BOOTSTRAPPING MAX TESTS IN THE PRESENCE OF WEAK IDENTIFICATION

John W. Dennis, III

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Economics.

Chapel Hill
2019

Approved by:

Jonathan B. Hill

Peter R. Hansen

Ju Hyun Kim

Valentin Verdier

Andrii Babii

ABSTRACT

JOHN W. DENNIS, III: Bootstrapping Max Tests in the Presence of Weak Identification.
(Under the direction of Jonathan B. Hill)

Traditional inference can be distorted in the presence of weakly identified parameters. I explore the effects of weak identification on inference in two different scenarios that have been previously unaddressed. First, I consider the effect of weak identification on a test for serial correlation, where I demonstrate that the distortion arising from weakly identified parameters propagates to test statistics that are not directly testing the parameter values. I show that existing tests can be extended to accommodate known sources of identification failure via a modification of the first order expansion utilized by the bootstrap. Second, I examine inference on a large dimensional parameter when some of the parameter elements may be weakly identified. Existing tests cannot simultaneously accommodate identification failure and a parameter vector with large dimension. In both scenarios, I provide testing procedures that accommodate weak identification, are based on a maximum value, and are implemented with a bootstrap. The efficacy of these testing procedures are explored in several Monte-Carlo simulations, and empirically relevant examples are discussed.

Caileigh, thank you.

ACKNOWLEDGMENTS

This project would have never been completed without the guidance and support of my advisor, Jonathan Hill. I am also grateful to my dissertation committee members, Peter Hansen, Ju Hyun Kim, Valentin Verdier, and Andrii Babii for their guidance and encouragement and for helpful comments from participants in the UNC - Chapel Hill Econometrics Dissertation Workshop.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND RELATIONSHIP WITH THE LITERATURE

I describe the use of max tests in the presence of weak identification. This collection informs the reader of some issues that can arise in common inferential analyses in Economics due to the presence of weak identification in the parameters and presents solutions to these problems within the framework of a convenient testing procedure. I consider two practical inferential problems, and for each I discuss how the presence of weak identification can lead to distorted inference and present a method to conduct inference in an appropriate manner.

The two classes of inferential problems I consider are white noise tests and tests on a large dimensional parameter. White noise tests fall within the class of model diagnostic tests; they are designed to aid the practitioner in determining if a particular model is able to capture all of the serial correlation present in the data. In this sense, they can be thought of as one of a group of tests that examines the adequacy of the model in describing the data. In the first paper, I present a white noise test that is appropriate for residuals from estimated models and is robust to parameter identification failure in the model.

The second class of tests has become a focus in the literature in recent years due to the vast quantities of data that have become available to researchers. In particular, researchers often have many variables in a dataset leading to many objects that must be estimated and tested. In the second paper, I present a test for many zero restrictions in a model with a large dimensional parameter when many of the parameter elements may be only weakly identified.

The frameworks presented in both of the problems I consider utilize a bootstrapping procedure to simulate the distribution of a maximum test statistic in the presence of weakly identified parameters. Here I discuss these topics and the relationship to the econometric literature broadly. I discuss these topics in more detail as they pertain to each of the two classes of problems that I

consider in the relevant chapters below.

Identification failure is present in both problems that I consider, but here I must be specific regarding the meaning of identification failure. Lewbel (ming) indicates the term identification failure appears in more than two dozen forms in the literature, but all share a common underlying meaning. In particular, an object is not identified if its true value cannot be uniquely determined in the population. I specifically use utilize the definition of Andrews and Cheng (2012a) to describe identification failure as the situation in which there is a known parametric source of identification failure for a parameter in the model under consideration. The framework of Andrews and Cheng (2012a) is convenient for the econometrician in describing parametric identification failure, as it allows a range of identification behaviors to exist between identification and non-identification.

Consider estimating scalar parameters (β, π) from the nonlinear function $Y_t = \beta g(X_t, \pi) + \varepsilon_t$ for some smooth non-linear function g. It is well known that when $\beta \neq 0$, π can be (strongly) identified, and when $\beta = 0$, π cannot be identified. In order to develop a unifying testing framework, we utilize a thought experiment which can be characterized by using the notion of drifting sequences of true parameters. Let $\beta = \beta_n$ be a sequence of true parameters, indexed by the sample size n, that are drifting to 0. Then the strength of identification of π is categorized by the speed at which $\beta_n \to 0$. When $\sqrt{n}\beta_n \to \infty$, we characterize π as being semi-strongly identified, and when $\sqrt{n}\beta_n \to b \in (0, \infty)$, we say π is weakly identified. In the latter case, our estimator $\hat{\pi}_n$ is not consistent for the true π_0, and converges instead to a random variable under certain conditions. Table 1 from Andrews and Cheng (2012a) details these categories. It is important to note that in this literature, the parametric source of identification failure is known. More recently, Han and McCloskey (2016) develop theory for the case in which the source of identification failure may be unknown. We focus on the former case and leave this extension for future research.

For the cases of non-identification and weak identification, the estimators for π are inconsistent. Further, in these cases the estimator for β is consistent; however, it is a function of $\hat{\pi}_n$ which converges to a random variable, resulting in a non-standard distribution for $\hat{\beta}_n$. This implies that the resulting test statistics will exhibit non-standard behavior, yielding distorted inference from classical tests. In this case, the asymptotic distribution of the test statistics will be nonstandard.

Table 1.1: Identification Categories: Andrews and Cheng's (2012a) Table I

Category	$\{\beta_n\}$ Sequence	Identification Property of π_0
I(a)	$\beta_n = 0 \; \forall n \geq 1$	Unidentified
I(b)	$\beta_n \neq 0$ and $n^{1/2}\beta_n \to b \in \mathbb{R}^{d_\beta}$	Weakly identified
II	$\beta_n \to 0$ and $n^{1/2}\|\|\beta_n\|\| \to \infty$	Semi-strongly identified
III	$\beta_n \to \beta_0 \neq 0$	Strongly identified

This poses problems for tests based on residuals from model estimation. Non-standard behavior of the estimators propagates through to the test statistic, yielding a non-standard distribution for the test statistic and resulting in potentially distorted inference from traditional tests.

Further, this is an issue for economic practitioners, as many commonly used models in Economics include parameters that may be unidentified in certain parts of the parameter space. Examples such as Dynamic Stochastic General Equilibrium models (Guerron-Quintana, Inoue, and Kilian, 2013; Andrews and Mikusheva, 2015), Smooth Transition AutoRegressive models (Terasvirta, 1994; Teräsvirta, 1998; van Dijk, Teräsvirta, and Franses, 2002; Andrews and Cheng, 2013), Probit models (Andrews and Cheng, 2012a, 2014) and Nonlinear Binary Choice Models (Andrews and Cheng, 2013), nonlinear instrumental variables models with possibly weak instruments (Andrews and Cheng, 2012a, 2014), ARMA models Andrews and Ploberger (1996); Andrews and Cheng (2012a); Dennis (2019), Regime Switching Models (Chen, Fan, and Liu, 2016) and Fuzzy Regression Discontinuity Designs (Feir, Lemieux, and Marmer, 2016), models based on moment conditions and GMM (Andrews and Cheng, 2014), and MiDAS Regressions (Ghysels, Hill, and Motegi, 2016b) have been shown to include model components that may not be identified in certain regions of the parameter space.

Missing from the analysis of Andrews and Cheng (2012a) is the ability to account for models with mixed identification strength, referring to models which may simultaneously include parameters from each from each of the identification categories (Cheng, 2015). Consider the simple model $Y_t = \beta_1 g(X_t, \pi_1) + \beta_2 g(X_t, \pi_2) + \varepsilon_t$ where ε_t is independent of X_t and with the null hypothesis

$H_0 : \beta = 0$. Under this null hypothesis, the π_j are unidentified nuisance parameters, so this framework is related to the literature on testing with nuisance parameters under the null (Davies, 1977, 1987; Andrews and Ploberger, 1994; Hansen, 1996; Stinchcombe and White, 1998; Ghysels and Guay, 2004; Andrews and Mikusheva, 2016). Nuisance parameters cause the test statistics to have non-standard distributions, which often do not have analytic expressions and must be simulated.

In this framework, however, each parameter π_j may exhibit its own degree of identification strength, so a uniformly valid test becomes necessary. Andrews and Cheng (2012a, 2013, 2014) discuss uniformly valid inference but do not allow for mixed identification strength. Cheng (2015) offers the first uniformly valid inference procedure for inference on sub-vectors of β allowing for mixed identification strength but limits her theory to additive nonlinear models.

Andrews and Cheng (2012a, 2013, 2014) and Cheng (2015) discuss inference under weak identification but do not consider large dimensional parameters or max test statistics, implementation of a bootstrap, or tests on objects from estimated models, such as white noise tests, that are not tests directly on the model parameters. In contrast, in the first paper, we consider white noise tests based on the maximum of a sequence of correlations that we implement with a bootstrap, and in the second paper, we construct a test based on the maximum of a sequence of estimated parameters from a high dimensional parameter.

The testing procedures that I consider in both classes of problems are based on max tests. When testing the maximum value in a sequence, we are often interested in determining if any of the parameter elements are different from zero. In considering only the maximum from the sequence of values, the max test statistic utilizes the most informative measure available from our data, eliminating issues that arise from low degrees of freedom and inversion of large or near singular covariance matrices when a large number of variables needs to be tested (Hill and Dennis, 2018; Ghysels, Hill, and Motegi, 2016a), or by combining noisy estimates, which occurs when calculating serial correlations at long lags (Hill and Motegi, 2018).

Statistics based on a maximum of a sequence of values is an extensively studied topic in the

literature[1] dating at least to Fisher and Tippet (1928) and Gnedenko (1943). See also Gumbel (1958) and Berman (1964). Typically in this literature, extreme value theory arguments appeal to the Extremal Types Theorem to determine the exact asymptotic distribution of the maximum statistic (de Haan, 1976). For example, Xiao and Wu (2014) provide a test for serial correlation for observed sequences using the maximum sample autocovariance and show that under suitable normalization, the test statistic converges in distribution to a Gumbel (type I extreme value) distribution. These arguments require that when the data are divided into blocks, the dependence between increasingly distant blocks decays at a sufficient rate as with a mixing condition.

Hill and Motegi (2018); Hill and Dennis (2018) argue that when estimating parsimonious models, allowing for general dependence in the data generating process, or residuals to be used in the max statistic, the classical extreme value theory arguments are no longer straight forward to prove and may require more stringent assumptions than are needed by other methods. Further, extreme value theoretic arguments for establishing the limiting distribution of the maximum of a sequence of values often relies on Gaussianity of the underlying sequence. Hill and Motegi (2018); Hill and Dennis (2018) develop theory that does not rely on Gaussianity and that allows the use of the dependent wild bootstrap (Shao, 2010, 2011a) to mimic the finite sample distribution of the max statistic.

For these reasons, I simulate the distribution of the test statistics with forms of a Wild, or Gaussian multiplier, bootstrap (Wu, 1986; Liu, 1988). Methods for bootstrapping high dimensional statistics have not been available until recently. Chernozhukov, Chetverikov, and Kato (2013, 2017) develop a theory that is able to both bypass the typical extreme value theoretic asymptotic arguments and deliver an impressive growth rate for the sequence being examined.[2] However, they require independence, and their theory is only appropriate for observed random variables and relies on Gaussian approximation that is not appropriate for approximations of non-Gaussian normalized summands. Zhang and Cheng (2018) extend the Gaussian approximation theory in Chernozhukov

[1] See Leadbetter, Lindgren, and Rootzèn (1983) and Resnick (1987) for textbook treatments.

[2] See also Belloni, Chernozhukov, Chetverikov, Hansen, and Kato (2018).

et al. (2013, 2017) to allow for dependence, but only allow for observed random variables. Zhang and Wu (2017) develop theory for a Gaussian approximation for high dimensional times series but only allow for observed sequences as well. The theory in Hill and Dennis (2018); Hill and Motegi (2018) is also able to bypass extreme value theoretic arguments, allows for dependence under the null, and is appropriate for residuals. For this reason, we rely on the theory developed in Hill and Motegi (2018); Hill and Dennis (2018).

This collection is organized as follows. Chapter 2 presents the test for serial correlation in models with weakly identified parameters. Chapter 3 presents the test for a large dimensional parameter when the parameter elements may exhibit mixed identification strength. Proofs of the results are collected in Appendices A and B.

CHAPTER 2

TESTING WHITE NOISE WHEN SOME PARAMETERS MAY BE WEAKLY IDENTIFIED

2.1 Introduction

We develop a bootstrapped white noise test for residuals that is based on the maximum correlation and is robust to parameter identification failure in the model. It is well known that the asymptotic and finite sample distributions of estimators are non-standard when the model contains parameters that are weakly identified, and that standard inference based on t or χ^2 distributions can be distorted. For example, Andrews and Cheng (2012a) demonstrate in their figures 1 and 2 that densities of the estimators from an ARMA(1,1) model can be quite different from normal when the AR and MA parameters are close to the same value. Further, Cheng (2015) shows in her table 1 that using standard normal critical values for tests on a parameter from an additive nonlinear model with a weakly identified parameter can generate large size distortions.

The impact of identification failure on the distributions of the estimators for a model can propagate beyond tests on the parameter values. When the test statistic is based on an estimated model, the usual method to either prove the asymptotic distribution of the test statistic or to implement a finite sample correction via a bootstrap is to utilize a first order expansion of the test statistic that involves the distribution of the parameter estimators. This enables inference on the test statistic to properly account for the impact of model estimation.

Traditional methods assume that the distribution of the estimators is normal, an assumption that is not true when some of the parameters are weakly identified. In particular, we show that the distribution of our white noise max test statistic differs under weak and strong identification, and we demonstrate that ignoring the effect of weakly identified parameters can lead to size distortions. We provide a robust procedure that allows a correctly sized, consistent test for the null hypothesis of uncorrelated errors when the strength of identification in the estimated model is not known.

In particular, this is an issue for economic practitioners engaging in model diagnostic activities, as many commonly used models in Economics include parameters that may be unidentified in certain parts of the parameter space. Examples such as Smooth Transition AutoRegressive models (Terasvirta, 1994; van Dijk et al., 2002; Andrews and Cheng, 2013), Probit models (Andrews and Cheng, 2012a, 2014) and Nonlinear Binary Choice Models (Andrews and Cheng, 2013), nonlinear instrumental variables models with possibly weak instruments (Andrews and Cheng, 2012a, 2014), ARMA models (Andrews and Ploberger, 1996; Andrews and Cheng, 2012a), Regime Switching Models (Chen et al., 2016) and Fuzzy Regression Discontinuity Designs (Feir et al., 2016), Dynamic Stochastic General Equilibrium models (Guerron-Quintana et al., 2013; Andrews and Mikusheva, 2015), models based on moment conditions and GMM (Andrews and Cheng, 2014), and MiDAS Regressions (Ghysels et al., 2016b) have been shown to include model components that may not be identified in certain regions of the parameter space. The models above are often used under the assumption of a white noise error term. The current paper focuses on testing if the error term is a white noise process while allowing for some model parameters to be unidentified in parts of the support of the parameter space. The test presented in this paper, then, can be viewed as a test of model adequacy for models such as those mentioned above, which may have identification failure in regions of the parameter space.

There are three key components that characterize this test. First, this test is a white noise correlation test for residuals that only requires uncorrelatedness under the null. Allowing for residuals requires that we account for the influence of the estimated parameters on our test statistic. In particular, we allow for models in which some parameters may be non- or weakly identified in the sense of Andrews and Cheng (2012a), leading to inconsistent estimators. Utilizing a first order expansion of our test statistic about the point of identification failure allows us to account for the influence of the estimated parameters without the need for a consistent estimator for the parameters that are not identified.

Second, the test statistic is formed using the maximum from an increasing sequence of sample correlations. Using the maximum correlation allows for a sharper statistic in the sense that, unlike

a traditional portmanteau test which utilizes the sum of all squared sample correlations, the maximum statistic only focuses on the most informative sample correlation and, therefore, mitigates both the issue of washing out a single non-zero correlation in an average when testing for potential correlations at many lags and issues stemming from noisy sample correlation estimates that can occur at long lags. Further, max statistics are convenient for high dimensional objects, as they do not require inversion of a large covariance matrix.

Traditional arguments for statistics of a maximum value rely on proving asymptotic convergence via the extremal types theorem (de Haan, 1976), but these arguments are typically for observed sequences, and bootstraps are not typically considered for finite sample improvement. Xiao and Wu (2014) discuss a similar maximum statistic and prove that under suitable normalizing constants their test statistic converges to a type I extreme value distribution; however, they do not allow for residuals, an important distinction that affects the extreme value theory asymptotic argument, and they do not prove the validity of their bootstrap. Further, the extreme value theory approach requires restrictions to ensure convergence that are not necessary under our bootstrapping approach. We bypass standard extreme value theory arguments by use of theory in Hill and Dennis (2018) and Hill and Motegi (2018) paired with the dependent wild bootstrap of Shao (2010, 2011a).

Finally, our test is robust against identification failure of the model parameters. In order to account for the influence parameter estimation on our test statistic, we incorporate a first order expansion that involves the distribution of the parameter estimators. Whether or not our model has weakly identified parameters affects the terms in the expansion of the test statistic, which correspondingly affects the limiting distributions and the objects that must be bootstrapped. This suggests that traditional model diagnostics on estimated models with weakly identified parameters may lead to tests with size distortions, as we demonstrate in our simulations. We show that modifying the first order expansion utilized by the bootstrap to account for the dependence of the test statistic upon the estimated parameters can mitigate this distortion. Namely, when a consistent estimator is not available, we perform our expansion about the point of identification failure. In practice, we do not know whether consistent estimators are available or not for our potentially

unidentified parameters. We construct our identification robust test by bootstrapping the test statistic under both scenarios and stitching the resulting critical values together with an identification pre-test as discussed in Andrews and Cheng (2012a).

Throughout the paper, we assume a general model for the residuals from a regression model, which we denote $\varepsilon_t(\theta)$, where θ are the parameters of the model. For clarity, we elaborate the details for our test with an ARMA model (e.g. $Y_t = \beta Y_{t-1} + \varepsilon_t - \pi\varepsilon_{t-1}$) and an additive nonlinear model, an example of which is the Smooth Transition Autoregressive model of Terasvirta (1994): $Y_t = \beta X_t \times g(Z_t, \pi) + \zeta X_t + \varepsilon_t$, where X_t typically contains lags of Y_t, g is a smooth, nonlinear function[1] and ε_t is the model error, which we estimate with the regression residuals $\varepsilon_t(\hat{\theta}_n)$. Our goal is to test if $\{\varepsilon_t\}$ is a white noise process:

$$H_0 : \rho(h) = 0 \; \forall h \in \mathbb{N} \;\text{ vs. }\; H_A : \rho(h) \neq 0 \;\text{ for some } h \in \mathbb{N}$$

where $\rho(h) = E(\varepsilon_t\varepsilon_{t-h})/E(\varepsilon_t\varepsilon_t)$. To test this hypothesis, we specifically consider the sample max correlation statistic (Hill and Motegi, 2018)

$$\hat{\mathcal{T}}_n = \sqrt{n} \max_{1\leq h\leq \mathcal{L}_n} |\hat{\rho}_n(h)|,$$

where $\{\mathcal{L}_n\}$ is a sequence of integers with $\mathcal{L}_n \to \infty$ as $n \to \infty$, $\mathcal{L}_n = o(n)$ allowing for a true white noise test.[2] We utilize the dependent wild bootstrap (Shao, 2010, 2011a) paired with an expansion of our test statistic to account for the dependence upon the estimated parameters. This allows our test to be appropriate for residuals as in Hill and Motegi (2018); however, the test in Hill and Motegi (2018) requires consistency of all parameter estimators and, thus, cannot accommodate models in which some parameters are weakly identified. Our test is designed specifically to accommodate such models.

[1] common examples are the logistic and exponential functions $g(z,\pi) = (1-\exp\{-\pi_1(z-\pi_2)\})^{-1}$ and $g(z,\pi) = 1-\exp\{-\pi_1(z-\pi_2)^2\}$ for $\pi_1 > 0$. See e.g. van Dijk et al. (2002).

[2] $\mathcal{L}_n = o(n)$ is necessary to provide Fischer consistency of the the sample correlation.

Our white noise test, however, is robust to parameter identification failure. Consider estimating scalar parameters (β, π) from the nonlinear model $Y_t = \beta_0 g(Z_t, \pi_0) + \varepsilon_t$ for some non-linear function g. It is well-known that π_0 can be (strongly) identified when $\beta_0 \neq 0$, and when $\beta_0 = 0$, π_0 cannot be identified. In order to accommodate non-identification, we adopt the identification unifying framework of Andrews and Cheng (2012a). This framework is characterized by the notion of drifting sequences of true parameters. Let $\beta \equiv \beta_n$ be a sequence of true parameters that are drifting to 0, the point of identification failure for this example. Andrews and Cheng (2012a) categorize the strength of identification of π_0 by the speed at which $\beta_n \to 0$. If $\beta_n \to 0$ slowly enough, then one can still consistently estimate π_0, and we say that π_0 is semi-strongly identified. However, if $\beta_n \to 0$ too quickly, then one cannot consistently estimate π_0, and we say that π_0 is weakly identified. Table 1 from Andrews and Cheng (2012a) details the rates associated with these categories. It is important to note that in this literature the source of identification failure is known; that is, our model tells us specifically that $\beta = 0$ results in identification failure. More recently, Han and McCloskey (2016) develop theory for the case in which the source of identification failure is unknown. We focus on the former case and leave this extension for future research.

Note that the estimator for β is consistent, regardless of the identification strength of π. However, the estimator for β is a function of the estimator for π, $\hat{\beta}_n \equiv \hat{\beta}_n(\hat{\pi}_n)$, and $\hat{\pi}_n$ converges to a random variable when π_0 is not consistently estimable, yielding a non-standard distribution for $\hat{\beta}_n$.[3] This poses problems for tests based on residuals from model estimation. Non-standard behavior of the estimators propagates through to the test statistic, yielding a non-standard distribution for the test statistic and resulting in potentially distorted inference from traditional tests. The limiting distribution of our test statistic can be categorized by whether π_0 is consistently estimable or not, so we group weak identification and non-identification together and refer to them as weak identification, and we collectively refer to strong and semi-strong identification as strong identification.

This paper is related to but different from the literature on hypothesis testing with a nuisance

[3] See e.g. figure 2 in Andrews and Cheng (2012a).

parameter. Davies (1977, 1987) provide early references for hypothesis tests with nuisance parameters under the null. See also Hansen (1996), Stinchcombe and White (1998), Ghysels and Guay (2004), and more recently Andrews and Mikusheva (2016). Andrews and Ploberger (1994) discuss optimal tests with a nuisance parameter under the null. Andrews and Ploberger (1996) develop a test for white noise against an ARMA(1,1) alternative since these models provide a parsimonious representation of a broad class of stationary time series. As noted by Nankervis and Savin (2010), Poterba and Summers (1988) show that many financial return series can be represented by ARMA(1,1) models. In their model, the ARMA(1,1) reduces to a white noise process under the null, making the MA coefficient a nuisance parameter.

