_n\right) &\to T\left(\psi_K^*\left(h_0\right), h_0\right) \text{ a.s. } \left[S^*\right], \\ 1\left(T\left(\psi_K^*\left(h_n\right), h_n\right) \leq x\right) &\to 1\left(T\left(\psi_K^*\left(h_0\right), h_0\right) \leq x\right) \text{ a.s. } \left[S^*\right], \\ P\left(T\left(\psi_K^*\left(h_n\right), h_n\right) \leq x\right) &\to P\left(T\left(\psi_K^*\left(h_0\right), h_0\right) \leq x\right). \\ c_{h_n}\left(1-\alpha\right) &\to c_{h_0}\left(1-\alpha\right) \end{aligned} \tag{C.16}$$

The fourth convergence result of $(C.16)$ holds by the third result and Lemma C.1. The third convergence result of $(C.16)$ holds by the second result and the bounded convergence theorem. The second convergence result of $(C.16)$ follows from the first result and Lemma C.1. □

Reference

Andrews, D. W. K. (1992): "Generic Uniform Convergence," *Econometric Theory*, 8, 241-257.

—— (1994): "Empirical Process Method in Econometrics," *Handbook of Econometrics*, Vol. IV. Edited by R. F. Engel and D. L. McFadden. Ch. 37, 2248-2294.

—— (1999): "Estimation When a Parameter Is on a Boundary," *Econometrica*, 67, 1341-1384.

—— (2000): "Inconsistency of the Bootstrap When a Parameter Is on the Boundary of the Parameter Space," *Econometrica*, 68, 399-405.

Andrews, D. W. K. and P. Guggenberger (2007): "Applications of Subsampling, Hybrid, and Size-Correction Methods," Cowles Foundation Discussion Paper No. 1608, Yale University.

—— (2009a): "Asymptotic Size and a Problem with Subsampling and with the m out of n Bootstrap," forthcoming in *Econometric Theory*.

—— (2009b): "Hybrid and Size-Corrected Subsampling Methods," forthcoming in *Econometrica*.

—— (2009c): "Validity of Subsampling and 'Plug-in Asymptotic' Inference for Parameters Defined by Moment Inequalities," forthcoming in *Econometric Theory*.

Andrews, D. W. K. and P. Jia (2008): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," Cowles Foundation Discussion Paper No. 1676, Yale University.

Andrews, D. W. K. and W. Ploberger (1994): "Optimal Tests when a Nuisance Parameter Is Present Only under the Alternative," *Econometrica*, 62, 1383-1414

Andrews, D. W. K. and G. Soares (2007): "Inference for Parameters Defined by Moment Inequalities using Generalized Moment Selection," Cowles Foundation Discussion Paper No. 1631, Yale University.

Chernozhukov V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75, 1243-1284.

Davies, R. B. (1977): "Hypothesis Testing when a Nuisance Parameter is Present Only under the Alternative," *Biometrika*, 64, 247-254.

—— (1987): "Hypothesis Testing when a Nuisance Parameter is Present Only under the Alternative," *Biometrika*, 74, 33-43.

Dufour, J.-M. (1997): "Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models," *Econometrica*, 65, 1365–1387.

Hansen, B. E. (1996): "Inference When a Nuisance Parameter Is Not Identified under the Null Hypothesis," *Econometrica*, 64, 413-430.

—— (2007): "Least Squares Model Averaging," *Econometrica*, 75, 1175-1189.

Kim, J. and D. Pollard (1990): "Cube Root Asymptotics," *Annals of Statistics*, 18, 191-219.

Mikusheva, A. (2007): "Uniform Inference in Autoregressive Models," *Econometrica*, 75, 1411-1452.

Romano, J. P. and A. M. Shaikh (2006): "Inference for the Identified Set in Partially Identified Econometric Models," Technical Report 2006–10, Department of Statistics, Stanford University.

—— (2008): "Inference for Identifiable Parameters in Partially Identified Econometric Models," *Journal of Statistical Planning and Inference*, 138, 2786-2807.

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557-586.

Tripathi, G. (1999): "A Matrix Extension of the Cauchy-Schwarz Inequality," *Economics Letters*, 63, 1-3.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*. New York: Springer.