Andrews and Cheng (2012a, 2013, 2014) and Cheng (2015) discuss inference under weak identification but do not consider max test statistics, implementation of a bootstrap, or tests on objects from estimated models, such as white noise tests, that are not tests directly on the model parameters. In contrast, we consider white noise tests based on the maximum of a sequence of correlations that we implement with a bootstrap.

White noise tests have a long history, dating in some form to at least Box and Pierce (1970) and Ljung and Box (1978). In addition to portmanteau tests, spectral tests (Hong, 1996; Shao, 2011a) are also widely considered in the literature. Many early tests for serial correlation are based on i.i.d. Gaussian assumptions and required a finite maximum lag length cutoff. We are specifically interested in true white noise tests, which are able to accommodate asymptotically infinitely many lags, as questions such as the efficient market hypothesis are related to true white noise tests (Hill and Motegi, 2019).

Further, and perhaps more importantly, serial uncorrelatedness is equivalent to independence under Gaussian assumptions, but it does not imply serial independence in general. Many questions in economics and finance such as financial predictability are related to a martingale difference sequence hypothesis, which itself implies serial uncorrelatedness but not serial independence. For example, a GARCH(1,1) process is a martingale difference sequence and is uncorrelated but serially dependent. Romano and Thombs (1996) showed that the traditional Box-Pierce statistic can be misleading under uncorrelated dependent errors. Francq, Roy, and Zakoian (2005) similarly show

that the asymptotic distribution of the correlation coefficients of residuals from ARMA processes do not follow the standard chi-square distribution when the errors are uncorrelated but dependent, and using chi-square critical values in this situation leads to distorted inference.

Often, the martingale difference sequence errors are modeled using GARCH processes, and standardized residuals are used to construct the sample serial correlation even though these tests do not have standard asymptotic distributions. Chen (2008) provides tests for autocorrelation specifically for models with GARCH based errors, but these tests assume that the model is correctly specified. Francq et al. (2005) and Nankervis and Savin (2010, 2012) develop tests that do not rely on a correctly specified model for the conditional variance. Further, Nankervis and Savin (2010, 2012) note that the assumption of martingale difference errors may be too restrictive. As a result, recent interest has focused on uncorrelated dependent time series (Nankervis and Savin, 2010, 2012; Shao, 2011a,b; Zhu and Li, 2015; Zhang, 2016; Hill and Motegi, 2018).

Our test is based on the maximum sample serial correlation, and when testing the maximum value in a sequence, we are most often interested in determining if any of the parameter elements are different from zero. In considering only the maximum from the sequence of values, the max test statistic utilizes the most informative measure available from our data, eliminating issues that arise from low degrees of freedom and inversion of large or near singular covariance matrices when a large number of variables needs to be tested (Hill and Dennis, 2018; Ghysels et al., 2016a), or by combining noisy estimates, which occurs when calculating serial correlations at long lags (Hill and Motegi, 2018).

Statistics based on a maximum of a sequence of values is an extensively studied topic in the literature[4] dating at least to Fisher and Tippet (1928) and Gnedenko (1943). See also Gumbel (1958) and Berman (1964). Typically in this literature, extreme value theory arguments appeal to the Extremal Types Theorem to determine the exact asymptotic distribution of the maximum statistic (de Haan, 1976). For example, Xiao and Wu (2014) provide a test for serial correlation for observed sequences using the maximum sample autocovariance and show that under suitable

[4]See Leadbetter et al. (1983) and Resnick (1987) for textbook treatments.

normalization, the test statistic converges in distribution to a Gumbel (type I extreme value) distribution. These arguments require that when the data are divided into blocks, the dependence between increasingly distant blocks decays at a sufficient rate as with a mixing condition.

Hill and Motegi (2018) and Hill and Dennis (2018) argue that when allowing for general dependence in the data generating process and residuals to be used in the max statistic, the classical extreme value theory arguments are no longer straight forward to prove and may require more stringent assumptions than are needed by other methods. Further, extreme value theoretic arguments for establishing the limiting distribution of the maximum of a sequence of values often relies on Gaussianity of the underlying sequence. Hill and Motegi (2018) and Hill and Dennis (2018) develop theory that does not rely on Gaussianity and that allows the use of the dependent wild bootstrap (Shao, 2010, 2011a) to mimic the finite sample distribution of the max statistic.

The bootstrapped white noise test in Hill and Motegi (2018) is based on the maximum serial correlation and allows for a weaker moment contraction property than that in Xiao and Wu (2014) and side-steps asymptotic extremal value theory arguments by exploiting convergence of $\{\sqrt{n}(\hat{\gamma}(h) - \gamma(h)) : \ 1 \leq h \leq \mathcal{L}\}$ to a Gaussian process for each $\mathcal{L} \in \mathbb{N}$ paired with arguments dating to Ramsey (1929). This method requires weaker conditions than the extreme value theoretic approach but results in the trade-off that an upper bound on the sequence $\mathcal{L}_n \to \infty$ cannot be provided.[5] Further, Hill and Motegi (2018) ignore the possibility of nuisance parameters and only allow for strong identification of all parameters in the model estimation step.

Methods for bootstrapping high dimensional statistics have not been available until recently. Chernozhukov et al. (2013, 2017) develop a theory that is able to both bypass the typical extreme value theoretic asymptotic arguments and deliver an impressive growth rate for the sequence being examined. However, they require independence, and their theory is only appropriate for observed random variables and relies on Gaussian approximation that is not appropriate for approximations

[5]Hill and Motegi (2018) address the issue of optimal lag selection with a data driven procedure, modified from the method of Escanciano and Lobato (2009). This procedure could be applied to the testing framework presented here; however, this is beyond the scope of this paper, as we seek to illustrate the effect of weak identification on the test.

of non-Gaussian normalized summands.[6] Zhang and Cheng (2018) extend the Gaussian approximation theory in Chernozhukov et al. (2013, 2017) to allow for dependence, but only allow for observed random variables. Zhang and Wu (2017) develop theory for a Gaussian approximation for high dimensional times series but only allow for observed sequences as well. The theory in Hill and Dennis (2018) and Hill and Motegi (2018) is also able to bypass extreme value theoretic arguments, allows for dependence under the null, and is appropriate for residuals. For this reason, we rely on the theory developed in Hill and Motegi (2018) and Hill and Dennis (2018).

For model estimation, we adopt the notation of Andrews and Cheng (2012a). Section 2.2 discusses the preliminary notation and assumptions needed to fit within their framework. Section 2.3 presents the main assumptions and results, and we present the bootstrap and prove its validity in section 2.4. Section 2.5 presents the Monte-Carlo simulations. All proofs and supporting lemmas are collected in the appendix.

2.2 Preliminary Notation and Assumptions

The true parameter is $\gamma = (\theta, \phi)$ with compact true parameter space

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

where $\theta = (\beta, \zeta, \pi) = (\psi, \pi)$, $\psi = (\beta, \zeta)$, and we assume ψ is always identified and ζ does not effect the identification of π, and ϕ is an additional parameter such that $\gamma = (\theta, \phi)$ completely determines the distribution of the data. For some $\gamma \in \Gamma$, expectation under the true distribution of $\{(Y_t, X_t, \varepsilon_t)\} = \{W_t : t \leq n\}$ is denoted E_γ.

Since the estimator $\hat{\pi}_n$ for π_n is inconsistent, we make use of the following concentrated criterion function $Q_n^c(\pi)$ and estimator $\hat{\psi}_n(\pi)$. Define $\hat{\psi}_n(\pi) \in \Psi(\pi)$ for a given $\pi \in \Pi$ by

$$Q_n(\hat{\psi}_n(\pi), \pi) = \inf_{\psi \in \Psi(\pi)} Q_n(\psi, \pi) + o_p(n^{-1})$$

[6] See also Belloni et al. (2018).

and define $\hat{\pi}_n \in \Pi$ by

$$Q_n^c(\hat{\pi}_n) = Q_n(\hat{\psi}_n(\hat{\pi}_n), \hat{\pi}_n) = \inf_{\pi \in \Pi} Q_n(\hat{\psi}_n(\pi), \pi) + o_p(n^{-1}).$$

Observe $(\hat{\psi}_n(\hat{\pi}_n), \hat{\pi}_n) = \hat{\theta}_n = \inf_{\theta \in \Theta} Q_n(\theta) + o_p(n^{-1})$.

We adopt the notation of Andrews and Cheng (2012a) in order to define cases that differentiate weak and (semi-)strong identification. The theory relies on the following drifting sequences of true parameters. Define the set of true drifting sequences as $\Gamma_0 = \{\{\gamma_n \in \Gamma : n \geq 1\} : \gamma_n \to \gamma_0 \in \Gamma\}$, and define the drifting cases:

(i) $\Gamma(\gamma_0, 0, b) = \{\{\gamma_n\} \in \Gamma_0 : \beta_0 = 0,\ n^{1/2}\beta_n \to b \in (\mathbb{R} \cup \{\pm\infty\})^{d_\beta}\}$

(ii) $\Gamma(\gamma_0, \infty, \omega_0) = \{\{\gamma_n\} \in \Gamma_0 : n^{1/2}\beta_n \to \infty,\ \beta_n/||\beta_n|| \to \omega_0,\ ||\omega_0|| = 1\}$.

In our model, the identification of π is based on whether or not the parameter $\beta = 0$. In terms of these drifting sequences, π_0 is not identified asymptotically when the limiting parameter $\beta_0 = 0$. Further, in the case that $\beta_0 = 0$, the speed at which $\beta_n \to \beta_0 = 0$ affects the asymptotic analysis. In particular, when $\beta_n \to 0$ fast enough, given by case (i) with $||b|| < \infty$, we say the parameter π_0 is *weakly* identified. In this case, the estimator $\hat{\pi}_n$ is not consistent. Case two gives the definitions of *semi-strong* identification, when $\beta_0 = 0$ and *strong* identification, when $\beta_0 \neq 0$.

In the (semi-)strong identification cases $\hat{\pi}_n$ is consistent, and we employ first order expansions around the true parameter θ_n. However, since $\hat{\pi}_n$ is not consistent under weak identification, an expansion around $\theta_n = (\psi_n, \pi_n)$ is not appropriate. Inspired by the expansion of the criterion function about the point of non-identification in Andrews and Cheng (2012a), we expand our test statistic about the point of non-identification in the weak identification case in order to deal with the inconsistency of $\hat{\pi}_n$. Recall the point of non-identification is $\beta_0 = 0$. Define $\psi_{0,n} = (0, \zeta_n)$ and $Q_{0,n} = Q_n(\psi_{0,n}, \pi)$.

Define

$$\xi(\pi; \gamma_0, b) = -\frac{1}{2}(G(\pi) + K(\pi, \pi_0)b)' H^{-1}(\pi)(G(\pi) + K(\pi, \pi_0)b)$$

where G is a mean zero Gaussian process, H is a Hessian, and K arises as a bias correction. Assume $\pi^*(\gamma_0, b) = \underset{\pi \in \Pi}{\text{argmin}}\ \xi(\pi; \gamma_0, b)$.

More specifically, under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$, the mean zero Gaussian process $\{G(\pi; \gamma_0) : \pi \in \Pi\}$ is defined as the limit of the process $\{G_n(\pi; \gamma_0) : \pi \in \Pi\}$ defined by

$$\begin{aligned} G_n(\psi_{0,n}, \pi) &= n^{1/2}\left\{ \frac{\partial}{\partial \psi} Q_n(\psi_{0,n}, \pi) - E_{\gamma_n} \frac{\partial}{\partial \psi} Q_n(\psi_{0,n}, \pi) \right\} \\ &= n^{-1/2} \sum_{t=1}^{n} \Big(m^{\psi}(W_t, \psi_{0,n}, \pi) - E_{\gamma_n} m^{\psi}(W_t, \psi_{0,n}, \pi) \Big) \end{aligned}$$

where $\frac{\partial}{\partial \psi} Q_n(\theta) = n^{-1} \sum_{i=1}^{n} m^{\psi}(W_t, \theta)$. $H(\pi; \gamma_0)$ is the nonstochastic symmetric $d_\psi \times d_\psi$ matrix valued function, continuous on Π that is the uniform (in π) limit of $H_n(\psi, \pi; \gamma_0) = \frac{\partial}{\partial \psi} \frac{\partial}{\partial \psi'} Q_n(\psi, \pi)$. Finally, $K_n(\theta; \gamma_0) = n^{-1} \sum_{t=1}^{n} \frac{\partial}{\partial \beta_0} E_{\gamma_0} m^{\psi}(W_t, \theta)$.

Assumption 1 (Weak Identification Objects)**.** *Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$,*

(i) $G_n(\cdot) \Rightarrow G(\cdot; \gamma_0)$*, where $G(\cdot; \gamma_0)$ is a mean zero Gaussian process indexed by $\pi \in \Pi$ with bounded continuous sample paths and a.s. p.d. covariance kernel* $\Omega(\pi, \tilde{\pi}; \gamma_0) \equiv E[G(\pi; \gamma_0) G(\tilde{\pi}; \gamma_0)']$ *for $\pi, \tilde{\pi} \in \Pi$.*

(ii) $\sup_{\pi \in \Pi} ||H_n(\psi_{0,n}, \pi) - H(\pi; \gamma_0)|| \xrightarrow{p} 0$ *for some nonstochastic symmetric $d_\psi \times d_\psi$ matrix-valued function $H(\pi; \gamma_0)$ on $\Pi \times \Gamma$ that is continuous on Π for all $\gamma_0 \in \Gamma$ and $\lambda_{min}(H(\pi; \gamma_0)) > 0$ and $\lambda_{min}(H(\pi; \gamma_0)) < \infty\ \forall \pi \in \Pi$ for all $\gamma_0 \in \Gamma$ with $\beta_0 = 0$.*

(iii) $K_n(\theta; \gamma)$ *exists for all $(\theta, \gamma) \in \Theta_\delta \times \Gamma_0$, $\forall n \geq 1$ and for some nonstochastic $d_\psi \times d_\beta$ matrix-valued function $K(\psi_0, \pi; \gamma_0)$ that is continuous on Π for all $\gamma_0 \in \Gamma$ with $\beta_0 = 0$, $K_n(\tilde{\psi}_n, \pi; \tilde{\gamma}_n) \to K(\psi_0, \pi; \gamma_0)$ uniformly over $\pi \in \Pi$ for all nonstochastic sequences $\{\tilde{\psi}_n\}$ and $\{\tilde{\gamma}_n\}$ such that $\tilde{\gamma}_n \to \gamma_0$ and $\tilde{\psi}_n \to \psi_0 = (0, \zeta_0)$.*

(iv) each sample path of the stochastic process $\{\xi(\pi; \gamma_0, b) : \pi \in \Pi\}$ is some set $\mathcal{A}(\gamma_0, b)$ with $P_{\gamma_0}(\mathcal{A}(\gamma_0, b)) = 1$ is minimized over Π at a unique point, denoted $\pi^*(\gamma_0, b)\ \forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$ and for all b with $||b < \infty$.

These assumptions are Assumptions C3, C4, and C5 Andrews and Cheng (2012a) which we borrow in order to retain generality. The objects G_n, H_n, and K_n are the objects that appear in our test statistic.

Example 1 (STAR(1) Model)**.** *Consider the model* $\varepsilon_t(\theta) = y_t - \beta x_t g(z_t, \pi) - \xi x_t$ *with true parameter* θ_n *so that* $\varepsilon_t(\theta_n) = \varepsilon_t$*. We estimate the model with least squares, so we have* $Q_n(\theta) = \frac{1}{2}\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t(\theta)^2$*. Define* $d_{\psi,t}(\pi) = \frac{\partial}{\partial\psi}\varepsilon_t(\psi,\pi) = -[x_t g(z_t,\pi), x_t]'$*. Then*

$$\hat{H}_n(\pi) = \frac{1}{n}\sum_{t=1}^{n} d_{\psi,t}(\pi)d_{\psi,t}(\pi)'$$
$$\hat{K}_n(\pi;\gamma_0) = -\frac{1}{n}\sum_{t=1}^{n} d_{\psi,t}(\pi)x_t g(z_t,\pi_0)$$

and

$$\begin{aligned} G_n(\pi) &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\Big\{\varepsilon_t d_{\psi,t}(\pi) - E_{\gamma_n}[\varepsilon_t d_{\psi,t}(\pi)]\Big\} \\ &\qquad - b'\frac{1}{n}\sum_{t=1}^{n}\Big\{x_t g(z_t,\pi_n)d_{\psi,t}(\pi) - E_{\gamma_n}[x_t g(z_t,\pi_n)d_{\psi,t}(\pi)]\Big\} \\ &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\Big\{\varepsilon_t d_{\psi,t}(\pi) - E_{\gamma_n}[\varepsilon_t d_{\psi,t}(\pi)]\Big\} + o_{p\pi}(1). \end{aligned}$$

Then $E(\varepsilon_t|x_t) = 0$ *a.s. and* $E(\varepsilon_t^2|x_t) = \sigma^2 \in (0,\infty)$ *a.s. under* H_0 *implies the covariance kernel for* $G(\pi)$ *is* $E[e_t^2 d_{\psi,t}(\pi)d_{\psi,t}(\tilde{\pi})'] = \sigma^2 H(\pi,\tilde{\pi})$*. Further, this implies that* $H^{-1/2}(\pi)G(\pi) \sim N(0,\sigma^2)$ *with covariance kernel* $\sigma^2 H^{-1/2}(\pi)H(\pi,\tilde{\pi})H^{-1/2}(\pi)$*.*

Example 2 (ARMA(1,1) Model)**.** *Consider the model* $y_t = (\beta_n + \pi_n)y_{t-1} + \varepsilon_t - \pi_n\varepsilon_{t-1}$*. This model is estimated by maximum likelihood, the limits are described by the following quantities.*

$$H_n(\pi) = \frac{1}{n}\sum_{t=1}^{n}\begin{pmatrix} \zeta_n^{-1}\Big(\sum_{j=0}^{\infty}\pi^j y_{t-j-1}\Big)^2 & \zeta_n^{-2}y_t\sum_{k=0}^{\infty}\pi^k y_{t-k-1} \\ \zeta_n^{-2}y_t\sum_{k=0}^{\infty}\pi^k y_{t-k-1} & -(1/2)\zeta_n^{-2} + \zeta_n^{-3}y_t^2 \end{pmatrix}$$

with limit

$$H(\pi;\gamma_0) = \begin{pmatrix} (1-\pi^2)^{-1} & 0 \\ 0 & (2\zeta_0^2)^{-1} \end{pmatrix}.$$

$K_n(\theta;\gamma_0)$ is complicated (see Andrews and Cheng (2012b), section C) and has limit $K(\pi;\gamma_0) = \begin{pmatrix} -(1-\pi_0\pi)^{-1} \\ 0 \end{pmatrix}$.

$$G_n(\pi) = n^{-1/2}\sum_{t=1}^{n} \begin{pmatrix} -\zeta_n^{-1} y_t \sum_{k=0}^{\infty} \pi^k y_{t-k-1} \\ -(1/2)\zeta_n^{-2}(y_t^2 - \zeta_n) \end{pmatrix} - \begin{pmatrix} -E_{\gamma_n}\zeta_n^{-1} y_t \sum_{k=0}^{\infty} \pi^k y_{t-k-1} \\ -E_{\gamma_n}(1/2)\zeta_n^{-2}(y_t^2 - \zeta_n) \end{pmatrix}$$

has the limit

$$G(\pi;\gamma_0) = \begin{pmatrix} \sum_{j=0}^{\infty} \pi^j Z_j \\ (1/2)\zeta^{-2}(E_{\gamma_0}(\varepsilon_t^2 - \zeta_0)^2)^{1/2} Z \end{pmatrix}$$

where $Z, Z_0, Z_1, \ldots$ are independent standard normal random variables. The covariance kernel of $G(\pi;\gamma_0)$ is $\begin{pmatrix} (1-\pi\tilde{\pi})^{-1} & 0 \\ 0 & (1/4)\zeta_0^{-4} E_{\gamma_0}(\varepsilon_t^2 - \zeta_0)^2 \end{pmatrix}$.

Finally, define the Gaussian process

$$\tau(\pi;\gamma_0,b) = -H^{-1}(\pi;\gamma_0)(G(\pi;\gamma_0) + K(\pi;\gamma_0)b) - (b,0).$$

We require additional objects for the case in which π_0 is (semi-)strongly identified.

Let $B(\beta) = \begin{pmatrix} I_{d_\psi} & 0_{d_\psi \times d_\pi} \\ 0_{d_\psi \times d_\pi} & ||\beta|| \cdot I_{d_\pi} \end{pmatrix}$.

Assumption 2 (Strong Identification Objects)**.** *Under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,*

(i) $G_n^\theta(\theta_n) = n^{1/2} B^{-1}(\beta_n)\frac{\partial}{\partial\theta} Q_n(\theta_n) \xrightarrow{d} G^\theta(\gamma_0) \sim N(0, V(\gamma_0))$ for some symmetric $d_\theta \times d_\theta$ matrix $V(\gamma_0)$ which is positive definite $\forall \gamma_0 \in \Gamma$.

(ii) $J_n(\theta_n) \equiv B^{-1}(\beta_n)\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta'}Q_n(\theta_n)B^{-1}(\beta_n) \xrightarrow{p} J(\gamma_0)$*, where* $J(\gamma_0)$ *is a* $d_\theta \times d_\theta$ *nonsingular and symmetric matrix.*

The previous assumption is Assumptions D2 and D3 from Andrews and Cheng (2012a), which detail the objects that appear in our expansions under the semi-strong and strong identification cases. The scaling matrix $B(\beta_n)$ is needed in order to eliminate singularity of the second derivative matrix when $\beta_n \to 0$.

We further assume that $\frac{\partial}{\partial\theta}Q_n(\theta) = n^{-1}\sum_{i=1}^{n} m^\theta(W_t, \theta)$, which also implies that $m^\psi(W_t, \theta) = \mathcal{S}_\psi m^\theta(W_t, \theta)$ for the $d_\psi \times d_\theta$ selection matrix $\mathcal{S}_\psi$ that selects the first ψ elements from the $d_\theta \times 1$ vector $m_t^\theta(\theta) \equiv m^\theta(W_t, \theta)$.

2.3 Assumptions and Main Results

2.3.1 Assumptions

Recall that $\varepsilon_t(\hat{\theta}_n)$ is our model for the regression error (e.g. $\varepsilon_t(\hat{\theta}_n) = Y_t - \hat{\beta}_n X_t \times g(X_t, \hat{\pi}_n) - \hat{\zeta}_n X_t$ in the nonlinear regression model), so under a correctly specified model with true parameter θ_n, we have $\varepsilon_t \equiv \varepsilon_t(\theta_n)$.

Assumption 3 (A)**.** *If* $\beta = 0$*, then* $\varepsilon_t(\theta)$ *does not depend on* π *for all* $\theta = (\beta, \zeta, \pi) = (0, \zeta, \pi) \in \Theta$ *for any true parameter* $\gamma^* \in \Gamma$*. Moreover,* $Q_n(\theta)$ *only depends on* π *through* $\varepsilon_t(\theta)$*.*

Remark 1. *Assumption 3 is similar to and indeed related to Assumption A in Andrews and Cheng (2012a). This restricts our attention to models in which the source of identification failure is known. Han and McCloskey (2016) extend the framework of Andrews and Cheng (2012a) to allow for cases in which the source of identification failure is not known; however, we do not allow for unknown sources of identification failure in our present white noise residual test.*

Our primary concern is in testing if $\{\varepsilon_t\}$ is a white noise process:

$$H_0 : \rho(h) = 0 \; \forall h \in \mathbb{N} \qquad \text{vs.} \qquad H_A : \rho(h) \neq 0 \text{ for some } h \in \mathbb{N}$$

Our test statistic is the sample max correlation statistic

$$\hat{\mathcal{T}}_n = \max_{1 \leq h \leq \mathcal{L}_n} \sqrt{n} |\hat{\rho}_n(h)|$$

where $\hat{\rho}_n(h) = E(\varepsilon_t \varepsilon_{t-h})/E(\varepsilon_t^2)$ and $\mathcal{L}_n$ is a sequence of integers with $\mathcal{L}_n \to \infty$ as $n \to \infty$ and $\mathcal{L}_n = o(n)$ to allow for a true white noise test.

We begin with assumptions on the estimator $\hat{\theta}_n$ that are standard results under weak identification (see e.g. Andrews and Cheng (2012a)). This allows us to maintain a great deal of generality with respect to the model that we are investigating. Define $\tau_n(\pi; \gamma_0, b) = -H_n^{-1}(\pi; \gamma_0)(G_n(\pi; \gamma_0) + K_n(\pi; \gamma_0)b) - (b, 0)$, and recall that

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

and

$$G_n^{\theta}(\theta_n) = n^{-1/2} B^{-1}(\beta_n) \sum_{i=1}^{n} m_t^{\theta}(\theta_n).$$

Assumption 4 (m). *(i) Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$, $m_t^{\psi}(\pi) \equiv m_t(\psi_{0,n}, \pi)$ is stationary, ergodic, $L_{p/2}$-bounded for some $p > 4$, and L_2-NED with size $-1/2$ on an α-mixing base $\{\nu_t\}$ with coefficients $\alpha_h^{\nu} = O(h^{-p/(p-4)-\iota})$ for tiny $\iota > 0$ for every $\pi \in \Pi$.*

(ii) Under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $m_t^{\theta} \equiv m_t(\theta_n)$ is mean zero, stationary, ergodic, $L_{p/2}$-bounded for some $p > 4$, and L_2-NED with size $-1/2$ on an α-mixing base $\{\nu_t\}$ with coefficients $\alpha_h^{\nu} = O(h^{-p/(p-4)-\iota})$ for tiny $\iota > 0$.

(iii) $m_t(\theta)$ is two times continuously differentiable and $E[\sup_{\theta \in \Theta} ||(\frac{\partial}{\partial \theta})^j m_t(\theta)||^2] < \infty$ for $j = 0, 1, 2$.

Remark 2. *Assumption 4 is a sufficient condition for Assumption 1(a) and 2(a). Smoothness (iii) ensures a stochastic equicontinuity property for a functional central limit theorem (see e.g. Andrews (1994)).*

Assumption 5 (Weak Id Estimator Limit). *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $||b|| < \infty$,

(i) $\sup_{\pi \in \Pi} ||\hat{\psi}_n(\pi) - \psi_n|| \xrightarrow{p} 0$

(ii) $\sup_{\pi \in \Pi} ||n^{1/2}(\hat{\psi}_n(\pi) - \psi_{0,n}) + H_n^{-1}(\psi_{0,n}, \pi)\frac{1}{\sqrt{n}} \sum_{t=1}^{n} m_t^{\psi}(\psi_{0,n}, \pi)|| \xrightarrow{p} 0$

Assumption 6 (Strong Id Estimator Limit). *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,

(i) $||\hat{\theta}_n - \theta_n|| \xrightarrow{p} 0$

(ii) $n^{1/2}B(\beta_n)(\hat{\theta}_n - \theta_n) = -J_n^{-1}(\gamma_0)n^{-1/2}B^{-1}(\beta_n) \sum_{i=1}^{n} m_t^{\theta}(\theta_n) + o_p(1)$

Following Hill and Motegi (2018), our test applies to near-epoch-dependent random variables.

Assumption 7 (W). *(i)* $\{x_t, y_t\}$ *are stationary, ergodic, and* $L_{2+\delta}$*-bounded for some* $\delta > 0$. *Denote by* $\mathcal{F}_t$ *the* σ*-field generated by* $\{x_t, y_t\}$.

(ii) ε_t *has* $E(\varepsilon_t) = 0$*, is stationary, ergodic,* L_p*-bounded for some* $p > 4$*, and* L_4*-NED with size* $-1/2$ *on an* α*-mixing base* $\{\nu_t\}$ *with coefficients* $\alpha_h^{\nu} = O(h^{-p/(p-2)-\iota})$ *for tiny* $\iota > 0$.

In order to establish the limiting distribution of our test statistic, we require some additional assumptions on the function $\varepsilon_t(\theta)$.

Assumption 8 (R0). *(i)* $\varepsilon_t(\theta)$ *is* $\mathcal{F}_t$*-measurable for each* θ *and three times continuously differentiable a.s. on an open convex set containing* Θ^*.

(ii) Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $||b|| < \infty$, $E[\sup_{\pi \in \Pi} \sup_{\psi \in \mathcal{N}_{\psi_0}} |(\frac{\partial}{\partial \psi})^j \varepsilon_t(\psi, \pi)|^4] < \infty$ *for* $j = 0, 1, 2, 3$ *and a compact set* $\mathcal{N}_{\psi_0}$ *containing* ψ_0.

(iii) Under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $E[\sup_{\theta \in \mathcal{N}_{\theta_0}} |(\frac{\partial}{\partial \theta})^j \varepsilon_t(\theta)|^4] < \infty$ *for* $j = 0, 1, 2, 3$ *and a compact set* $\mathcal{N}_{\theta_0}$ *containing* θ_0.

The following two assumptions are technical conditions that are necessary to establish uniform LLNs for the derivatives that appear in the mean value expansion of the covariance under weak and strong identification, respectively. We require more conditions under weak identification due to the inconsistency of the estimator $\hat{\pi}_n$. In particular, the expansion in the weak identification case

is about $\psi_{0,n}$ rather than the true parameter θ_n, leading to the need to add and subtract $\varepsilon_t\varepsilon_{t-h}$ in the proof, hence the need for Assumption 9(v) which ensures the associated bias term is bounded.

Assumption 9 (Rw)**.** *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $||b|| < \infty$*,*

(i) the non-stochastic function $\mathcal{D}_n(h,\pi) = \mathcal{D}_n(h,\psi_{0,n},\pi) \equiv E_{\gamma_n}[\frac{\partial}{\partial\psi}(\varepsilon_t(\psi,\pi)\varepsilon_{t-h}(\psi,\pi))]\Big|_{\psi=\psi_{0,n}}$ *exists and is differentiable a.s. on an open, convex set* Π *containing the true parameter space* Π^**.*

(ii) $\sup_{\pi\in\Pi} ||\frac{\partial}{\partial\pi}\left(\frac{1}{n}\sum_{t=1+h}^{n}\frac{\partial}{\partial\psi}[\varepsilon_t(\psi,\pi)\varepsilon_{t-h}(\psi,\pi)]\Big|_{\psi=\psi_{0,n}} - \mathcal{D}_n(h,\psi_{0,n},\pi)\right)|| = O_p(1).$

(iii) The non-stochastic function $\tilde{\mathcal{D}}_n(h,\psi,\pi) = E_{\gamma_n}\left[\frac{\partial}{\partial\psi}\frac{\partial}{\partial\psi'}\left(\varepsilon_t(\psi,\pi)\varepsilon_{t-h}(\psi,\pi)\right)\right]$ *is continuous at* $\psi_{0,n}$ *and is differentiable a.s. on an open, convex set* Θ_0 *containing the true parameter space* Θ^**.*

(iv) $\sup_{\pi\in\Pi}\sup_{\psi\in\Psi(\pi)} ||\frac{\partial}{\partial\theta}\mathcal{Z}_n(h,\psi,\pi)||$
$= \sup_{\pi\in\Pi}\sup_{\psi\in\Psi(\pi)} ||\frac{\partial}{\partial\theta}\left(\frac{1}{n}\sum_{t=1+h}^{n}\frac{\partial}{\partial\psi}\frac{\partial}{\partial\psi'}[\varepsilon_t(\psi,\pi)\varepsilon_{t-h}(\psi,\pi)] - \tilde{\mathcal{D}}_n(h,\psi_{0,n},\pi)\right)|| = O_p(1).$

(v) $E_{\gamma_n}\left[\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi) - \varepsilon_t\varepsilon_{t-h}\right] = O(1/\sqrt{n})$

Assumption 10 (Rs)**.** *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$*,*

(i) the non-stochastic function $\mathcal{D}_n^\theta(h) = \mathcal{D}_n^\theta(h,\theta_n) \equiv E_{\gamma_n}[\frac{\partial}{\partial\theta}(\varepsilon_t(\theta)\varepsilon_{t-h}(\theta))]\Big|_{\theta=\theta_n}$ *exists.*

(ii) The non-stochastic function $\tilde{\mathcal{D}}_n^\theta(h,\theta) = E_{\gamma_n}\left[\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta'}\left(\varepsilon_t(\theta)\varepsilon_{t-h}(\theta)\right)\right]$ *is continuous at* θ_n *and is differentiable a.s. on an open, convex set* Θ_0 *containing the true parameter space* Θ^**.*

(iii) $\sup_{\theta\in\Theta} ||\frac{\partial}{\partial\theta}\mathcal{Z}_n^\theta(h,\theta)|| = \sup_{\theta\in\Theta} ||\frac{\partial}{\partial\theta}\left(\frac{1}{n}\sum_{t=1+h}^{n}\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta'}[\varepsilon_t(\theta)\varepsilon_{t-h}(\theta)] - \tilde{\mathcal{D}}_n(h,\theta_n)\right)|| = O_p(1).$

Remark 3. *Assumption 9(i) implies that* $\varepsilon_t(\psi,\pi)\varepsilon_{t-h}(\psi,\pi)$ *is stationary and ergodic, and Assumptions 9(i) and (ii) imply* $\frac{\partial}{\partial\psi}[\varepsilon_t(\psi,\pi)\varepsilon_{t-h}(\psi,\pi)]$ *is stationary and ergodic since the derivative is a measurable transformation. Assumption 9(iv) is a technical condition that is necessary in order to establish stochastic equicontinuity for a uniform law of large numbers (e.g. Newey (1991)). Considering the least squares case with* $h = 0$ *would indicate that* $\frac{\partial}{\partial\psi'}\frac{\partial}{\partial\psi}[\varepsilon_t(\psi,\pi)^2]$ *is the Hessian*

of the objective function, so the assumption does not appear to be very restrictive. In many example applications, these conditions hold as a result of the conditions needed to establish the asymptotic results for the estimators (e.g. Assumption 5).

Remark 4. *Assumption 9(v) must be verified for the chosen model $\varepsilon_t(\theta)$. It is often easy to verify in specific models. Further, recall that $\varepsilon_t(\psi_{0,n}, \pi)$ does not depend on π under Assumption 3; however, $\varepsilon_t(\psi_{0,n}, \pi)$ does depend on the true parameter π_n. Thus, we only require the quantity to be $O(1/\sqrt{n})$ and do not require uniformity over Π. For example, consider the two example models (a) STAR(1) and (b) ARMA(1,1).*

Example 3 (Scalar Non-linear Regression Model)**.** *The Scalar Non-linear Regression Model is $\varepsilon_t(\theta) = y_t - \beta x_t g(x_t, \pi) - \xi x_t$. Then $\varepsilon_t(\theta_n) = y_t - \beta_n x_t g(x_t, \pi_n) - \xi_n x_t = \varepsilon_t$ and $\varepsilon_t(\psi_{0,n}, \pi) = y_t - \xi_n x_t = \beta_n x_t g(x_t, \pi) + \varepsilon_t$. To verify Assumption 9(v), use $\sqrt{n}\beta_n \to b$, stationarity, ergodicity, moment bound assumptions and the construction of the model to see that under H_0, $\sqrt{n}\Big[E_{\gamma_n}(\varepsilon_t(\psi_{0,n}, \pi)\varepsilon_{t-h}(\psi_{0,n}, \pi)) - E(\varepsilon_t\varepsilon_{t-h})\Big] = b\zeta^{h-1}E[\varepsilon_t^2 g(x_t, \pi_0)] + o_p(1)$.*

Example 4 (ARMA(1,1))**.** *The ARMA(1,1) model, $y_t = (\beta + \pi)y_{t-1} + \varepsilon_t - \pi\varepsilon_{t-1}$, can be written $\varepsilon_t(\theta) = y_t - \beta \sum_{j=1}^{\infty} \pi^{j-1} y_{t-j}$. Then $\varepsilon_t(\beta_n, \pi_n) = y_t - \beta_n \sum_{j=1}^{\infty} \pi_n^{j-1} y_{t-j} \equiv \varepsilon_t$ and $\varepsilon_t(0, \pi_n) = y_t = \beta_n \sum_{j=1}^{\infty} \pi_n^{j-1} y_{t-j} + \varepsilon_t$. To verify Assumption 9(vi), we can show that under H_0,*

$$\sqrt{n}\Big[E_{\gamma_n}(\varepsilon_t(\psi_{0,n}, \pi)\varepsilon_{t-h}(\psi_{0,n}, \pi)) - E(\varepsilon_t\varepsilon_{t-h})\Big] = b \sum_{j=1}^{h} \pi_n^{j-1} E_{\gamma_n}[y_{t-j}\varepsilon_{t-h}] + o_p(1).$$

2.3.2 Main Results

Due to the inconsistency of $\hat{\pi}_n$ under weak identification, we must consider the two cases (i) $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$, which we colloquially refer to as weak identification, and (ii) $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, which we refer to as strong identification, in the analysis of our test statistic. We operate on a first order expansion of our test statistic that differs depending on the identification case, so we refer to the approximations $r_t^{\theta}(h)$ and $r_t^{\psi}(h, \pi)$, defined under strong and weak identification respectively:

$$r_t^{\theta}(h) = \frac{\varepsilon_t\varepsilon_{t-h} - E[\varepsilon_t\varepsilon_{t-h}] - \mathcal{D}^{\theta}(h)' J^{-1}(\gamma_0) m_t^{\theta}}{E[\varepsilon_t^2]}$$

$$r_t^{\psi,n}(h,\pi) = \frac{\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi) - E[\varepsilon_t\varepsilon_{t-h}] - \mathcal{D}(h,\pi)'H^{-1}(\pi;\gamma_0)m_t^{\psi}(\psi_{0,n},\pi)}{E[\varepsilon_t^2]}$$

Define under strong and weak identification, respectively, $z_t^{\theta}(h) = r_t^{\theta}(h) - \rho(h)r_t^{\theta}(0)$ and $z_t^{\psi,n}(h,\pi) = r_t^{\psi,n}(h,\pi) - \rho(h)r_t^{\psi,n}(0,\pi)$.

Define $\mathcal{Z}_n^{\theta}(h) = \frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t^{\theta}(h)$ and $\mathcal{Z}_n^{\psi}(h,\pi) = \frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t^{\psi,n}(h,\pi)$.

Assumption 11. *Let $\mathcal{L}, K \in \mathbb{N}$, and let $\lambda = [\lambda_h]_{h=1}^{\mathcal{L}} \in \mathbb{R}^{\mathcal{L}}$ and $a \in \mathbb{R}^K$. Then*

(i) Take $\{\pi_1,\ldots,\pi_K\} \in \Pi^{\otimes K}$. Under $\{\gamma_n\} \in \Gamma(\gamma_0,0,b)$ with $||b|| < \infty$,

$$\liminf_{n\to\infty} \inf_{\lambda'\lambda=1} \inf_{a'a=1} \inf_{\pi\in\Pi} E\Big[\Big(\sum_{h=1}^{\mathcal{L}}\sum_{k=1}^{K} \lambda_h a_k \mathcal{Z}_n^{\psi}(h,\pi_k)\Big)^2\Big] > 0,$$

and

(ii) under $\{\gamma_n\} \in \Gamma(\gamma_0,\infty,\omega_0)$, $\liminf_{n\to\infty} \inf_{\lambda'\lambda=1} E[(\sum_{h=1}^{\mathcal{L}} \lambda_h \mathcal{Z}_n^{\theta}(h))^2] > 0$.

Remark 5. *Non-degenerate asymptotic variance in a standard assumption in the literature. See e.g. Hill and Motegi (2018).*

The following Lemma provides the approximations that are used to bootstrap the test statistic.

Lemma 2.3.1. *Let Assumptions 3 - 11 hold. For some non-unique sequence of positive integers $\{\mathcal{L}_n\}$ with $\mathcal{L}_n \to \infty$ and $\mathcal{L}_n = o(n)$,*

(a) under $\{\gamma_n\} \in \Gamma(\gamma_0,0,b)$ with $||b|| < \infty$,

$$\Big|\max_{1\le h\le\mathcal{L}_n} \sup_{\pi\in\Pi}(\sqrt{n}|\hat{\rho}_n(h;\pi) - \rho(h)|) - \max_{1\le h\le\mathcal{L}_n} \sup_{\pi\in\Pi}(|\mathcal{Z}_n^{\psi}(h,\pi)|)\Big|$$
$$\le \max_{1\le h\le\mathcal{L}_n} \sup_{\pi\in\Pi}(|\sqrt{n}(\hat{\rho}_n(h;\pi) - \rho(h)) - \mathcal{Z}_n^{\psi}(h,\pi)|) \xrightarrow{p} 0.$$

(b) under $\{\gamma_n\} \in \Gamma(\gamma_0,\infty,\omega_0)$,

$$\Big|\max_{1\le h\le\mathcal{L}_n} (\sqrt{n}|\hat{\rho}_n(h) - \rho(h)|) - \max_{1\le h\le\mathcal{L}_n} (|\mathcal{Z}_n^{\theta}(h)|)\Big| \le \max_{1\le h\le\mathcal{L}_n} (|\sqrt{n}(\hat{\rho}_n(h) - \rho(h)) - \mathcal{Z}_n^{\theta}(h)|) \xrightarrow{p} 0.$$

The limiting distribution of the test statistic under strong identification is established in a similar fashion to Hill and Motegi (2018); however, illuminating the limiting distribution under weak identification require that we decompose $r_t^{\psi,n}(h,\pi)$ by adding and subtracting $\varepsilon_t\varepsilon_{t-h}$ and $\mathcal{D}(h,\pi)'H^{-1}(\pi;\gamma_0)E_{\gamma_n}[m_t^{\psi}(\psi_{0,n},\pi)]$, and then performing a mean value expansion on $E_{\gamma_n}[m_t^{\psi}(\psi_{0,n},\pi)]$ about $\gamma_{0,n}$.[7] This yields the following quantities:

$$\begin{aligned} r_t^{\psi,n}(h,\pi) &= \frac{\varepsilon_t\varepsilon_{t-h} - E[\varepsilon_t\varepsilon_{t-h}]}{E[\varepsilon_t^2]} - \frac{\mathcal{D}(h,\pi)'H^{-1}(\pi;\gamma_0)\big(m_t^{\psi}(\psi_{0,n},\pi) - E_{\gamma_n}[m_t^{\psi}(\psi_{0,n},\pi)]\big)}{E[\varepsilon_t^2]} \\ &\quad + \frac{\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi) - \varepsilon_t\varepsilon_{t-h}}{E[\varepsilon_t^2]} \\ &\quad - \frac{\mathcal{D}(h,\pi)'H^{-1}(\pi;\gamma_0)\big(\beta_n\frac{\partial}{\partial\tilde{\beta}}E_{\tilde{\gamma}_n}[m_t^{\psi}(\psi_{0,n},\pi)]\big)}{E[\varepsilon_t^2]} \\ &= r_t^{1,\psi,n}(h,\pi) + r_t^{2,\psi,n}(h,\pi) \end{aligned}$$

where $r_t^{1,\psi,n}(h,\pi) = \frac{\varepsilon_t\varepsilon_{t-h} - E[\varepsilon_t\varepsilon_{t-h}]}{E[\varepsilon_t^2]} - \frac{\mathcal{D}(h,\pi)'H^{-1}(\pi;\gamma_0)\big(m_t^{\psi}(\psi_{0,n},\pi) - E_{\gamma_n}[m_t^{\psi}(\psi_{0,n},\pi)]\big)}{E[\varepsilon_t^2]}$ and $r_t^{2,\psi,n}(h,\pi) = \frac{\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi) - \varepsilon_t\varepsilon_{t-h}}{E[\varepsilon_t^2]} - \frac{\mathcal{D}(h,\pi)'H^{-1}(\pi;\gamma_0)\big(\beta_n\frac{\partial}{\partial\tilde{\beta}}E_{\tilde{\gamma}_n}[m_t^{\psi}(\psi_{0,n},\pi)]\big)}{E[\varepsilon_t^2]}$.

Next, define $z_t^{i,\psi,n}(h,\pi) = r_t^{i,\psi,n}(h,\pi) - \rho(h)r_t^{i,\psi,n}(0,\pi)$ for $i = 1,2$, and observe that $z_t^{\psi,n}(h,\pi) = z_t^{1,\psi,n}(h,\pi) + z_t^{2,\psi,n}(h,\pi)$. Finally, define $\mathcal{Z}_n^{i,\psi}(h,\pi) = \frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t^{i,\psi,n}(h,\pi)$ for $i = 1,2$. We show in Lemma A.1.2 that $\mathcal{Z}_n^{1,\psi}(h,\pi)$ converges weakly to a Gaussian process and $\mathcal{Z}_n^{2,\psi}(h,\pi)$ converges uniformly in probability to a mean component. This leads to the following theorem stating the limit of the test statistic under H_0.

Theorem 2.3.2. *Let H_0 and Assumptions 3, 7, and 8 hold.*

(a) Let $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$, and additionally let Assumptions 1, 4(i), 5, 9, and 11(i) hold. Let $\{\mathcal{Z}^{\psi}(h,\pi) : h \in \mathbb{N}, \pi \in \Pi\}$ be a Gaussian process with finite mean $\lim_{n\to\infty}\sqrt{n}E_{\gamma_n}(r_t^{2,\psi,n}(h,\pi)) < \infty$ and variance $\lim_{n\to\infty}\frac{1}{n}\sum_{s,t=1}^{n}E_{\gamma_n}[r_s^{1,\psi,n}(h,\pi)r_t^{1,\psi,n}(h,\pi)] < \infty$ and covariance kernel $\lim_{n\to\infty}\frac{1}{n}\sum_{s,t=1}^{n}E_{\gamma_n}[r_s^{1,\psi,n}(h,\pi)\, r_t^{1,\psi,n}(\tilde{h},\tilde{\pi})]$. Then for some non-unique

[7] See Andrews and Cheng (2012a,b), especially Lemma 9.1.

sequence of positive integers $\{\mathcal{L}_n\}$ *with* $\mathcal{L}_n \to \infty$ *and* $\mathcal{L}_n = o(n)$,

$$\sup_{\pi \in \Pi} \left| \max_{1 \leq h \leq \mathcal{L}_n} (\sqrt{n}|\hat{\rho}_n(h,\pi) - \rho(h)|) - \max_{1 \leq h \leq \mathcal{L}_n} (|\frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} r_t^{\psi,n}(h,\pi)|) \right| \xrightarrow{p} 0 \qquad \textit{and}$$

$$\left| \max_{1 \leq h \leq \mathcal{L}_n} |\frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} r_t^{\psi,n}(h,\hat{\pi}_n)| - \max_{1 \leq h \leq \mathcal{L}_n} |\mathcal{Z}^{\psi}(h, \pi^*(b,\gamma_0))| \right| \xrightarrow{p} 0.$$

(b) Let $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$*, and additionally let Assumptions 2, 4(ii), 6, 10, and 11(ii) hold. Let* $\{\mathcal{Z}^{\theta}(h) : h \in \mathbb{N}\}$ *be a zero mean Gaussian process with variance* $\lim_{n\to\infty} \frac{1}{n} \sum_{s,t=1}^{n} E[r_s^{\theta}(h) r_t^{\theta}(h)] < \infty$ *and covariance kernel* $\lim_{n\to\infty} \frac{1}{n} \sum_{s,t=1}^{n} E[r_s^{\theta}(h) r_t^{\theta}(\tilde{h})]$*. Then for some non-unique sequence of positive integers* $\{\mathcal{L}_n\}$ *with* $\mathcal{L}_n \to \infty$ *and* $\mathcal{L}_n = o(n)$,

$$\left| \max_{1 \leq h \leq \mathcal{L}_n} (\sqrt{n}|\hat{\rho}_n(h) - \rho(h)|) - \max_{1 \leq h \leq \mathcal{L}_n} (|\frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} r_t^{\theta}(h)|) \right| \xrightarrow{p} 0 \qquad \textit{and}$$

$$\left| \max_{1 \leq h \leq \mathcal{L}_n} |\frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} r_t^{\theta}(h)| - \max_{1 \leq h \leq \mathcal{L}_n} |\mathcal{Z}^{\theta}(h)| \right| \xrightarrow{p} 0.$$

The limiting distribution of the test statistic under true $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ is the maximum of a Gaussian process. However, the limiting distribution of the test statistic under true $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$ is the maximum of a Gaussian process on Π evaluated at $\pi^*(\gamma_0, b)$. Further, the bias term $\lim_{n\to\infty} \sqrt{n} E_{\gamma_n}(r_t^{2,\psi,n}(h,\pi))$ is present under weak identification, whereas the limiting distribution under strong identification has mean zero. This mandates a different method for implementing the bootstrap under each identification scenario, as we describe in Section 2.4.1.

The limiting processes differ under weak and strong identification due to the inconsistency of $\hat{\pi}_n$ and the nonstandard limiting process of $\hat{\psi}_n$ under the case in which π_0 is weakly identified. In particular, when π_0 is weakly identified, we must expand around $\psi_{0,n}$, the subvector of the true parameter θ_n with β_n evaluated at the point of non-identification of π_0.

The finding that the limiting distribution of the white noise test statistic differs depending upon whether or not a consistent estimator is available for the potentially unidentified parameters gives us reason to believe that standard testing procedures may yield distorted inference. In particular,

it is well known that the distributions of the parameter estimators differ depending upon the identification strength of potentially unidentifiable parameters in the model, and standard inference on the parameters based on t or χ^2 distributions can be distorted when the model contains parameters that are weakly identified. Andrews and Cheng (2012a) demonstrate that densities of the estimators from an ARMA(1,1) model can be quite far from normal when the AR and MA parameters are close to the same value, and Cheng (2015) shows that using standard normal critical values for tests on a parameter from an additive nonlinear model with a weakly identified parameter can generate large size distortions. As shown in our Theorem 2.3.2, the impact of identification failure on the distributions of the estimators for a model can be noticed beyond tests on the parameter values. Our ARMA(1,1) model simulations indicate that that this difference can manifest itself in an empirically relevant way, leading to size distortions in white noise tests that ignore the effect of potential identification failure in the model.

2.3.3 Critical Values

Manufacturing a test that is robust to identification failure requires that we account for the possibility that our test statistic falls within the limiting distributions given by either identification regime. To this end, our bootstrapping procedure provides critical values under both situations. We then construct an identification robust critical value for our test statistic by stitching together the critical values found under each identification scenario using methods detailed in Andrews and Cheng (2012a).

We employ two types of robust critical values: least favorable (LF) and identification-category selection (ICS) critical values. The LF critical values always take the larger of the critical values found under each identification category, whereas the ICS critical values employ a data driven first step to determine if $b = \lim_{n\to\infty} \sqrt{n}\beta_n$ is finite or infinite.

Least Favorable Critical Values

Let $c_{1-\alpha}^{(w)}$ be the critical value under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$, and let $c_{1-\alpha}^{(s)}$ be the critical value under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$. The least favorable critical value is $c_{1-\alpha}^{(LF)} = \max\{c_{1-\alpha}^{(w)}, c_{1-\alpha}^{(s)}\}$. We reject the null hypothesis when $\hat{T}_n > c_{1-\alpha}^{(LF)}$.

Data Dependent Critical Values

The least favorable critical values can be improved by use of an identification-category-selection procedure outlined in Andrews and Cheng (2012a). The ICS procedure uses the available data to determine if b is finite, and hence whether $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$ or $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$. The LF critical value is used if the selection procedure suggests that b is finite, and the critical value under $||b|| = \infty$ is used otherwise. The statistic used for category selection is

$$\mathcal{A}_n = (n\hat{\beta}_n' \hat{\Sigma}_{\beta\beta,n}^{-1} \hat{\beta}_n / d_\beta)^{1/2}$$

where $\hat{\Sigma}_{\beta\beta,n}$ is the upper left $d_\beta \times d_\beta$ block of $\hat{\Sigma}_n = \hat{J}_n^{-1} \hat{V}_n \hat{J}_n^{-1}$, the estimator of the covariance matrix $\Sigma_n(\gamma_0) = J^{-1}(\gamma_0) V(\gamma_0) J^{-1}(\gamma_0)$.

Let $\{\kappa_n : n \geq 1\}$ be a sequence of constants that diverge to infinity as $n \to \infty$. We compare the statistic $\mathcal{A}_n$ to this sequence of tuning parameters in order to determine the identification category. Since $\mathcal{A}_n = O_p(1)$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$, this procedure consistently selects this category when $\kappa_n \to \infty$.

Assumption 12. *$\kappa_n \to \infty$ and $\kappa_n / n^{1/2} \to 0$.*

Assumption 12 is Assumption K in Andrews and Cheng (2012a).

The ICS critical value is $c_{1-\alpha}^{(ICS)} = \begin{cases} c_{1-\alpha}^{(LF)} & \text{if } \mathcal{A}_n \leq \kappa_n \\ c_{1-\alpha}^{(s)} & \text{if } \mathcal{A}_n > \kappa_n. \end{cases}$

Asymptotic Size

Let F_γ be the distribution function of $\mathcal{W}_t$ under some $\gamma \in \Gamma^*$, for the true parameter space Γ^*, and let P_γ denote probability under F_γ. For any critical value, $c_{1-\alpha,n}$, the asymptotic size of the test is the maximum rejection probability over Γ^* such that the null hypothesis is true:

$$AsySz = \limsup_{n\to\infty} \sup_{\gamma \in \Gamma^*} P_\gamma(\mathcal{T}_n > c_{1-\alpha,n} | \ H_0).$$

Assumption 13. *If $c^{(\cdot)}_{1-\alpha,n}$ is (i) LF or (ii) ICS, then assume that Andrews and Cheng's (2012a) Assumption (i) LF or (ii) K and V3 hold, respectively.*

Theorem 2.3.3. *Under Assumptions 12 and 13 and H_0, the LF and ICS critical values $c^{(\cdot)}_{1-\alpha,n}$ satisfy $AsySz = \alpha$.*

The proof of Theorem 2.3.3 is omitted as it follows directly from Andrews and Cheng (2012a).

2.4 Bootstrap Critical Value Computation

Standard arguments for critical value computation rely on computation of the exact limiting distribution of the test statistic by appealing to the extremal types theorem (see e.g. Xiao and Wu (2014); de Haan (1976)). Recent work for high dimensional statistics has focused on by-passing extreme value theory but has been limited by not allowing for dependence or residuals or by only allowing for Gaussian approximation (Chernozhukov et al., 2013, 2017; Zhang and Cheng, 2018; Zhang and Wu, 2017). Theory in Hill and Motegi (2018) and Hill and Dennis (2018) allows for dependence under the null, residuals, and does not require Gaussianity. Here, we side-step the extreme value theory asymptotics by using the approach found in Hill and Motegi (2018) and Hill and Dennis (2018) paired with the dependent wild bootstrap of Shao (2010, 2011a).

The wild bootstrap is a multiplier bootstrap. Wu (1986) and Liu (1988) detail the classic wild bootstrap for iid sequences. Hansen (1996) allows for adapted martingale difference sequences, and Shao (2010, 2011a) allows for dependent sequences. Shao (2010) uses iid random draws as weights with a kernel function, but does not allow for a truncated kernel. Shao (2011a) uses a truncated kernel function.

In order to compute robust critical values, we consider least favorable (LF) and information criteria selection (ICS) critical values. This involves computation of critical values under both weak and strong identification. We follow Shao (2011a) to compute critical values under each scenario. Below, we first elucidate the algorithm used to perform the bootstrap under each identification scenario. Then we discuss critical value computation.

2.4.1 Bootstrap Algorithm

Here we detail the bootstrap algorithm for computing critical values under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$ (weak identification) and $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ (strong identification).

First, we draw standard normal random variables with perfect dependence within blocks and independence across blocks. This sequence of random normals forms the Gaussian multiplier used in the wild bootstrap of Shao (2011a). It is important to note that the multiplier random variables need not be Gaussian; however Gaussianity greatly simplifies arguments in the proofs.

Begin by selecting a block size k_n s.t. $1 \leq k_n \leq n$, $k_n \to \infty$, and $k_n/n \to 0$. Define blocks by $\mathbb{B}_s = \{(s-1)k_n + 1, \ldots, sk_n\}$ for $s = 1, \ldots, n/k_n$. Generate iid $N(0,1)$ random variables $\{\tilde{\xi}_1, \ldots, \tilde{\xi}_{n/k_n}\}$ and define $z_t = \tilde{\xi}_s$ if $t \in \mathbb{B}_s$.

The bootstrapped critical values must be computed separately for each identification scenario. The algorithms, which we detail next, are similar under weak and strong identification. They differ due to the different expansions needed in order to account for the impact of parameter estimation on the residuals.

Weak Identification

Under strong identification, the bootstrap only needs to replicate the randomness underlying $\varepsilon_t \varepsilon_{t-h}$ and m_t^θ. However, under weak identification, an additional source of randomness is present due to the inconsistency of $\hat{\pi}_n$. Therefore, the bootstrap under weak identification must also replicate the underlying randomness from the limiting distribution $\pi^*(b, \gamma_0)$. Further, the additional bias terms arise in the first order expansion under the case of weak identification that are not present under strong identification. The bootstrap will replicate these sources of randomness in two steps.

First, we simulate a random draw, $\pi^*_{(bs)}(b, \gamma_0)$, from the distribution $\pi^*(b, \gamma_0)$ by using the draws z_t mentioned above. Next, we use $\pi^*_{(bs)}(b, \gamma_0)$ to construct the components of our test statistic under weak identification, which are functions of π. Then we again use the draws z_t to construct the wild bootstrap version of the test statistic. Finally, b and γ_0 are nuisance parameters that we must deal with. The algorithm is as follows.

First, compute $\hat{H}_n(\pi)$ and $\hat{K}_n(\pi; \gamma_0)$. For example, in the STAR(1) model, $\hat{H}_n(\pi) = \frac{1}{n}\sum_{t=1}^n d_{\psi,t}(\pi) d_{\psi,t}(\pi)'$ and $\hat{K}_n(\pi; \gamma_0) = \frac{1}{n}\sum_{t=1}^n d_{\psi,t}(\pi) x_t' g(z_t, \pi_0)$.

Next, recall that $\Omega(\pi, \tilde{\pi}; \gamma_0) = E[G(\pi; \gamma_0)G(\tilde{\pi}; \gamma_0)]$ is the a.s. p.d. covariance kernel for $G(\cdot; \gamma_0)$. Recall also that $G_n(\pi; \gamma_n) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} m_t^{\psi}(\pi)$. Compute $\hat{G}_n^{(bs)}(\pi) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} z_t \big[m_t^{\psi}(\pi) - \frac{1}{n} \sum_{t=1}^{n} m_t^{\psi}(\pi)\big]$. In the STAR(1) model,
$\hat{G}_n^{(bs)}(\pi) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} z_t \big[d_{\psi,t}(\pi)\varepsilon_t(\hat{\psi}_{0,n}(\pi), \pi) - \frac{1}{n} \sum_{t=1}^{n} \big[d_{\psi,t}(\pi)\varepsilon_t(\hat{\psi}_{0,n}(\pi), \pi)\big]$.

Define

$$\xi_n^{(bs)}(\pi; \gamma_0, b) = -\frac{1}{2}\Big(\hat{G}_n^{(bs)}(\pi) + \hat{K}_n(\pi; \gamma_0)b\Big)'(\hat{H}_n(\pi))^{-1}\Big(\hat{G}_n^{(bs)}(\pi) + \hat{K}_n(\pi; \gamma_0)b\Big),$$

and compute $\pi_{(bs)}^*(\gamma_0, b) = \underset{\pi \in \Pi}{\operatorname{argmin}}\ \xi_n^{(bs)}(\pi; \gamma_0, b)$.

Now use $\pi_{(bs)}^*(\gamma_0, b)$ and $\hat{\psi}_{0,n}$ to compute the quantities

$$\mathcal{G}_n(\pi_{(bs)}^*) = (\hat{H}_n(\pi_{(bs)}^*))^{-1}\Big[m_t^{\psi}(\hat{\psi}_{0,n}, \pi_{(bs)}^*) - \frac{1}{n}\sum_{t=1}^{n} m_t^{\psi}(\hat{\psi}_{0,n}, \pi_{(bs)}^*)\Big]$$

and

$$\hat{\mathcal{D}}_n(h, \pi_{(bs)}^*) = \frac{1}{n}\sum_{t=1+h}^{n} [d_{\psi,t}(\pi_{(bs)}^*)\varepsilon_{t-h}(\hat{\psi}_{0,n}, \pi_{(bs)}^*) + d_{\psi,t-h}(\pi_{(bs)}^*)\varepsilon_t(\hat{\psi}_{0,n}, \pi_{(bs)}^*)].$$

Define

$$\begin{aligned}\hat{\mathcal{E}}_{t,h}(\psi, \pi) = \varepsilon_t(\psi, \pi)\varepsilon_{t-h}(\psi, \pi) - \mathcal{G}_n(\pi_{(bs)}^*)'\hat{\mathcal{D}}_n(h, \pi_{(bs)}^*) \\ - \frac{1}{n}\sum_{t=1+h}^{n} [\varepsilon_t(\psi, \pi)\varepsilon_{t-h}(\psi, \pi) - \varepsilon_t\varepsilon_{t-h}],\end{aligned}$$

and the draws $\{z_t\}$ to define

$$\begin{aligned}\hat{\rho}_n^{(w)}(h; \gamma_n, b) = \frac{1}{n^{-1}\sum_{t=1}^{n}\varepsilon_t^2(\hat{\theta}_n)} \times \Bigg\{\frac{1}{n}\sum_{t=1+h}^{n} z_t\Big(\hat{\mathcal{E}}_{t,h}(\hat{\psi}_{0,n}, \pi_{(bs)}^*) - \frac{1}{n}\sum_{t=1+h}^{n}\hat{\mathcal{E}}_{t,h}(\hat{\psi}_{0,n}, \pi_{(bs)}^*)\Big) \\ - ((\hat{H}_n(\pi_{(bs)}^*))^{-1}\hat{K}_n(\pi; \gamma_n)\frac{b}{\sqrt{n}})'\hat{\mathcal{D}}_n(h, \pi_{(bs)}^*) \\ + \frac{1}{n}\sum_{t=1+h}^{n} [\varepsilon_t(\hat{\psi}_{0,n}, \pi)\varepsilon_{t-h}(\hat{\psi}_{0,n}, \pi) - \varepsilon_t\varepsilon_{t-h}]\Bigg\}.\end{aligned}$$

Observe that we subtract $\frac{1}{n}\sum_{t=1+h}^{n}[\varepsilon_t(\psi,\pi)\varepsilon_{t-h}(\psi,\pi)]$ from the component that will multiplied by the random variable z_t, and then add $\frac{1}{n}\sum_{t=1+h}^{n}[\varepsilon_t(\psi,\pi)\varepsilon_{t-h}(\psi,\pi)]$ after. Recall that the distribution of the test statistic has mean $\lim_{n\to\infty}\sqrt{n}E_{\gamma_n}[\varepsilon_t(\psi_{0,n},\pi^*)\varepsilon_{t-h}(\psi_{0,n},\pi^*)]$, hence the reason for adding the quantity after multiplication by z_t. We must subtract the quantity prior to multiplication as failing to do so will bias the variance of the bootstrapped distribution, since the variance in the distribution of the test statistic is based on $\varepsilon_t\varepsilon_{t-h}$, but our critical values are constructed using $\varepsilon_t(\hat{\psi}_{0,n},\hat{\pi}_n)\varepsilon_{t-h}(\hat{\psi}_{0,n},\hat{\pi}_n)$.

We now define the bootstrapped test statistic $\hat{\mathcal{T}}_n^{(w)}(\gamma_n,b) = \max_{1\leq h\leq\mathcal{L}_n}|\sqrt{n}\ \hat{\rho}_n^{(w)}(h;\gamma_n,b)|$. Critical value construction, which is based on order statistics of repeated draws of $\hat{\mathcal{T}}_n^{(w)}(\gamma_n,b)$, is defined below. Observe that the bootstrapped test statistic is a function of the nuisance parameters (π_n,b) which cannot be consistently estimated. Therefore, we will define the α-level critical value $c_{n,1-\alpha}^{(w)} = \sup_{\pi_n,b} c_{n,1-\alpha}^{(w)}(\gamma_n,b)$.

Strong Identification

The procedure for generating bootstrapped test statistics under the case of (semi-) strong identification is similar to that under weak identification; however, $\hat{\pi}_n$ consistently estimates the true value π_n in this case. Due to this, we simply use the full estimated parameter vector $\hat{\theta}_n$, as the layer involving generation of random draws from the distribution $\pi^*(\gamma_0,b)$ is not warranted. In addition, the test statistic has mean zero, so the bootstrap test statistic follows a simpler construction.

Compute $\hat{J}_n(\hat{\theta}_n)$ and $\hat{\mathcal{D}}_n^{\theta}(h,\hat{\theta}_n) = \frac{1}{n}\sum_{t=1+h}^{n}[d_{\theta,t}(\hat{\theta}_n)\varepsilon_{t-h}(\hat{\theta}_n) + d_{\theta,t-h}(\hat{\theta}_n)\varepsilon_t(\hat{\theta}_n)]$. Define $\hat{\mathcal{E}}_{t,h}(\theta) = \varepsilon_t(\theta)\varepsilon_{t-h}(\theta) - (B(\hat{\beta}_n)^{-1}\hat{\mathcal{D}}_n^{\theta}(h,\theta))'(\hat{J}_n(\hat{\theta}_n))^{-1}m_t^{\theta}(\theta)$. Use the draws $\{z_t\}$ to define

$$\hat{\rho}_n^{(s)}(h) = \frac{1}{n^{-1}\sum_{t=1}^{n}\varepsilon_t^2(\hat{\theta}_n)} \times \left\{\frac{1}{n}\sum_{t=1+h}^{n} z_t\Big(\hat{\mathcal{E}}_{t,h}(\hat{\theta}_n) - \frac{1}{n}\sum_{t=1+h}^{n}\hat{\mathcal{E}}_{t,h}(\hat{\theta}_n)\Big)\right\}$$

Finally, define the bootstrapped test statistic $\hat{\mathcal{T}}_n^{(s)} = \max_{1\leq h\leq\mathcal{L}_n}|\sqrt{n}\ \hat{\rho}_n^{(s)}(h)|$. Observe that there are no nuisance parameters in this case.

Remark 6. *Note that our bootstrapped test statistics rely on sample versions of the first order expansion $\varepsilon_t\varepsilon_{t-h} - \mathcal{D}^{\theta}(h)J^{-1}m_t^{\theta}$ as in Hill and Motegi (2018). It is incorrect to simply use*

$\varepsilon_t(\hat{\theta}_n)\varepsilon_{t-h}(\hat{\theta}_n)$ *without the term from the first order expansion since the bootstrap multiplier random variables* z_t *are mean zero and independent of the data. One can show that* $\frac{1}{n}\sum_{t=1+h}^{n} z_t\varepsilon_t(\hat{\theta}_n)\varepsilon_{t-h}(\hat{\theta}_n) = \frac{1}{n}\sum_{t=1+h}^{n}\varepsilon_t\varepsilon_{t-h} + o_p(1/\sqrt{n})$*, whereas first order arguments show* $\frac{1}{n}\sum_{t=1+h}^{n}\varepsilon_t(\hat{\theta}_n)\varepsilon_{t-h}(\hat{\theta}_n) = \frac{1}{n}\sum_{t=1+h}^{n}\varepsilon_t\varepsilon_{t-h} + O_p(1/\sqrt{n})$*. Ignoring the first order expansion term therefore results in loss of information from the estimator* $\hat{\theta}_n$*.*

The same argument applies to the case of weak identification. The fact that the limiting distribution of the estimator differs under weak identification is precisely why Hill and Motegi (2018) cannot accommodate weakly identified models.

Critical Value Computation

Repeat the above procedures $i = 1, \ldots, M$ times to define $\{\hat{\mathcal{T}}_{n,i}^{(s)}\}_{i=1}^{M}$ and $\{\hat{\mathcal{T}}_{n,i}^{(w)}(\gamma_n, b)\}_{i=1}^{M}$. For $k = w, s$, define the order statistics $\{\hat{\mathcal{T}}_{n,(i)}^{(k)}\}_{i=1}^{M}$ such that $\hat{\mathcal{T}}_{n,(1)}^{(k)} \leq \hat{\mathcal{T}}_{n,(2)}^{(k)} \leq \cdots \leq \hat{\mathcal{T}}_{n,(M)}^{(k)}$. The approximate α-level critical values under weak and strong identification, respectively, are $\hat{c}_{n,1-\alpha}^{(w)}(\gamma_n, b) = \hat{\mathcal{T}}_{n,[(1-\alpha)\cdot M]}^{(w)}(\gamma_n, b)$ and $\hat{c}_{n,1-\alpha}^{(s)} = \hat{\mathcal{T}}_{n,[(1-\alpha)\cdot M]}^{(s)}$. Observe that the α-level critical value under weak identification is nuisance parameter dependent. Therefore, define the approximate α-level critical value $\hat{c}_{n,1-\alpha}^{(w)} = \sup_{\pi_n, b}\hat{c}_{n,1-\alpha}^{(w)}(\gamma_n, b)$.

Theorem 2.4.1. *Let Assumptions 1 - 11 hold, and let the number of bootstrap samples* $M_n \to \infty$*. Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $||b|| < \infty$*, let* $k = w$ *and under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$*, let* $k = s$*. There is a non-unique sequence of positive integers* $\{\mathcal{L}_n\}$ *with* $\mathcal{L}_n \to \infty$ *and* $\mathcal{L}_n = o(n)$ *such that* $\hat{c}_{1-\alpha,n}^{(k)} \xrightarrow{p} c_{1-\alpha}^{(k)}$*.*

Moreover, under the alternative hypothesis, $P(\hat{T}_n > \hat{c}_{1-\alpha,n}^{(k)}) \to 1$ *for* $k = w$ *and* $k = s$*.*

The bootstrapped critical values are constructed assuming the null hypothesis is true. Theorem 2.4.1 shows that the bootstrapped critical values are consistent for the critical values from the null limiting distribution of the test statistic. This indicates that under H_0, the test achieves correct size asymptotically when the correct (e.g. either weak or strong) critical value is used. Theorem 2.3.3 then allows us to construct correctly sized tests by combining these critical values into LF and ICS critical values.

Further, since the tests based on weak and strong critical values are both consistent against

the alternative hypothesis, tests based on the LF and ICS critical values are also consistent against the alternative. Note that the critical values are dependent upon the sequence $\{h : 1 \leq h \leq \mathcal{L}_n\}$ because they are computed using the sample correlations out to lag $\mathcal{L}_n$. This dependence is suppressed in the notation of Theorem 2.4.1.

2.5 Simulations

In this section we perform Monte-Carlo experiments to demonstrate the benefits of our robust max-correlation test. We perform simulations assuming that the true nuisance parameters are known, which we call infeasible simulations, and we perform simulations using a grid of π and b to calculate correlation expansions under weak identification, which we call feasible simulations. We simulate $J = 1000$ samples to be used in infeasible simulations and $J = 500$ samples to be used in feasible simulations of size $n \in \{100, 250, 500, 1000\}$ from the following processes:

$$\begin{aligned} &\text{STAR(1):} \quad y_t = \beta_n y_{t-1}(1 + \exp(-10(y_{t-1})))^{-1} + .5y_{t-1} \\ &\text{ARMA(1,1):} \quad y_t = (\beta_n + .5)y_{t-1} + \varepsilon_t - .5\varepsilon_{t-1} \end{aligned}$$

for values of $\beta_n \in \{0, .3/\sqrt{n}, .3\}$ that satisfy identification failure, weak identification, and strong identification, respectively. The STAR(1) model is estimated via least squares, and the ARMA(1,1) model is estimated with an ARMA(1,1) filter by QML. Recall that when $\beta_n = 0$, the ARMA(1,1) model reduces to the process $y_t = \varepsilon_t$; hence π_n is not identifiable.

Both the ARMA and Smooth Transition Regression (STR) models have been highly useful models to practitioners for decades. In particular, Andrews and Ploberger (1996) note that AMRA(1,1) models provide parsimonious representations of many different stationary time series. Poterba and Summers (1988) show that many mean-reverting financial time series can be represented by ARMA(1,1) models, and Taylor (2005) shows that the ARMA model can be used to represent certain price-trend models.

STR models have spawned a broad class of time series representations. The Logistic STAR (LSTAR) model that we examine in these simulations has been used to model business cycle asymmetry with regimes associated with recession and expansions (Teräsvirta and Anderson, 1992;

Skalin and Teräsvirta, 2001). When the variable controlling the speed of transition approaches infinity, the logistic function approaches an indicator function; hence, the LSTAR model itself nests many Threshold Autoregressive (TAR) models as a special case. See Hansen (2011) for a survey of the history of TAR models in Economics.

Changing the transition function leads to other classes of STR models that have been used to model other phenomena. Exponential STAR models have been used to explain the nonlinear dependence of the real exchange rate on the size of the deviation from purchasing power parity, and higher order logistic functions have been used to allow multiple switches between regimes. Multiple regime models (van Dijk and Franses, 1999) have also been explored to describe business cycle nonlinearity, and Multiple Regime STAR models have been shown to nest many models as special cases, including certain types of artificial neural networks.[8] Taking the transition variable to be time leads to the time-varying STAR model (Lundbergh, Teräsvirta, and van Dijk, 2000), which is related to structural instability of a time series, and has been used to study seasonal patterns in industrial production (van Dijk, Strikholm, and Teräsvirta, 2001). STAR models have been extended to allow cointegration and error-correction models in vector frameworks as well.[9]

With ARMA, STR and STAR models and similar models, it is common to assume that the errors ε_t are martingale difference sequences.[10] This leads to a natural model diagnostic test in the form of a test of no serial correlation in the errors.

Let $\nu_t \sim$ iid $N(0,1)$. For the error ε_t, we consider several different processes: 3 null hypothesis models and several alternative hypothesis models designed to provide alternatives that vary in difficulty for the tests to discover. We consider the following 3 null hypothesis models (H_0):

iid: $\varepsilon_t = \nu_t$

[8] Artificial neural networks can be used to approximate continuous functions to an arbitrary degree of accuracy. See Kuan and White (1994) for a review of ANNs.

[9] e.g. Taylor, van Dijk, Franses, and Lucas (2000) use a Smooth Transition Error Correction Model to examine the relationship between spot and futures prices of the FTSE 100, Anderson (1997) and van Dijk and Franses (2000) examine the term structure of interest rates, Swanson (1999) and Rothman, van Dijk, and Franses (2001) examine the relationship between money and output, and Dwyer, Locke, and Yu (1996), Martens, Kofman, and Vorst (1998) and Tsay (1998) examine spot and futures prices of the S&P 500.

[10] see e.g. Nankervis and Savin (2010), Terasvirta (1994); Teräsvirta (1998) and the review by van Dijk et al. (2002).

GARCH(1,1): $\varepsilon_t = \sigma_t \nu_t, \quad \sigma_t^2 = 1 + .3\varepsilon_{t-1}^2 + .6\sigma_{t-1}^2$

Bilinear: $\varepsilon_t = .5\varepsilon_{t-1}\nu_{t-1} + \nu_t$

and the following alternative hypotheses (H_A):

AR(2): $\varepsilon_t = .5\varepsilon_{t-2} + \nu_t$

MA(1): $\varepsilon_t = .5\nu_{t-1} + \nu_t$

MA(10): $\varepsilon_t = .5\nu_{t-10} + \nu_t$

MA(21): $\varepsilon_t = .5\nu_{t-21} + \nu_t$

In order to examine power against distant alternatives for the longer sample lengths, we additionally consider a MA(50) alternative when $n = 500$ and a MA(100) alternative when $n = 1000$.

We perform the identification-robust Max Correlation test using both Least Favorable (LF) and Identification Category Selection (ICS) critical values. Futher, we perform Hong's (1996) Standardized Ljung-Box Q test (LBQ), Shao's (2011a) Cramér von Mises Test (CvM), and Nankervis and Savin's (2010) representation of Andrews and Ploberger's (1996) sup-LM test (supLM). For all tests, we report rejection frequencies for both feasible and infeasible tests based on ICS, LF, strong identification critical values only (S), and critical values obtained without using a first order expansion (NoX). Note that the Max Correlation test in Hill and Motegi (2018) is simply the Max Correlation utilizing the strong identification critical values only, so the tests labeled "MC S" correspond to this test. Additionally, for comparison with Hill and Motegi (2018), we report p-value based rejection frequencies in the supplemental appendix.

To satisfy space requirements, a selection of critical value based rejection frequency tables for $n = 100$ and $n = 500$ under weak id are presented in section 2.5.1. All tables are presented in the supplemental appendix.

Max Correlation Test The max-correlation tests require a lag length $\mathcal{L}_n$. We used a fixed $\mathcal{L}_n = 5$ and sample size dependent $\mathcal{L}_n \in \{[n^{1/3}], [\sqrt{n}/(\ln(n)/4)], [\sqrt{n}/(\ln(n)/5)], [\sqrt{n}-1], [.5n/\ln(n)]\}$. This implies that $\mathcal{L}_n \in \{4, 5, 8, 9, 10\}$ when $n = 100$, $\mathcal{L}_n \in \{5, 6, 11, 14, 22\}$ when $n = 250$, $\mathcal{L}_n \in$

$\{5, 7, 14, 17, 21, 40\}$ when $n = 500$, $\mathcal{L}_n \in \{5, 9, 18, 22, 30, 72\}$ when $n = 1000$. Additionally, we use $\mathcal{L}_n = [n/\ln(n)]$ when $n = 500, 1000$ leading to lag lengths 80 and 144, respectively. We expect visible size distortions with the longer lag selections for the larger sample sizes; in particular, we expect the sampling error at longer lags to lead the max statistic to exhibit larger variance yielding under-sized tests. The test statistic is $\sqrt{n}\max_{1\leq h\leq \mathcal{L}_n}|\hat{\rho}_n(h)|$.

Infeasible critical values, for which the nuisance parameters are known, are computed by dependent wild bootstrap with $M = 1000$ bootstrap samples for $T = 100, 250$ and $M = 500$ bootstrap samples for $T = 500, 1000$ using the first order expansion given in Lemma 2.3.1. Due to the greater computational requirements for computing the feasible critical values, for which all correlations expansions under weak identification must be computed over a grid of π and b, we use $M = 500$ bootstrap samples. For the DWB block size, we use $k_n = [\sqrt{n}]$, where $[\cdot]$ is the truncation operator.

Standardized Ljung-Box Q Test The standardized Ljung-Box statistic is $\mathbb{N}_n = (2\mathcal{L}_n)^{-1/2}\sum_{h=1}^{\mathcal{L}_n} w_n(h)\{n\hat{\rho}_n^2(h) - 1\}$ where $w_n(h) = (n+2)/(n-h)$. Under Hong's (1996) assumptions, $\mathbb{N}_n \xrightarrow{d} N(0,1)$ when the null hypothesis is true and the $\{n\hat{\rho}_n^2(h) : 1 \leq h \leq \mathcal{L}_n\}$ are asymptotically independent. This asymptotic independence typically fails for models with martingale difference errors, so we expect this test to exhibit size distortions when ε_t is dependent under H_0 as with GARCH and bilinear errors. We perform a bootstrapped version of the test using the Lemma 2.3.1 first order expansion.

Cramér-von-Mises Test Shao's (2011a) statistic is $\mathcal{C}_n = \int_0^\pi S_n^2(\lambda)d\lambda$ where $S_n(\lambda) = \sum_{h=1}^{n-1}\sqrt{n}\hat{R}_n(h)\psi_h(\lambda)$, $\hat{R}_n(h)$ is the sample covariance, and $\psi_n(\lambda) = (h\pi)^{-1}\sin(h\lambda)$ if $h \neq 0$ and $\psi_n(\lambda) = \lambda/(2\pi)$ if $h = 0$. We approximate the integral with the sum $\sum_{i=1}^{3124} S_n^2(\lambda_i)$ and discretization of λ by the grid $\lambda_i = 0, .001, \ldots, 3.141$. We again implement the test with the DWB based on the correlation expansion detailed in Lemma 2.3.1.

Sup LM Test The sup-LM test is based on Nankervis and Savin's (2010) representation of Andrews and Ploberger's (1996) sup-LM statistic $\mathcal{AP}_n(n-1) = \sup_{\lambda\in\Lambda} LM_n(\lambda, n-1)$, where $LM_n(\lambda, n-1) = n(1-\lambda^2)\big[\sum_{h=1}^{n-1}\lambda^{h-1}\hat{\rho}_n(h)\big]^2$. The generalized AP test of Nankervis and Savin

(2010) are consistent against all nonwhite noise alternatives and have good power against nonseasonal alternatives compared to many other tests in the literature. We compute $\mathcal{AP}_n(\mathcal{L}_n)$ with the discretized parameter space $\Lambda = [-.8, -.795, \ldots, 0, .005, \ldots, .8]$. We implement the test with the DWB based on the correlation expansion in Lemma 2.3.1.

2.5.1 Simulation Results: STAR(1) Model

Recall from remark 6 that ignoring the first order expansion term results in loss of information from the estimator $\hat{\theta}_n$, since the multipliers used in the bootstrap are mean zero and independent of the data. Initial observations from the STAR(1) model indicate that the tests constructed without use of the first order expansion may under perform in some situations. In particular, it appears that the rejection frequency is usually lower than its expansion based counterparts. An exception to this is found in the case of bilinear errors when we see that the NoX-based tests give higher rejection frequencies, but this could be an empirical artifact due to a small number of simulations. This lower rejection frequency is not noticeable under every alternative hypothesis scenario that we consider; however, it is noticeable under the distant, weak alternatives considered in particular for the sup LM and CvM tests. The sup LM and CvM tests perform poorly against such alternatives in general.

Next, recall that the least favorable (LF) critical values are constructed by always taking the larger of the critical values found assuming weak and strong identification, so we see that tests based on these critical values tend to have lower rejection frequencies than other tests. The critical values attained under weak identification tend to be larger than those attained under strong identification, so the effect is particularly prominent when the truth is strong identification. When π is weakly identified, tests based on the LF critical values have rejection frequencies similar to those based on the ICS critical values.

Interestingly, tests based on the LF critical values are the only tests that do not tend to over-reject the null hypothesis under bilinear errors. In fact, in this situation, these tests always have near zero rejection frequencies. It seems that under this specification, not only do the weak identification critical values tend to be large under strong identification, but the strong identification critical values also tend to be large under weak identification. This has the effect of causing the LF critical

values to always be large, leading to very small rejection frequencies. Due to these issues and those mentioned for the NoX based tests, we will not mention the LF or NoX based tests in the discussion that follows.

Tests based on Identification-Category-Selection (ICS) critical values tend to perform fairly well under the specifications tested here. Empirical size tends to be close to or less than nominal, and power under the alternative hypothesis tends to be relatively high. Exceptions include the cases in which we specify bilinear errors. Recall that bilinear $\{\varepsilon_t\}$ is a non-mds white noise process; hence it is a null hypothesis specification. However, our tests tend to exhibit much larger than nominal rejection frequencies in this case. This does not seem to represent a disadvantage when compared to tests based on other critical value constructions, as they all tend to perform poorly for bilinear errors.

Table 2.1: White Noise Test Simulations - STAR model without expansion

	No Id	Weak Id	Strong Id
MC NoX	0.04	0.03	0.04
LBQ NoX	0.03	0.02	0.02
sup LM NoX	0.02	0.01	0.02
CvM NoX	0.00	0.00	0.02

STAR model without expansion. Rejection Frequencies: Infeasible CV based Tests, STAR(1) model with iid errors, $\alpha = 0.05$, $T = 500$, $\mathcal{L}_n = 5$, $J = 500$.

Table 2.2: White Noise Test Simulations - Least Favorable Critical Values

	$T = 100$	$T = 500$
MC LF	0.02	0.01
LBQ LF	0.02	0.01
sup LM LF	0.00	0.00
CvM LF	0.01	0.01

STAR model under Strong Identification with LF CVs. Rejection Frequencies: Infeasible CV based Tests, STAR(1) model with iid errors, Strong Identification, $\alpha = 0.05$, $\mathcal{L}_n = 5$, $J = 500$.

Recall that since the sup LM and CvM tests do not require a maximum lag, so we use a lag length of $n - 1$ for these tests. the MC and LBQ tests, however, require specification of a lag length, and we provide a range of lags to check finite sample performance. The most obvious

observation is that we experience rejection frequency shrinkage across all specifications and critical value constructions as $\mathcal{L}_n$ increases, holding n fixed. This affects specifications under both the null and alternative hypotheses. It appears that this is an artifact of increasingly noisy estimates at longer lags making their way into the critical values through the dependent wild bootstrap.

Now, recall that one of the primary reasons for considering a max test in this paper is that the max test tends to offer sharper estimates at longer lags when compared to tests based on averages of correlations, as the max test only relies on the most informative sample correlation estimate. Due to this, we expect the issue of rejection frequency shrinkage to be less pronounced for the max correlation test than for the LBQ test, a result that we do find in our simulations. The infeasible ICS statistics tend to have close to nominal size at the shortest lag length tested, $\mathcal{L}_n = 5$.

Table 2.3: White Noise Test Simulations - Size Shrinkage

	$\mathcal{L}_n = 5$	$\mathcal{L}_n = 22$	$\mathcal{L}_n = 40$	$\mathcal{L}_n = 80$
MC ICS	0.04	0.02	0.01	0.01
LBQ ICS	0.02	0.00	0.00	0.00

Size Shrinkage in STAR-GARCH model under Weak Identification with ICS. Rejection Frequencies: Infeasible CV based Tests, STAR(1) with GARCH(1,1) errors, Weak Identification, $\alpha = 0.05$, $T = 500$, $J = 500$.

For specifications in which H_0 is false, we see that MC as comparable pwoer to LBQ and sometimes smaller power than sup LM and CvM when the dependence under H_A is easily detectable; that is for error specifications with close dependence such as the AR(2) and MA(1). However, for specifications with weak distant dependence, the max correlation test dominates the other tests. This is due to the fact that the max correlation test utilizes the most informative sample correlation estimate rather than smoothing over many possibly noisy estimates.

The maximum correlation test constructed using only the strong identification critical values (MC S) is the test of Hill and Motegi (2018). We include tests constructed by using only strong identification critical values in oder to compare the identification robust critical value based tests to these tests and in particular to the test of Hill and Motegi (2018). We find that when π is strongly identified, the ICS based tests are comparable to the tests that use only the critical values

constructed under strong identification. This seems to indicate that the ICS selection procedure works well in practice.

However, we also find that when π is non- or weakly-identified, the ICS based tests perform similarly to the tests based on the strong identification critical values only. Hence, based on these specifications, it appears that MC ICS does not dominate MC S under weak or non identification of π. It is possible that this is an artifact of the chosen model specifications, as the simulations based on the ARMA(1,1) model detailed below indicate that when using only the strong identification expansion when the truth is weak identification results in size distortions under the null.

Table 2.4: White Noise Test Simulations - STAR model under Strong Identification

	H_0 True		H_0 False			
	Infeasible		Infeasible		Feasible	
	iid	GARCH(1,1)	AR(2)	MA(10)	AR(2)	MA(10)
	$\mathcal{L}_n = 5$	$\mathcal{L}_n = 5$	$\mathcal{L}_n = 5$	$\mathcal{L}_n = 10$	$\mathcal{L}_n = 5$	$\mathcal{L}_n = 10$
MC ICS	0.05	0.03	0.84	0.74	0.36	0.36
LBQ ICS	0.04	0.03	0.77	0.57	0.33	0.28
sup LM ICS	0.06	0.04	0.69	0.11	0.32	0.06
CvM ICS	0.00	0.00	0.00	0.00	0.00	0.01
MC LF	0.03	0.02	0.57	0.47	0.36	0.36
LBQ LF	0.02	0.02	0.53	0.31	0.33	0.28
sup LM LF	0.02	0.01	0.47	0.05	0.32	0.06
CvM LF	0.00	0.00	0.00	0.00	0.00	0.01
MC S	0.06	0.05	0.92	0.83	0.92	0.83
LBQ S	0.06	0.05	0.84	0.68	0.83	0.66
sup LM S	0.10	0.07	0.75	0.18	0.74	0.16
CvM S	0.01	0.01	0.01	0.01	0.01	0.01
MC NoX	0.03	0.03	0.87	0.81	0.86	0.79
LBQ NoX	0.04	0.02	0.75	0.61	0.75	0.59
sup LM NoX	0.04	0.03	0.63	0.08	0.65	0.07
CvM NoX	0.00	0.00	0.00	0.00	0.00	0.00
MC W	0.03	0.02	0.57	0.47	0.38	0.43
LBQ W	0.02	0.02	0.53	0.32	0.39	0.40
sup LM W	0.02	0.02	0.49	0.05	0.41	0.41
CvM W	0.13	0.13	0.10	0.15	0.37	0.41

Rejection Frequencies: Feasible CV based Tests, STAR1, Strong Id, $\alpha = 0.05$, $T = 100$, $\beta_n = 0.300$, $J = 500$.

Table 2.5: White Noise Test Simulations - STAR model under Weak Identification

	H_0 True		H_0 False			
	Infeasible		Infeasible		Feasible	
	iid	GARCH(1,1)	AR(2)	MA(10)	AR(2)	MA(10)
	$\mathcal{L}_n = 5$	$\mathcal{L}_n = 5$	$\mathcal{L}_n = 5$	$\mathcal{L}_n = 10$	$\mathcal{L}_n = 5$	$\mathcal{L}_n = 10$
MC ICS	0.04	0.03	0.89	0.78	0.24	0.20
LBQ ICS	0.03	0.03	0.83	0.63	0.26	0.13
sup LM ICS	0.06	0.05	0.79	0.12	0.26	0.02
CvM ICS	0.00	0.00	0.00	0.00	0.00	0.00
MC LF	0.04	0.03	0.84	0.74	0.24	0.20
LBQ LF	0.03	0.03	0.75	0.59	0.26	0.13
sup LM LF	0.06	0.05	0.70	0.11	0.26	0.02
CvM LF	0.00	0.00	0.00	0.00	0.00	0.00
MC S	0.06	0.05	0.93	0.82	0.93	0.82
LBQ S	0.05	0.05	0.86	0.71	0.86	0.67
sup LM S	0.11	0.07	0.84	0.18	0.84	0.15
CvM S	0.00	0.00	0.01	0.01	0.00	0.00
MC NoX	0.02	0.03	0.85	0.80	0.88	0.79
LBQ NoX	0.02	0.02	0.74	0.63	0.74	0.58
sup LM NoX	0.02	0.02	0.65	0.05	0.69	0.03
CvM NoX	0.00	0.00	0.00	0.00	0.00	0.00
MC W	0.04	0.03	0.84	0.75	0.26	0.23
LBQ W	0.04	0.03	0.76	0.60	0.29	0.18
sup LM W	0.06	0.05	0.72	0.11	0.29	0.15
CvM W	0.07	0.08	0.09	0.09	0.21	0.15

Rejection Frequencies for Feasible CV based Tests. STAR(1) model, Weak Id, $\alpha = 0.05$ $T = 100$, $\beta_n = 0.030$, $J = 500$.

2.5.2 Simulation Results: ARMA(1,1) Model

Similar to the results for the STAR(1) model discussed above, we find that tests based on critical values constructed by ignoring the first order expansion tend to be overly conservative. Such tests still have power against the alternatives studied; however, power is lower than their expansion based counterparts. The ARMA model also experiences a similar issue with the NOX based tests giving higher rejection frequencies under bilinear errors.

Further, the results for LF tests appear more favorable for the ARMA model than for the STAR model. In particular, sizes appear to be similar to the ICS counterparts when the true identification case is weak ($\beta_n = .3/\sqrt{n}$) or non ($\beta_n = 0$) identification. However, the LF tests appear to be too conservative when the truth is strong identification. Under strong identification and the null hypothesis, sizes are typically considerably smaller than nominal, and the rejection frequencies

are low enough under the AR(2) errors that all LF based tests are in danger of failing to reject the alternative hypothesis. For MA(10) errors, empirical power of the LF based tests are lower than their counterparts, but only the sup LM and CvM tests give negligible power.

ICS based tests tend to perform well across most of our testing specifications. Namely, for iid and GARCH errors, the ICS based tests are smaller than or close to nominal across all identification scenarios, with the MC ICS test giving rejection frequencies closest to the nominal levels. The sup LM and CvM tests are often very conservative, giving very small empirical sizes.

In particular, under the null hypothesis specifications and strong identification, the ICS based tests appear to be equivalent to their S based counter parts. However, under weak and non identification, the ICS based tests continue to provide empirical sizes that are smaller than or close to nominal, while their S based counterparts tend to exhibit empirical size distortions. For example, at the nominal level $\alpha = .10$ for iid errors and weak identification, the MC ICS statistic has empirical size .112 while the MC S statistic gives an empirical size of .172 at lag length $\mathcal{L}_n = 5$, and the MC ICS statistic has empirical size of .106 and the MC S statistic has empirical size of .144 at lag length $\mathcal{L}_n = 10$. The CvM test is an exception here, as the CvM S test does not tend to exhibit empirical size distortions; however, the CvM test is very conservative with CvM ICS often exhibiting rejection frequencies far below nominal. This results in lower empirical power for the CvM test as well, with CvM ICS and CvM S exhibiting rejection frequencies of .066 and .07 under the alternative hypothesis given by MA(10) errors and strong identification. The sup LM test has a similar issue with being overly conservative.

Under the alternative hypothesis, the MC tests have the overall highest empirical power out of the tests considered. The LBQ test performs decently as well. With S based tests providing higher rejection frequencies than ICS based counterparts, leading to size distortions under the null when identification is weak, we might expect the rejection frequencies of S based tests to be considerably higher than ICS counterparts under alternative specifications. While S based tests do tend to have higher rejection frequencies than ICS based tests, overall, the empirical power of ICS based tests is comparable to their S based counterparts. For example, under the MA(10) alternative specification with lag length $\mathcal{L}_n = 10$ when identification is weak, the MC ICS test has rejection frequency .904

and the MC S test has rejection frequency .928.

The AR(2) alternative specification appears to be fairly easy to pick up by the tests under consideration, as they exhibit reasonably large empirical power. The MA(10) alternative provides a better distinction among the tests.

Table 2.6: White Noise Test Simulations - ARMA under Strong Identification

	H_0 True		H_0 False			
	Infeasible		Infeasible		Feasible	
	iid	GARCH(1,1)	AR(2)	MA(10)	AR(2)	MA(10)
MC ICS	0.05	0.03	0.56	0.76	0.51	0.72
LBQ ICS	0.04	0.04	0.52	0.54	0.47	0.48
sup LM ICS	0.02	0.05	0.50	0.06	0.46	0.05
CvM ICS	0.01	0.02	0.41	0.03	0.00	0.01
MC S	0.05	0.03			0.58	0.76
LBQ S	0.04	0.05			0.54	0.55
sup LM S	0.02	0.05			0.52	0.06
CvM S	0.02	0.03			0.41	0.03

Rejection Frequencies: Robust Tests, Identified, ARMA model, $\alpha = 0.05$, $T = 100$, $\beta = 0.300$, $J = 500$.

Table 2.7: White Noise Test Simulations - ARMA model under Weak Identification

	H_0 True		H_0 False			
	Infeasible		Infeasible		Feasible	
	iid	GARCH(1,1)	AR(2)	MA(10)	AR(2)	MA(10)
	$\mathcal{L}_n = 5$	$\mathcal{L}_n = 5$	$\mathcal{L}_n = 5$	$\mathcal{L}_n = 10$	$\mathcal{L}_n = 5$	$\mathcal{L}_n = 10$
MC ICS	0.05	0.05	0.89	0.81	0.52	0.15
LBQ ICS	0.03	0.04	0.85	0.63	0.47	0.11
sup LM ICS	0.01	0.04	0.85	0.04	0.45	0.01
CvM ICS	0.00	0.01	0.87	0.01	0.01	0.00
MC S	0.09	0.10			0.90	0.85
LBQ S	0.08	0.07			0.88	0.73
sup LM S	0.09	0.09			0.87	0.11
CvM S	0.05	0.05			0.88	0.12

Rejection Frequencies: Robust Tests, Nearly Unidentified, ARMA model, $\alpha = 0.05$, $T = 100$, $\beta = 0.030$, $J = 500$.

2.6 Empirical Analysis

Predictability of stock returns continues to be an active area of research. Campbell, Lo, and MacKinlay (1997) use Box-Pierce tests to analyze serial correlation in stock return indices. Nankervis and Savin (2010, 2012) reproduce this analysis with their sup LM test, a generalized

version of the test of Andrews and Ploberger (1996). However, Nankervis and Savin (2010) note the limitation that their sup LM test is not appropriate for residuals, specifically from an ARMA model. Our test, on the other hand, is designed specifically to be appropriate for residuals from estimated models.

In a related, but different line of research, trading rules are utilized in an attempt to exploit low levels of dependence in financial return series. Some of these rules are based on modeling the return series. Taylor (2005) describes a trading rule based on modeling the dependence in financial returns with an ARMA(1,1) model and informing the trading decision based on forecasts generated from that model.

We perform an exercise similar in spirit to that performed by Campbell et al. (1997) and Nankervis and Savin (2010) in that we are interested in testing for serial correlation in financial return series. However, we motivate our exercise by modeling the return series with an ARMA(1,1) model and using our robust test for serial correlation as a model adequacy test on the residuals from the model.

A more appropriate model for this context would be an ARMA-GARCH model; however, the notion of market efficiency should lead us to believe that the model is over-parameterized in the sense that the ARMA parameters are zero. Since GARCH residuals are MDS, we believe that adding a GARCH component to the model unnecessarily complicates the exposition of this test. Because of this, we simply this analysis which would other-wise require a more realistic model in order to demonstrate the use of the test.

Our data are annual, monthly, daily value-weighted and equally-weighted CRSP NYSE/AMEX stock return indices for the period July 1962 to December 2005, the same data as used by Nankervis and Savin (2010). Additionally, we perform our analysis on the same sub-sample periods used by Nankervis and Savin (2010).

We compare results from the same tests as illustrated in the simulations. The sup LM S test is the test of Nankervis and Savin (2010), appropriately modified to be used with residuals but ignoring the possibility of weak identification. The sup LM ICS test is the same test but constructed to account for the possibility of weak identification using the Identification Category Selection

procedure.

We model the return series with the ARMA(1,1) model

$$y_t = (\beta + \pi)y_{t-1} + \varepsilon_t - \pi\varepsilon_{t-1},$$

which is estimated via Quasi-Maximum Likelihood. Recall that π is unidentified when $\beta = 0$, and π is weakly identified when β is statistically close to 0. Tables for the monthly data analysis of value-weighted returns are provided here for illustrative purposes. The first table shows the estimated coefficients for β along with the corresponding standard errors.

It appears that the $\hat{\beta}_n$ are all statistically insignificantly different from zero. This situation could ordinarily give the practitioner pause in modeling these series with the model used. One's first reaction may be to assume that the ARMA model is over-parameterized, yielding $y_t = \varepsilon_t$. However, as discussed in the simulation section 2.5, this does not mean that the ARMA(1,1) model is not adequate for the purpose of modeling the return series.

Further, while we expect the notions of market efficiency and no arbitrage to manifest as a common root in the ARMA model, as practitioners we cannot say for certain that there is no dependence in the series. This is exactly the empirical situation in which one must worry about potential weak identification. That is, we want to model the potentially low levels of dependence that may exist in the series, and we must account for the non-standard inference that results from the parameter π being statistically near identification failure.

The goal of this empirical exercise is to perform a model diagnostic activity on the resulting estimated model in order to determine if we are willing to believe that the model has captured any serial correlation that might have existed in the series. Hence, for these sub-sample periods, we conduct the serial correlation tests, appropriately modified for residuals, as outlined in the simulation section.

Table 2.8: White Noise Test Empirical Exercise

White Noise Test Empirical Exercise: Estimated βs

	1	2	3	4	5	6	7
Start Date	31-Jul-1962	31-Jul-1962	31-Oct-1978	31-Jul-1962	31-Jan-1995	31-Oct-1978	29-Jan-1988
End Date	30-Dec-1994	29-Sep-1978	30-Dec-1994	30-Dec-2005	30-Dec-2005	30-Dec-2005	30-Dec-2005
n	389	194	194	521	131	326	215
$\hat{\beta}_n$	0.08	0.10	0.03	0.07	0.09	0.04	0.051
$\hat{se}$	(0.91)	(0.74)	(0.24)	(0.64)	(0.82)	(0.24)	(0.62)

White Noise Test Empirical Exercise: Test Statistics

MC ICS	2.97	1.37	2.47	2.68	1.23	$2.40^{\star\star}$	0.98
LBQ ICS	1.52	-0.51	1.12	1.54	-0.95	$1.81^{\star\star}$	-0.94
sup LM ICS	2.53	1.02	1.07	6.43	2.88	1.83	4.88
MC S	$2.97^{\star\star}$	1.37	$2.47^{\star}$	$2.68^{\star\star}$	1.23	$2.40^{\star\star}$	0.98
LBQ S	$1.52^{\star}$	-0.51	1.12	$1.54^{\star}$	-0.95	$1.81^{\star\star}$	-0.94
sup LM S	2.53	1.02	1.07	$6.43^{\star\star}$	2.88	1.83	$4.88^{\star\star}$

Estimated βs and White Noise Test Statistics for the Empirical Exercise using VWRETD, monthly data, with an ARMA model where the identification is unknown. $\mathcal{L}_n = 5$. The suffix ICS indicates the Robust CV based tests, and S indicates the corresponding non-robust tests. Significance levels: * = .1, ** = .05, *** = .01

Note that we are not, as is done in previous research (Campbell et al., 1997; Nankervis and Savin, 2010), examining serial correlation in the raw return series. Here we model the raw returns with an ARMA(1,1) model and examine the residuals from this estimated model for serial correlation. Recall that the tests denoted with *S* are based on the procedure that ignores identification failure, and the tests denoted *ICS* are based on the procedure that accommodates identification failure. Note also that the table displays the test statistics, which should be, and are, the same across testing procedures; recall that holding a particular test fixed, the testing procedures only differ in the critical values leading to the differences in statistical significance shown in the table.

Observe first the tests that ignore identification failure. In particular, the results from columns 1, 4, and 6 would indicate to the practitioner that there is evidence to conclude that the modeling exercise does not adequately capture serial correlation in the raw return series. However, from the table detailing the estimated $\hat{\beta}$'s and their associated standard errors, the results from this paper should lead the practitioner to believe that our modeling activity may be contaminated by identification failure. Since this batch of tests do not accommodate identification failure, we should be wary of their implications.

Indeed, turning to the batch of tests that accommodate identification failure, we see that, in general, the identification failure itself may be to blame for the rejection of the null hypothesis found with the *S* based tests.

2.7 Conclusion

A long line of literature documents the effect of parameter identification failure on the distributions of the estimators. Here we demonstrate that the resulting nonstandard estimator distribution can propagate to a test statistic for serial correlation conducted on the residuals from the estimated model. Naively using tests for serial correlation that do not accommodate identification failure can lead to distorted inference when identification failure is present as a feature of the modeling activity. In this case, the practitioner is in danger of concluding that the modeling activity did not capture the serial correlation in her data, when in fact, this conclusion was induced by failing to account for the failure in parameter identification. We provide a testing procedure that accommodates parametric identification failure in a white noise test on the residuals from an estimated

model by simulating the resulting nonstandard distribution in order to conduct valid inference.

Our simulations indicate that this testing procedure is able to correct the over rejection of the null hypothesis that occurs when ignoring identification failure in finite samples. We document that while the procedure does involve a loss in empirical power in finite samples, the procedure still appears to be useful for certain types of alternative. Finally, we present an empirical exercise in which we demonstrate why we expect to encounter identification failure, the danger of ignoring this issue, and how this testing procedure accounts for identification failure.

CHAPTER 3

TESTING MANY ZERO RESTRICTIONS UNDER MIXED IDENTIFICATION STRENGTH

3.1 Introduction

Traditional inference is distorted when estimating a model with either a large dimensional parameter or a parameter that may not be identifiable. These issues have both been well studied in isolation; however, there are situations in Economics in which practitioners may wish to estimate and conduct inference on a large dimensional parameter when some of the parameters might not be identified. We provide a testing framework that accommodates large dimensional parameters when some or many of the parameters may only be weakly identified. Specifically, we demonstrate that parameter identification failure results in size distortions in a high dimensional setting, and we develop an inference procedure that corrects the size distortion in tests for large dimensional parameters for which a subset may be weakly identified.

Testing in a high dimensional framework is a challenge for standard tests and a topic of recent interest. Many solutions have been proposed to accommodate estimation and inference in a high dimensional setting; however, we focus our attention on a particular class of tests. This class of tests, referred to as Max tests, conducted on the maximum estimator from a sequence of parsimoniously constructed sub-models have been shown to have nice properties when testing high dimensional parameters (Ghysels et al., 2016a; Hill and Dennis, 2018). These sub-models which we describe in detail in section 3.3 form the foundation of our testing framework.

It is also known that tests on parameters that may not be identifiable require non-standard inference (Andrews and Cheng, 2012a; Cheng, 2015). In particular, many parameters in our framework may or may not be identifiable, requiring the analysis of mixed identification strength in our setting. The effect of mixed identification strength on inference has not yet been studied in a high dimensional setting.

Examples in the empirical literature demonstrate the need to examine the meeting point of these two topics. Some of these examples include the use of additive non-linear models to study non-linear mean reversion in exchange rate dynamics (Taylor, Peel, and Sarno, 2001; Kilic, 2016) and linear instrumental variables models with many weak instruments (Andrews and Stock, 2007; Belloni, Chernozhukov, and Hansen, 2014b), particularly when estimated with limited information maximum likelihood (Andrews and Cheng, 2012b). We discuss examples in greater detail in section 3.6.

We provide initial simulation results in Section 3.3 demonstrating that traditional tests tend to over-reject the null hypothesis that the parameter vector of interest is zero when this parameter has large dimension and some model parameters are only weakly identified. When the parameter of interest has large dimension but parameter identification failure is not present, the Wald test exhibits size distortions, but the Max test does not; this reiterates the result in Hill and Dennis (2018). However, when combined with parameter identification failure, both tests tend to over-reject the null hypothesis. This paper synthesizes ideas in Cheng (2015) and Hill and Dennis (2018) to develop a test on a large dimensional parameter that reduces size distortions when weakly identified parameters may be present in the model.

We consider tests of parameter zero restrictions in models for which the parametric source of identification failure is known. For this class of models, identification failure can occur under the null hypothesis, leading to the inclusion of nuisance parameters under the null, or weakly identified parameters unrelated to the null hypothesis may be present as a feature of the model. In order to account for the presence of unidentified parameters, we operate under the unifying framework of Andrews and Cheng (2012a) and Cheng (2015). Unlike Andrews and Cheng (2012a), our framework accommodates models with mixed identification strength originating from many parametric sources of identification failure. The concept of testing under mixed identification strength is explored by Cheng (2015) for the case of an additive nonlinear model with a small dimensional parameter. Our procedure is applicable to a broader class of models within the framework of M-estimation.

Further, the dimension of the subset of the parameter vector on which we conduct the test is

allowed to increase with the sample size, thereby creating a high dimensional testing framework. This is an avenue that has not been explored in conjunction with weakly identified models. Estimation and inference on parameters with a large dimension must be handled in a non-standard way. That is, traditional methods for conducting inference on high dimensional parameters will result in size distortions (Hill and Dennis, 2018), and if the dimension of the parameter is large enough, traditional estimation methods may fail entirely. Estimation either proceeds with a shrinkage estimator[1] (Tibshirani, 1996) paired with a sparsity assumption with inference conducted only on the non-zero parameters (Belloni, Chernozhukov, and Hansen, 2014a), or estimation and inference proceeds via carefully selected parsimonious models (Ghysels et al., 2016a; Hill and Dennis, 2018). We adopt the latter framework to provide a correctly sized testing procedure for high dimensional parameters in models with mixed identification strength.

We formally introduce the Max Test in Section 3.3. For the interested reader, Section 3.2 discusses the related literature. For inference, we adopt the notation of Cheng (2015).[2] Section 3.4.1 presents the main assumptions and estimation results[3] that provide the joint limit theory for the parsimonious estimators which is then used to inform the limit theory for the Max Test in Section 3.5. Section 3.4.2 discusses the link between the hypothesized model and the parsimonious models that is needed in order to provide a valid test. Section 3.5 discusses the Max Test and the inference procedure used to calculate p-values. We collect several examples that can analyzed using this framework in Section 3.6. Section 3.7 details the Monte-Carlo simulations, and the final section concludes.

3.2 Relationship with the Literature

Consider estimating scalar parameters (β, π) from the nonlinear function $Y_t = \beta g(X_t, \pi) + \varepsilon_t$ for some smooth non-linear function g. It is well known that when $\beta \neq 0$, π can be (strongly)

[1] e.g. the LASSO.

[2] Section B.1 in the Appendix discusses in detail the notation needed to utilize her sequential procedure.

[3] The results in Section 3.4.1 are based upon results derived in Section B.2 in the Appendix which provide the assumptions and estimation results that generalize Cheng's (2015) results to a broader class of models and hence are of independent interest.

identified, and when $\beta = 0$, π cannot be identified. In order to develop a unifying testing framework, we utilize a thought experiment which can be characterized by using the notion of drifting sequences of true parameters. Let $\beta = \beta_n$ be a sequence of true parameters, indexed by the sample size n, that are drifting to 0. Then the strength of identification of π is categorized by the speed at which $\beta_n \to 0$. When $\sqrt{n}\beta_n \to \infty$, we characterize π as being semi-strongly identified, and when $\sqrt{n}\beta_n \to b \in (0, \infty)$, we say π is weakly identified. In the latter case, our estimator $\hat{\pi}_n$ is not consistent for the true π_0, and converges instead to a random variable under certain conditions. Table 1 from Andrews and Cheng (2012a) details these categories. It is important to note that in this literature, the parametric source of identification failure is known. More recently, Han and McCloskey (2016) develop theory for the case in which the source of identification failure may be unknown. We focus on the former case and leave this extension for future research.

For the cases of non-identification and weak identification, the estimators for π are inconsistent. Further, in these cases the estimator for β is consistent; however, it is a function of $\hat{\pi}_n$ which converges to a random variable, resulting in a non-standard distribution for $\hat{\beta}_n$. This implies that the resulting test statistics will exhibit non-standard behavior, yielding distorted inference from classical tests. In this case, the asymptotic distribution of the test statistics will be nonstandard.

In particular, this is an issue for economic practitioners, as many commonly used models in Economics include parameters that may be unidentified in certain parts of the parameter space. Examples such as Dynamic Stochastic General Equilibrium models (Guerron-Quintana et al., 2013; Andrews and Mikusheva, 2015), Smooth Transition AutoRegressive models (Terasvirta, 1994; Teräsvirta, 1998; van Dijk et al., 2002; Andrews and Cheng, 2013), Probit models (Andrews and Cheng, 2012a, 2014) and Nonlinear Binary Choice Models (Andrews and Cheng, 2013), nonlinear instrumental variables models with possibly weak instruments (Andrews and Cheng, 2012a, 2014), ARMA models Andrews and Ploberger (1996); Andrews and Cheng (2012a); Dennis (2019), Regime Switching Models (Chen et al., 2016) and Fuzzy Regression Discontinuity Designs (Feir et al., 2016), models based on moment conditions and GMM (Andrews and Cheng, 2014), and MiDAS Regressions (Ghysels et al., 2016b) have been shown to include model components that may not be identified in certain regions of the parameter space.

Missing from the analysis of Andrews and Cheng (2012a) is the ability to account for models with mixed identification strength, referring to models which may simultaneously include parameters from each from each of the identification categories (Cheng, 2015). Consider the simple model $Y_t = \beta_1 g(X_t, \pi_1) + \beta_2 g(X_t, \pi_2) + \varepsilon_t$ where ε_t is independent of X_t and with the null hypothesis $H_0 : \beta = 0$. Under this null hypothesis, the π_j are unidentified nuisance parameters, so this framework is related to the literature on testing with nuisance parameters under the null (Davies, 1977, 1987; Andrews and Ploberger, 1994; Hansen, 1996; Stinchcombe and White, 1998; Ghysels and Guay, 2004; Andrews and Mikusheva, 2016). Nuisance parameters cause the test statistics to have non-standard distributions, which often do not have analytic expressions and must be simulated.

In this framework, however, each parameter π_j may exhibit its own degree of identification strength, so a uniformly valid test becomes necessary. Andrews and Cheng (2012a, 2013, 2014) discuss uniformly valid inference but do not allow for mixed identification strength. Cheng (2015) offers the first uniformly valid inference procedure for inference on sub-vectors of β allowing for mixed identification strength but limits her theory to additive nonlinear models. The theory presented here is applicable to M-estimation problems and hence is appropriate for a much larger class of models.

Andrews and Cheng (2012a, 2013, 2014) and Cheng (2015) do not consider large dimensional parameters or max tests. In contrast, we construct a test based on the maximum of a sequence of estimated parameters from a high dimensional parameter. Inference in models with many parameters is typically conducted with an imposed sparsity assumption by forcing a large number of the parameters to be equal to zero with a penalized estimator such as LASSO or Ridge (Tibshirani, 1996) in a way that precludes inference on those parameters. As a result, valid inference can only be conducted on the remaining non-zero parameters.

Further, the LASSO sets exactly equal to zero any parameter that cannot be statistically distinguished from zero. Belloni, Chernozhukov, Hansen, and Kozbur (2016), Leeb and Pötscher (2008) and Pötscher (2009) note that this can be problematic for conducting inference with approximately sparse models that include both variables with small but nonzero coefficients and strong predictors, as the LASSO will exclude the variables with small but nonzero coefficients, which can lead

to omitted variable bias and irregular sampling behavior. Recent work focusing on this inference issue has relied on 'desparsification' (van de Geer, Bühlmann, Ritov, and Dezeure, 2014; Caner and Kock, 2018) or 'debiasing' (Belloni et al., 2014b; Wooldridge and Zhu, ming) the LASSO estimator; however, using these procedures to conduct inference when some parameters are weakly identified has not been studied.

Our approach differs in that we estimate a collection of parsimonious models by considering each parameter in turn and evaluating the maximum of the estimated values, thereby allowing inference on all parameters Ghysels et al. (2016a); Hill and Dennis (2018). In this sense, our test can be thought of as a *pre*-test on a high dimensional parameter vector for variable inclusion.[4] Alternatively, one may prefer to frame the max test as method for testing the sparsity assumption on a given model. In general, however, we may simply have a desire to test a large subset of our parameters based on economic reasoning. Further, our procedure does not require sub-Gaussian related moment conditions, typically cited for use of the LASSO; however, this results in the trade-off that the rate of allowable parameter inclusion be limited to $o(n)$ rather than the much less restrictive rates typically allowed by the LASSO and related estimators.

When testing the maximum value in a sequence, we are most often interested in determining if any of the parameter elements are different from zero. In considering only the maximum from the sequence of values, the test statistic utilizes the most informative measure available from our data, eliminating issues that arise from low degrees of freedom and inversion of large or near singular covariance matrices when a large number of variables needs to be tested, for example (Hill and Dennis, 2018; Ghysels et al., 2016a), or by combining noisy estimates which occurs when calculating serial correlations at long lags (Hill and Motegi, 2018; Dennis, 2019).

Statistics based on a maximum of a sequence of values is an extensively studied topic in the literature (see the textbook treatments by Leadbetter et al. (1983); Resnick (1987)) dating at least to Fisher and Tippet (1928) and Gnedenko (1943). See also Gumbel (1958) and Berman (1964).

[4] As noted in Antoine and Renault (2015), pre-testing approaches are sometimes criticized, as a correct approach to inference should account for error induced from the pre-testing stage, an argument similar to that posed by Leeb and Pötscher (2008) regarding post-selection inference with the LASSO. The pre-testing approach is only one interpretation of the max test.

Typically in this literature, extremal value theory arguments appeal to the Extremal Types Theorem to determine the exact asymptotic distribution of the maximum statistic. For example, Xiao and Wu (2014), who provide a max test for serial correlation, show that under suitable normalization, their test statistic converges in distribution to a Gumbel distribution. See de Haan (1976), who provides a standard approach to proving the Extremal Types Theorem.

These arguments require that when the data are divided into blocks, the dependence between increasingly distant blocks decays at a sufficient rate. Hill and Dennis (2018) argue that when estimating parsimonious models, the classical extreme value theory arguments are no longer straight forward to prove, as the estimators from the parsimonious models may exhibit some degree of asymptotic dependence due to omitted variables. For this reason, we rely on a different method proved in Hill and Dennis (2018) to prove the validity of our bootstrap.

Methods for bootstrapping high dimensional statistics have not been available until recently. Chernozhukov et al. (2013, 2017) develop a theory that is able to both bypass the typical extreme value theoretic asymptotic arguments and deliver an impressive growth rate for the sequence being examined. However, they require independence and their theory is only appropriate for observed random variables and relies on Gaussian approximation that is not appropriate for approximations of non-Gaussian normalized summands.[5] Zhang and Cheng (2018) extend the Gaussian approximation theory in Chernozhukov et al. (2013, 2017) to allow for dependence, but only allow for observed random variables. Zhang and Wu (2017) develop theory for a Gaussian approximation for high dimensional times series but only allow for observed sequences as well. The theory in Hill and Dennis (2018) is also able to bypass extreme value theoretic arguments, allows for dependence under the null, and is appropriate for residuals. For this reason, we rely on the theory developed in Hill and Dennis (2018); however, this theory results in the trade-off that a precise upper bound on the sequence $\mathcal{L}_n \to \infty$ cannot be provided, and the allowed growth rate is simply shown to be $o(n)$.[6]

[5] see also Belloni et al. (2018).

[6] Hill and Motegi (2018) address the issue of optimal lag selection with a data driven procedure, modified from the method of Escanciano and Lobato (2009). This procedure could be applied to the testing framework presented here;

We discuss several relevant examples in section 3.6 including tests for omitted nonlinearity, relevant to studying Purchasing Power Parity (Rogoff, 1996; Taylor et al., 2001), and linear instrumental variables estimation with many instruments (Belloni et al., 2014a,b, 2016). In an empirical example, we conduct tests of linearity against an additive nonlinear alternative. In particular, the class of Smooth Transition Auto-Regressive (STAR) models that we examine in the simulations has been used to model business cycle asymmetry with regimes associated with recession and expansions (Teräsvirta and Anderson, 1992; Skalin and Teräsvirta, 2001), and nonlinear mean reversion in exchange rate dynamics (Taylor et al., 2001; Kilic, 2016). In general, smooth transition models have been used to study many phenomena. Further, certain STAR models nest many Threshold Autoregressive (TAR) models as a special case. See Hansen (2011) for a survey of the history of TAR models in Economics, and for a more detailed account of the impact of STR models on Economics and Finance, see the review by van Dijk et al. (2002). We now formally introduce the max test.

3.3 The Max Test

First, to fix notation, let θ be the model parameter, and denote the criterion function by $Q_n(\theta)$ where n is the sample size. We are interested in testing the null hypothesis that a subvector λ of θ is the zero vector:

$$H_0 : \lambda_0 = 0_k \qquad \text{vs.} \qquad H_A : \lambda_{i,0} \neq 0 \text{ for some } i.$$

where the dimension of λ, k, is potentially large. Observe that the null hypothesis is true if and only the largest element of λ in absolute value is zero. For this reason, we base our test on the statistic $\max_i |\hat{\lambda}_i|$.

The large dimension of λ poses a problem for both estimation and inference procedures. The max test is constructed by estimating each parameter of interest separately in smaller dimension models, which we call parsimonious or sub-models, as each of the smaller dimension models contains only a single element from the parameter of interest. Whereas the full model may contain

however, this is beyond the scope of this paper, as we seek to illustrate the effect of weak identification on the test.

a parameter of large dimension, each parsimonious model contains only a small dimensional parameter, resulting in the need to estimate a large number of parsimonious models and combine the results in a creative way.

This solves the estimation problem faced with a large dimensional parameter by reducing the dimensionality in the estimation step; however, this induces the trade-off in the form of the need to combine estimates from a large number of models into a meaningful statistic. In this manner, the max test is able to accommodate inference on parameters of large dimension by reducing the estimation step to many finite dimensional models. Due to this, the max test rests upon the foundation of these carefully constructed parsimonious models, which we describe in detail in this section.

To distinguish the sub-models from the full model, let $\theta_{(i)} \in \Theta_{(i)}$ denote the vector of parameters and $Q_{(i),n}(\theta_{(i)})$ denote the criterion function for the ith parsimonious model. We will be more explicit about the construction of these parsimonious models in a moment, but for now note that the construction of these parsimonious models results in k different sub-models that must be estimated, each of which contains only a single element λ_i from the parameter of interest, λ. We estimate each parsimonious model with

$$\hat{\theta}_{(i)} = \underset{\theta_{(i)} \in \Theta_{(i)}}{\operatorname{argmin}} \; Q_{(i),n}(\theta_{(i)}).$$

By construction of these sub-models, each parsimonious model i contains only a single element λ_i from the vector λ. We collect the relevant $\hat{\lambda}_{(i)}$'s from the estimation of each of the $i = 1, \ldots, k$ parsimonious models and form the test statistic

$$\hat{\mathcal{T}}_n = \max_{1 \leq i \leq k_n} \left| \mathcal{N}_{(i),\lambda,n} \mathcal{W}_{(i),n} \hat{\lambda}_{(i)} \right|$$

where $\mathcal{N}_{(i),\lambda,n}$ gives the appropriate standardization as described in section 3.4.1, $\mathcal{W}_{(i),n}$ is a weighting term that we assume is uniformly consistent for some constant $W_{(i)}$, and $k_n \to \mathring{k} \geq d_\lambda$ where $\mathring{k}$ is allowed to be ∞. Since each $\hat{\lambda}_{(i)}$ may be scaled differently, the $\mathcal{W}_{(i),n}$ can provide the appropriate

scaling; we will use $\mathcal{W}_{(i),n}$ as the inverse of the standard error.

Max tests are natural choices for high dimensional statistics in part because they do not require inverting a potentially large or near singular covariance matrix, and they use the most informative sample estimate from the sequence, ignoring the others as though they are zero. This method is particularly adept at picking out non-zero values from a large dimensional object, even when only a small number of the parameter values being tested are different from zero.

There are always trade-offs, and ignoring all but the most informative sample estimate can result in information loss. In particular, we expect a test based on the maximum value to underperform relative to a test based on combining many values, such as a Wald or portmanteau, for alternatives that involve a large number of small valued parameters as with a test for correlation in a model with a weak but flat correlation structure. The practitioner should be cognizant of these trade-offs when testing his or her models. For a discussion of this issue, we refer the reader to Hansen (2005).

3.3.1 Parsimonious Models

The many smaller dimension parsimonious models form the foundation of the max test, and we describe here the notation used in the construction and estimation of these parsimonious models. We partition our parameter vector $\theta = (\delta', \lambda', \tilde{\delta}')'$ where we are interested in the parameter λ; specifically, recall that we test the hypothesis $H_0 : \lambda_0 = 0_{d_\lambda}$ where the dimension of λ, d_λ, is potentially large. The parameter δ is a vector of nuisance parameters that are present in every parsimonious model, and $\tilde{\delta}$ is an additional vector of nuisance parameters that are tied to λ and are described later.

Since the null hypothesis is true if and only if $\lambda_{i,0} = 0$ for every i, the max test operates by estimating the restricted parameter $\theta_{(i)} = (\delta', \lambda_i', \tilde{\delta}_i')'$ from the ith parsimoniously constructed model with loss function $Q_{(i)}(\theta_{(i)}) = Q([\theta]_{(i)})$ where $[\theta]_{(i)} = (\delta', 0, \ldots, 0, \lambda_i, 0, \ldots, 0, \tilde{\lambda}_i, 0, \ldots, 0)'$. That is, the ith parsimonious model is constructed by restricting all elements $\lambda_j = 0$ for $j \neq i$ and $i = 1, \ldots, \mathring{k}$ where $\mathring{k} \geq d_\lambda$.

The associated $\tilde{\delta}_j$ elements are also set equal to zero for convenience; however, this is not critical, as the model does not depend upon $\tilde{\delta}_j$ when $\lambda_j = 0$. In this fashion, $\tilde{\delta}$ are nuisance

parameters that appear only when the relevant element from λ is not zero.[7] Hence, we simply parameterize the parsimonious models in a way that allows an extra nuisance parameter to appear when the parsimonious model calls for it.

Observe that the parameter δ is present in every parsimonious model i for every $i = 1, \ldots, \mathring{k}$; however, the estimates may differ between parsimonious models. For this reason, it is important to keep track of the estimators from each parsimonious model. We use the subscript (i) to indicate that the parameter estimators, estimates, and (pseudo-)true values belong to parsimonious model i.[8]

Since the parsimonious models are constructed by omitting variables that are not relevant when the null hypothesis is true, there must be at least one parsimonious model that is missing a relevant variable under the alternative. That is, the estimator from the ith parsimonious model minimizes the parsimonious loss function, which by definition is a restricted version of the true loss function and may omit relevant variables when the alternative hypothesis is true; hence we cannot say in general that the ith parsimonious estimator can consistently estimate the true parameter. Instead, we say that the ith parsimonious estimator $\hat{\theta}_{(i)} = (\hat{\delta}'_{(i)}, \hat{\lambda}'_{(i)}, \hat{\tilde{\delta}}'_{(i)})'$ estimates $\theta_{(i),n}$ which may not be equal to the relevant components of the restricted parameter $[\theta_n]_{(i)}$ with $\lambda_j = 0$ and $\tilde{\delta}_j = 0$ for every $j \neq i$. This is a well known issue (see e.g. White (1981)), and we establish conditions that provide a valid test.

3.3.2 Max Test Framework

We demonstrate a simplified exposition of the max test with the model

$$Y_t = \beta_1 g(X_t, \pi_1) + \beta_2 g(X_t, \pi_2) + \cdots + \beta_p g(X_t, \pi_p) + \varepsilon_t.$$

where $E[\varepsilon_t | X_t] = 0$ and for some non-linear function g that is a non-degenerate random variable $g(X_t, \pi_i)$ for every π_i. When p is restricted to be small, this is the same model analyzed by Cheng

[7] That is, we could parameterize the parsimonious models so that all nuisance parameters are included in $\delta_{(i)}$; however, this would involve changing the size of the parameter $\delta_{(i)}$ with every i. While this may add clarity in one dimension, we believe that it detracts from clarity in another.

[8] Here we mean pseudo-true in the sense of White (1981).

(2015).

Here, we may want to test the null hypothesis $H_0 : \lambda = 0_k$ where λ is a sub-vector of β, and $k \leq p$ is potentially large. Two issues are evident in this simple model. First, λ has a large dimension. This is known to result in problematic estimation and inference. Second, the sub-vector of π corresponding to λ is not identified under the null hypothesis. For example, when $\lambda_1 = \beta_1 = 0$, π_1 is not identified.

Table 3.1: Max Test Initial Simulations

Model	Linear	Non-linear	Linear	Non-linear
k_λ	1	1	20	20
Wald Test	0.04	0.14	0.22	0.85
Max Test	0.05	0.11	0.05	0.12
Max t-Test	0.05	0.11	0.06	0.17

Max Test Initial Simulations - Rejection Frequencies, $J = 1000$, $\alpha = 0.05$, $n = 200$, $k_\lambda \in \{1, 20\}$

The table above details rejection frequencies for the standard Wald and Max tests under a few model specifications for the nominal size $\alpha = .05$. The null hypothesis is $H_0 : \lambda = 0_k$ where k is the dimension of the parameter being tested. The linear model, presented for reference, takes $g_i(X_t, \pi_i) \equiv X_{i,t}$ and hence does not contain weakly identified parameters, but the non-linear model does contain parameter identification failure under the null hypothesis. We make note of the following observations. In the first column with low k_λ and no weak identification, we see that neither the Wald nor the Max Test have size distortions. For the second column, a low k_λ is paired with weakly identified parameters to demonstrate that both standard tests have size distortions. Cheng (2015) corrects this issue for a Wald Test on a subset of parameters from the additive non-linear model.

The final two columns demonstrate the influence of a large dimensional parameter on the tests. Observe that the standard Wald test over-rejects the null hypothesis when a large dimensional parameter is present, but the Max Test has size close to nominal; this reiterates the result in Hill and Dennis (2018). However, both tests exhibit size distortions when the model contains a large dimensional parameter and parametric identification failure is present. We solve these two issues

by testing the maximum λ_j by using a sequence of carefully constructed smaller dimension models and using a limiting distribution that is robust to the presence of weak identification.

The max test is built upon the foundation of these carefully constructed smaller dimension models, which we call parsimonious models and which are defined by imposing the null hypothesis for all but one λ_j at a time. For the simplified additive nonlinear model given above when we wish to test the entire vector β,[9] we construct k different parsimonious models:

$$\begin{aligned} Y_t &= \lambda_1 g(X_t, \pi_1) + \nu_{1,t} \\ &\vdots \\ Y_t &= \lambda_k g(X_t, \pi_k) + \nu_{k,t} \end{aligned}$$

where $\nu_{i,t} = \sum_{j \neq i} \lambda_j g(X_t, \pi_j) + \varepsilon_t$. Observe that under the null hypothesis, $\nu_{i,t} = \varepsilon_t$ for every i; hence, $E[\nu_{i,t}|X_t] = 0$ for every i when the null is true. Each parsimonious model in this example is estimated by non-linear least squares with the criterion $Q_{i,n}(\theta_{(i)}) = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \lambda_i g(X_t, \pi_i))^2$. The max test statistic is then constructed by collecting the $\hat{\lambda}_j$, and calculating the appropriately standardized maximum value

$$\hat{\mathcal{T}}_n = \max_{1 \leq i \leq k} |\mathcal{N}_i \hat{\lambda}_i|$$

where $\mathcal{N}_i = \sqrt{n}$ in this example.

Whereas we give the example above to demonstrate the features and construction of the max test, one of the strengths of our test is its ability to accommodate a broad class of models. In its general form, the max test is appropriate for models estimated with M-estimators with criterion of the form $Q_n = \frac{1}{n}\sum_{t=1}^{n} m_t(\theta)$ which includes nonlinear least squares and maximum likelihood. The parsimonious models are then defined by restrictions on the criterion function

[9] That is, we let $\lambda = \beta$ and $k = p$ in this simplified exhibition.

$Q_{(i),n} = \frac{1}{n}\sum_{t=1}^{n} m_{(i),t}(\theta_{(i)})$ where

$$m_{(i),t}(\theta_{(i)}) = m_t\big((\delta, 0, \ldots, \lambda_i, 0, \ldots, \tilde{\delta}_i)\big).$$

Next, we briefly introduce a topic from the empirical literature and describe how this topic fits within the max text framework. We discuss this example and others in greater detail in section 3.6.

3.3.3 Empirical Example

Studies of the effect of transaction costs on Purchasing Power Parity (PPP) suggest that exchange rate adjustments resemble a unit root process within a band and a stationary process outside of that band (Taylor et al., 2001; Obstfeld and Taylor, 1997). Taylor et al. (2001) allow a smooth transition at the boundary of the band with the Smooth Transition Auto-Regressive (STAR) model:

$$q_t = \sum_{j=1}^{p} \beta_j q_{t-j} + \sum_{j=1}^{p} \beta_j^* q_{t-j} h(\gamma; q_{t-d}) + \varepsilon_t$$

where h is the exponential transition function

$$h(\gamma; q_{t-d}) = 1 - \exp(-\gamma(q_{t-d})^2)$$

and q_t is the demeaned log real exchange rate. Similarly, Kilic (2016) examines the first differenced model

$$\Delta q_t = \Big[\beta_0^* + \sum_{j=1}^{p} \beta_j^* \Delta q_{t-j}\Big] h(\gamma_d, \Delta q_{t-d}) + u_t.$$

Two issues are illustrated with these models. First, the unknown value of d must be selected from a potentially large number of available lags. Second, parameter identification failure occurs when $\gamma_d = 0$.

The first issue seems to imply that the assumed model is simply a restriction that many $\gamma_j = 0$

in a larger model that takes the form of the model above summed over many d.

$$\Delta q_t = \sum_{d=1}^{k} \left(\left[\beta_{d,0} + \sum_{j=1}^{p} \beta_{d,j} \Delta q_{t-j} \right] h(\gamma_d, \Delta q_{t-d}) \right) + \varepsilon_t$$

where we can relax the model above by allowing β_d to represent a potentially different vector of parameters for each d. A common null hypothesis is that of no (omitted) nonlinearity:

$$H_0 : \lambda = 0_k$$

where λ is a sub-vector of $\gamma = (\gamma_1, \ldots, \gamma_d, \ldots)$.

Recall the parsimonious models form the foundation of the max test. Conveniently, the parsimonious models are already given as they are implicitly utilized within the literature. That is, for the model studied by Kilic (2016), the parsimonious models are simply given by

$$\Delta q_t = \left[\beta^*_{d,0} + \sum_{j=1}^{p} \beta^*_{d,j} \Delta q_{t-j} \right] h(\gamma_d, \Delta q_{t-d}) + u_{d,t}$$

for each $d = 1, \ldots, k$. Each of these models is estimated, the $\hat{\lambda}_i$'s are collected, and the max test statistic may be calculated as $\hat{\mathcal{T}} = \max_{1 \leq i \leq k} |\mathcal{N}_i \hat{\lambda}_i|$.

We expand this example and discuss additional examples in section 3.6. Next, we begin developing the theory for the max test statistic. First, we develop the limit theory for the parsimonious estimators, and we discuss the linkage between the parameters under the null hypothesis that is necessary to ensure that the limit theory for the max test will follow.

3.4 Assumptions and Preliminary Results

3.4.1 Limit Theory for Parsimonious Estimators

Recall that the max test is constructed as the maximum of an estimated parameter vector where each element from the parameter vector is estimated individually from a parsimoniously constructed model consisting only of that element and any nuisance parameters. This implies that if there are $\mathring{k}$ elements in the parameter vector being tested, then there are $\mathring{k}$ different parsimonious

models, each omitting $\mathring{k} - 1$ elements from the parameter vector.

Section 3.4.2 discusses the assumption needed to ensure the test is valid and consistent. This section details the joint limit theory for the parsimonious estimators that forms the basis for inference using the max test statistic. The limit theory for the parsimonious estimators follows from limit theory developed for M-estimation of models with mixed identification strength in sections B.2 and B.3 in the Appendix. Cheng (2015) develops similar theory for the special case of additive nonlinear models, and she requires correct model specification. The theory developed here is suitable for a broader class of models and assumes correct model specification only under the null hypothesis, a necessary requirement for the max test which is constructed from misspecified parsimonious models.

We first describe the data and the true parameter space. We require strongly mixing stationary sequences and compact parameter spaces that include the point of identification failure in the interior.

Assumption 14. *The observations* $\{W_t = (Y_t', X_t', Z_t')' : t \leq n\}$ *are strictly stationary as are* $\{\varepsilon_t\}$. $\{W_t\}$ *is strongly mixing with mixing coefficient* $\alpha(j)$ *such that* $\sum_{j=1}^{\infty} \alpha(j)^{\delta/(2+\delta)} < \infty$ *for some* $\delta > 0$.

For each parsimonious model i, the estimator $\hat{\theta}_{(i),n}$ minimizes the criterion function $Q_{(i),n}(\theta) \equiv Q_{(i),n}(\theta_{(i)}; W_t) = \frac{1}{n}\sum_{t=1}^{n} m_{(i),t}(\theta_{(i)}; W_t)$ over $\theta_{(i)} \in \Theta_{(i)} = \Delta \times \Lambda_i \times \tilde{\Lambda}_i$ where Λ_i collects the parameter spaces from Θ that are exclusive to model i and included in the null hypothesis, $\tilde{\Lambda}_i$ collects the parameter spaces that are exclusive to model i but not included in the null hypothesis, and Δ collects the parameter spaces for the parameters that are estimated in every model.

Additionally, for the true parameter space Θ^* and the optimization parameter space Θ, a parameter vector $\theta \in \Theta^*$ can be partitioned into three subvectors $\theta = (\beta', \zeta', \pi')'$ where the parameters β and ζ are always strongly identified, and the identification strength of π is determined by β. ζ does not affect the identification of π or β. For the observations $\{W_t = (Y_t', X_t', Z_t')' : t \leq n\}$, $\{Z_t\}$ are the variables associated with parameter ζ which are not associated with β or π. The variables X_t are associated with β and π but not with ζ. For any $\theta \in \Theta^*$, we denote by F_γ the distribution

of $\{W_t : t \leq n\}$ and E_γ its expectation, where $\gamma = (\theta, \phi) \in \Gamma$ and $\phi \in \Phi^*$ is a possibly infinite dimensional nuisance parameter such that the distribution is fully characterized by γ.

Hence, the joint estimator parameter space $\Theta = \mathcal{B}_1 \times \cdots \times \mathcal{B}_p \times \mathcal{Z} \times \Pi$. We require $\mathcal{B}_j, \mathcal{Z}, \Pi$ to be compact for every j and the true parameter space Θ^* to be contained in the interior of Θ. For example, consider the additive nonlinear regression model (Cheng, 2015)

$$Y_t = \zeta' Z_t + \sum_{j=1}^{p} \beta_j g(X_t, \pi_j) + \varepsilon_t$$

with the null hypothesis $H_0 : \beta_j = 0 \; \forall j$. Then $\Delta = \mathcal{Z}$, $\Lambda_i = \mathcal{B}_i$, and $\tilde{\Lambda}_i = \Pi_i$. If the null hypothesis is $H_0 : \zeta_j = 0 \; \forall j$. Then $\Delta = \mathcal{B} \times \Pi$, $\Lambda_i = \mathcal{Z}_i$, and $\tilde{\Lambda}_i = \emptyset$.

Assumption 15. *The true value θ^* belongs to the set $\Theta^* = \mathcal{B}_1^* \times \cdots \times \mathcal{B}_p^* \times \mathcal{Z}^* \times \Pi^*$ where $\mathcal{B}_j^*$ is compact and includes 0 for each j. Π^* and $\mathcal{Z}^*$ are compact. For any $\theta \in \Theta^*$, the distribution of $\{W_t\}$ is given by F_γ, where $\gamma = (\theta', \phi')' \in \Gamma$, and $\phi \in \Phi^*$ is an possibly infinite dimensional nuisance parameter that fully characterizes the distribution. Φ^* is a compact metric space with a metric that induces weak convergence on bivariate distributions (W_t, W_{t+m}) for every $t, m \geq 1$.*

The theory developed in Andrews and Cheng (2012a) and subsequent papers utilizes a $\beta \in \mathcal{B}$ and $\pi \in \Pi$ such that individual elements of β do not affect the identification of individual elements of π. Their theory requires a single drifting rate for the parameter β. Cheng (2015) extends this theory to allow individual components of β to affect mutually exclusive components in π; however, her framework is restricted to the class of additive non-linear models.

Assumption 16. *For every $\mathcal{B}_j$ there is a $\Pi_j = \otimes_{i=1}^{q_j} \Pi_i$ such that $m_t(\theta; w)$ does not depend upon $\pi_j \in \Pi_j$ iff $\beta_j = 0$. β_i for $i \neq j$ does not affect the identification of π_j. ζ does not affect the identification of β or π, and the identification of ζ is not affected by β or π.*

This assumption requires that for every element π_j of π, the identification status of π_j is determined by one and only one element β_j from β. This allows us to define the identification problem precisely and build a uniform estimation theory around it. Clearly the additive nonlinear model of Cheng (2015) satisfies this assumption. Let L be the lag operator, and consider also the

ARMA(2,2) model

$$(1-(\beta_1+\pi_1)L)(1-(\beta_2+\pi_2)L)Y_t = (1-\pi_1 L)(1-\pi_2 L)\varepsilon_t$$

written in this form to show that when $\beta_i = 0$ for some $i = 1, 2$ then a common root is present. Specifically, one can see that whether $\beta_i = 0$ or not affects only the identification status of π_i and does not affect the identification status of any π_j for $j \neq i$.

To allow for uniformity over $\gamma \in \Gamma$, all true parameters are indexed by the sample size n. That is, the true $\gamma_n = (\theta_n', \phi_n')'$ where $\theta_n = (\beta_n', \zeta_n', \pi_n')'$ with $\beta_n = (\beta_{1,n}', \ldots, \beta_{p,n}')'$ and $\pi_n = (\pi_{1,n}', \ldots, \pi_{p,n}')'$. These parameters drift to the limiting values $\theta_n \to \theta_0 = (\beta_0', \zeta_0', \pi_0')' \in \Theta^*$ and $\gamma_n \to \gamma_0 \in \Gamma$.

Assumption 3 requires that the identification strength of π_j, $i = 1, \ldots, p$, be determined by the rate at which $||\beta_{j,n}||$ converges to 0 as $n \to \infty$, with π_j being strongly identified only if $\beta_{j,n} \to \beta_{j,0} \neq 0$. In the case that $\beta_{j,0} = 0$, the speed at which $\beta_{j,n} \to \beta_{j,0} = 0$ affects the asymptotic analysis. In particular, when $||\beta_{j,n}|| \to 0$ fast enough, given by case (i) below, we say the parameter $\pi_{j,0}$ is *weakly* identified. In this case, the estimator $\hat{\pi}_{j,n}$ is not consistent. Hence, following Cheng (2015), we divide the space of drifting sequences into three identification categories of π_j:

(i) Weak Identification: $\beta_{j,n} \to 0$ with $n^{1/2}\beta_{j,n} \to b_j \in \mathbb{R}^{d_{\beta_j}}$

(ii) Semi-Strong Identification: $\beta_{j,n} \to 0$ with $n^{1/2}||\beta_{j,n}|| \to \infty$

(iii) Strong Identification: $\beta_{j,n} \to \beta_j \neq 0$.

Observe that the case $\beta_{j,n} = 0 \ \forall n$ is allowed under case (i); hence this case includes non-identification. The category (ii) of semi-strong identification is necessary for uniform results in Cheng's (2015) work.

Following these categorical definitions, she groups subvectors of π by the identification category above and the rate of convergence to zero for subvectors in the semi-strong identification category. This grouping allows a convenient inductive argument to be used to prove estimation results.

(i) All $||\beta_{j,n}||$ that have non-zero limit are put in the first group. If all $||\beta_{j,n}||$ have zero limits, the first group is empty.

(ii) All $||\beta_{j,n}||$ that are $O(n^{-1/2})$ are put in the last group.

(iii) For those that converge to 0 but at a rate slower than $n^{-1/2}$, members in group k converge to 0 slower than members in group k' for any $k' > k$ and members in the same group converge to 0 at the same rate.

The first group is associated with strong identification, the last group is associated with weak identification, and the middle groups are associated with semi-strong identification, ordered by the rate of convergence. Note that the group index k is a property associated with the drifting sequence $\{\beta_{j,n} : n \geq 1\}$. Therefore the group index k does not change with the sample size n. See Cheng (2015) for details.

Next, suppose there are K groups and $\beta_{k_1}, \ldots, \beta_{k_{p_k}}$ are the elements in group k. Let $l_k = \{k_1, \ldots, k_{p_k}\}$ denote the indices for group k. Use the subscript l_k to denote a sub-vector associated with group k:

$$\beta_{l_k} = (\beta'_{k_1}, \ldots, \beta'_{k_{p_k}})' \in \mathbb{R}^{d_k}$$
$$\text{and} \quad \pi_{l_k} = (\pi'_{k_1}, \ldots, \pi'_{k_{p_k}})' \in \mathbb{R}^{d_{\pi_{l_k}}}.$$

$\beta_{l_k,n}$ denotes the true value of β_{l_k} when the sample size is n and $\beta_{l_k,0}$ denotes its limit. In particular, the grouping rule implies that $||\beta_{l_{k'},n}|| = o(||\beta_{l_k,n}||)$ for $k' > k$ between groups and $||\beta_{j',n}||$ converges at the same rate as $||\beta_{j,n}||$ for any $j, j' \in l_k$ and $k = 1, \ldots, K-1$. In the presence of weak identification, $\beta_{l_k,n} = O(n^{-1/2})$ for $k = K$. If all regressors are in the semi-strong or strong identification category, then we denote $l_K = \emptyset$.

Finally, we describe one more partition of the vectors β and π based on the grouping notation above that will be used to sequentially analyze the limiting behavior of the estimators.

Consider $\pi_{(i),l_k}$, and denote $\pi_{(i),k^-}$ as the elements of π in the previous groups $l_1, \ldots, l_{k-1}$ and

$\pi_{(i),k^+}$ as the elements of π in the subsequent groups $l_{k+1}, \ldots, l_K$.

$$\pi_{k^-} = (\pi'_{l_1}, \ldots, \pi'_{l_{k-1}})' \quad \text{and} \quad \pi_{k^+} = (\pi'_{l_{k+1}}, \ldots, \pi'_{l_K})'.$$

Observe that $\pi = (\pi'_{k^-}, \pi'_{l_k}, \pi'_{k^+})'$, and that the identification strength of these subvectors are in decreasing order by definition. The same notation will apply to β, where we can note that the subvectors in $\beta = (\beta'_{k^-}, \beta'_{l_k}, \beta'_{k^+})'$ have smaller magnitude by definition.

It is important to note that π_{l_1} is strongly identified. All strongly identified elements of π are included in this group in order to analyze them together with the strongly identified parameters β and ζ. The semi-strongly identified and weakly-identified elements of π are analyzed using the sequential procedure outlined in Cheng (2015). If no elements of π are strongly identified, $l_1 = \emptyset$ and π_{l_1} disappears.

We require that each parsimonious model estimation identify a pseudo-true value of the parameter θ in the sense of White (1981). Denote E_{γ_0} as expectation taken under true parameter γ_0.

Assumption 17. *1. if $l_K = \emptyset$, then $E_{\gamma_0}(m_t(\theta_{(i)}; W_t))$ is minimized uniquely by $\theta_{(i)} = \theta_{(i),0} \in \Theta^*_{(i)}$.*

2. if $l_K \neq \emptyset$, then $E_{\gamma_0}(m_t(\psi_{(i),K^-}, \pi_{(i),K}; W_t))$ is minimized uniquely by $\psi_{(i),K^-} = \psi_{(i),K^-,0} \in \Psi^*_{(i),K^-}$ for every $\pi_{(i),K} \in \Pi_{(i),K}$.

Note that $\theta_{(i),0}$ and $\psi_{(i),K^-,0}$ are not, in general, equal to their true model counterparts θ_0 and $\psi_{K^-,0}$. Further, this implies that the pseudo-true drifting sequences $\theta_{(i),n}$ and $\psi_{(i),K^-,n}$ need not be equal to their true model counter parts either. Together with the identification link discussed in section 3.4.2, we are able to establish that $\lambda^*_i = 0$ for every i if and only if the null hypothesis $H_0 : \lambda_0 = 0$ is true, where λ^* is a subvector of $\theta_{(i),0}$ or $\psi_{(i),K^-,0}$ and λ_0 is the corresponding subvector from θ_0 or $\psi_{K^-,0}$.

Additionally, we require some technical moment conditions to establish uniform laws of large numbers, stochastic equicontinuity, and convergence in law of our estimators.

Assumption 18. *For each parsimonious model i, the function $m_{(i),t}(\theta_{(i)};\cdot)$ is measurable with respect to $\sigma(W_t)$, the sigma field generated by $\{W_t\}$, for every $\theta_{(i)} \in \Theta_{(i)}$. Further, $m_{(i),t}(\theta_{(i)})$ is three times continuously differentiable, and for some $\delta > 0$*

1. $\sup_{\theta_{(i)}\in\Theta_{(i)}} E_{\gamma_0}|m_{(i),t}(\theta_{(i)})|^{2+\delta} < \infty$
2. $\sup_{\theta_{(i)}\in\Theta_{(i)}} \lim_{n\to\infty} E_{\gamma_n}|\left[B(\beta_{(i),K^-})^{-1}\nabla_{\psi_{(i),K^-}} m_{(i),t}(\theta_{(i)})\right]_j|^{2+\delta} < \infty$
3. $\sup_{\theta_{(i)}\in\Theta_{(i)}} \lim_{n\to\infty} E_{\gamma_n}|\left[B(\beta_{(i),K^-})^{-1}\left(\nabla^2_{\psi_{(i),K^-}} m_{(i),t}(\theta_{(i)})\right)B(\beta_{(i),K^-})^{-1}\right]_{r,s}|^{2+\delta} < \infty$
4. $\sup_{\theta_{(i)}\in\Theta_{(i)}} \lim_{n\to\infty} E_{\gamma_n}|\left[\frac{\partial}{\partial\psi'_{(i),k^-}} vec\left(B(\beta_{(i),K^-})^{-1}\nabla^2_{\psi_{(i),K^-}} m_{(i),t}(\theta_{(i)})B(\beta_{(i),K^-})^{-1}\right)\right]_{r,s}|^{2+\delta} < \infty$

where $[A]_{i,j}$ denotes the i,jth element of the matrix A.

This moment assumptions are standard when dealing with strongly mixing sequences of random variables.[10] In particular, for the nonlinear additive model and ARMA model, a $2+\delta$th moment on m_t implies slightly more than a 4th moment on $\{W_t, \varepsilon_t\}$. The normalization $B(\beta_{K^-})^{-1}$ is needed due to the mixed rates of convergence of the estimators of the elements of π. Specifically, $\nabla_{\psi_{K^-}} m_t(\theta)$ can often be written in the form $B(\beta_{K^-})\nabla_{\psi_{K^-}}\tilde{m}_t(\theta)$, so this condition can be expressed as a uniform moment condition on $\nabla_{\psi_{K^-}}\tilde{m}_t(\theta)$. For example, in the additive nonlinear regression model with $p = 1$

$$\nabla_{\psi_{K^-}} m_t(\theta) = -\begin{pmatrix} g(X_t,\pi) \\ Z_t \\ \beta\frac{\partial}{\partial\pi}g(X_t,\pi) \end{pmatrix}\varepsilon_t(\theta) = -B(\beta_{K^-})\begin{pmatrix} g(X_t,\pi) \\ Z_t \\ \frac{\partial}{\partial\pi}g(X_t,\pi) \end{pmatrix}\varepsilon_t(\theta).$$

A similar discussion applies to the further derivatives of m_t.

Within each parsimonious model, the consistency and asymptotic distributions of the estimators follow from the theory developed in section B.2 in the appendix. The theory developed in the appendix does not require the models to be correctly specified; instead, it only requires that each

[10] See e.g. Davidson (1994).

misspecified model is uniquely minimized in population by some pseudo-true value. The following lemmas and theorem then follow with only a change of notation.

Lemma 3.4.1 (Consistency for Strong Identification Groups)**.** *Suppose Assumptions 1-5 hold. Then under* $\gamma_n \to \gamma_0$*,*

$$\sup_{\pi^+_{(i),1} \in \Pi^+_{(i),1}} ||\hat{\zeta}_{(i)}(\pi^+_{(i),1}) - \zeta_{(i),n}|| \xrightarrow{p} 0$$

$$\sup_{\pi^+_{(i),1} \in \Pi^+_{(i),1}} ||\hat{\beta}_{(i)}(\pi^+_{(i),1}) - \beta_{(i),n}|| \xrightarrow{p} 0$$

$$\sup_{\pi^+_{(i),1} \in \Pi^+_{(i),1}} ||\hat{\pi}_{(i),l_1}(\pi^+_{(i),1}) - \pi_{(i),l_1,n}|| \xrightarrow{p} 0$$

We require some additional assumptions in order to establish consistency of the estimator for the semi-strong identification groups. Consistency of this estimator is established with an inductive argument that proceeds in the order of decreasing identification strength. Along each step, the more strongly identified estimators are concentrated out and consistency is established uniformly over the subsequent groups by use of a mean value expansion about the point of sequential identification failure. This induces a bias that appears in the limit of the concentrated criterion function and must be accounted for in the proof. The next assumption ensures that the term describing this bias exists and is well behaved.

Assumption 19. *i) For every* i *and every* $k = 1, \ldots, K$*,*

$$\mathcal{K}_{(i),k}(\psi_{(i),k^-}, \pi_{(i),l_k}, \pi_{(i),k^+}; \gamma_0) = \frac{\partial}{\partial \beta'_{(i),k,0}} E_{\gamma_0} \nabla_{\psi_{(i),k^-}} m_t(\theta_{(i)})$$

exists for every $(\theta_{(i)}, \gamma_0) \in \Theta_{(i),\eta} \times \Gamma_0$*, where* $\theta_{(i)} = (\psi_{(i),k^-}, \pi_{(i),l_k}, \pi_{(i),k^+})$*.*

ii) For every i *and each* $k = 1, \ldots, K$*,* $\mathcal{K}_{(i),k}(\theta_{(i)}; \gamma)$ *is continuous at* $(\psi^0_{(i),k^-}, \pi_{(i),l_k}, \pi_{(i),k^+}; \gamma^0)$ *uniformly over* $\pi_{(i),l_k}, \pi_{(i),k^+} \in \Pi_{(i),l_k} \times \Pi_{(i),k^+}$ *for every* $\gamma^0 \in \Gamma$ *such that* $\psi^0_{(i),k^-}$ *is a subvector of* γ^0*.*

This assumption describes how the mean of the score, $E_{\gamma_0}\nabla_{\psi_{(i),k^-}} m_t(\theta_{(i)})$, changes as the true $\beta_{(i),0}$ changes and is similar to Assumption S4 in Andrews and Cheng (2013) and Assumption C5 in Andrews and Cheng (2012a). Cheng (2015) does not state this assumption explicitly; however, she implicitly derives and utilizes the quantity $\mathcal{K}_{(i),k}(\psi_{(i),k^-}, \pi_{(i),l_k}, \pi_{(i),k^+}; \gamma_0)$ in her proof of consistency of her estimator and its limiting distribution.

Finally, we need (sequential) limiting matrix of the second derivative of the criterion function to be non-singular, a standard assumption in the analysis of M-estimation.

Assumption 20. *For each i and k, $\lambda_{min}(H_{(i),k}(\pi_{(i),l_k}, \pi_{(i),k^+}; \gamma_0)) \geq \varepsilon$ for some $\varepsilon > 0$.*

The next lemma provides consistency results for the estimators of the semi-strongly identified parameters. Though the statement of the lemma is similar to that in Cheng (2015), the lemma presented here, and in particular its counterpart in Appendix B.2, is applicable to a broader class of models than her additive nonlinear model.

In order to facilitate the analysis, define the concentrated criterion function

$$Q_n^c(\pi_{l_k}, \pi_{k^+}) = Q_n(\hat{\psi}_{k^-}(\pi_{l_k}, \pi_{k^+}), \pi_{l_k}, \pi_{k^+})$$

where $\psi_{k^-} = (\beta', \zeta', \pi'_{k^-})'$ collects the parameters that have been concentrated out, and the true values of these parameters are denoted with the additional subscripts $\psi_{k^-,n} = (\beta'_n, \zeta'_n, \pi'_{k^-,n})'$ and $\psi_{k^-,0} = (\beta'_0, \zeta'_0, \pi'_{k^-,0})'$ where the latter gives the limit of the drifting sequence: $\psi_{k^-,n} \to \psi_{k^-,0}$. Further, we use the superscript 0 notation to define

$$\psi^0_{k^-,n} = (\beta'_{k^-,n}, \beta^{0\prime}_{l_k}, \beta^{0\prime}_{k^+}, \zeta'_n, \pi'_{k^-,n})'$$

to be the parameter vector consisting of the concentrated out parameters evaluated at the point of sequential identification failure $\beta^0_{l_k} = 0$ and $\beta^0_{k^+} = 0$. See Appendix B.1 for details.

Lemma 3.4.2 (Consistency for Semi-Strong Identification Groups)**.** *Suppose Assumptions 1-7 hold. Then under $\gamma_n \to \gamma_0$, for $k = 2, \ldots, K-1$,*

(a) the concentrated sample criterion function satisfies

$$
\begin{aligned}
&||\beta_{(i),l_k,n}||^{-2}\Big(Q^c_{(i),n}(\pi_{(i),l_k},\pi_{(i),k^+}) - Q_{(i),n}(\psi^0_{(i),k^-,n})\Big) \\
&\xrightarrow{p} -\frac{1}{2}(\omega'_{(i),k,0}, 0'_{d_{(i),k^+}})\mathcal{K}_{(i),k}(\pi_{(i),l_k},\pi_{(i),k^+};\gamma_0)'[H_{(i),k}(\pi_{(i),l_k},\pi_{(i),k^+};\gamma_0)]^{-1} \\
&\qquad\qquad \times \mathcal{K}_{(i),k}(\pi_{(i),l_k},\pi_{(i),k^+};\gamma_0)(\omega'_{(i),k,0}, 0'_{d_{(i),k^+}})', \qquad (3.1)
\end{aligned}
$$

where $\omega_{(i),k,0} = \lim_{n\to\infty} \beta_{(i),l_k,n}/||\beta_{(i),l_k,n}||$ *is the angle parameter*

(b) the estimator of $\pi_{(i),l_k,n}$ *satisfies*

$$
\sup_{\pi_{(i),k^+}\in\Pi_{(i),k^+}} ||\hat{\pi}_{(i),l_k}(\pi_{(i),k^+}) - \pi_{(i),l_k,n}|| \xrightarrow{p} 0
$$

(c) the estimator of $\psi_{(i),k^-} = (\beta'_{(i)}, \zeta'_{(i)}, \pi'_{(i),l_1}, \ldots, \pi'_{(i),l_{k-1}})'$ *satisfies*

$$
||\beta_{(i),l_k,n}||^{-1}\begin{pmatrix}
\hat{\beta}_{(i),k^-}(\pi_{(i),k^+}) - \beta_{(i),k^-,n} \\
\hat{\beta}_{(i),l_k}(\pi_{(i),k^+}) - \beta_{(i),l_k,n} \\
\hat{\beta}_{(i),k^+}(\pi_{(i),k^+}) \\
\hat{\zeta}_{(i)} - \zeta_{(i),n} \\
B^*(\beta_{(i),k^-,n})(\hat{\pi}_{(i),k^-}(\pi_{(i),k^+}) - \pi_{(i),k^-,n})
\end{pmatrix} \xrightarrow{p} 0,
$$

uniformly over $\pi_{(i),k^+} \in \Pi_{(i),k^+}$ *where*

$B^*(\beta_{(i),k^-,n}) = diag\{(1_{d_{\pi_{(i),l_1}}}||\beta_{(i),l_1}||, \ldots, 1_{d_{\pi_{(i),l_{k-1}}}}||\beta_{(i),l_{k-1}}||)'\}$.

The method of proof is similar to that in Cheng (2015); in particular, it exploits an inductive argument along the parameter grouping in order of decreasing identification strength in order to prove consistency in a sequential manner. Part (a) relies on sequentially concentrating out the estimator group associated with the current inductive step, a procedure that Cheng (2015) refers to as sequential peeling of the criterion function. At each sequential step, the concentrated criterion function is analyzed with a second order mean value expansion about the point of sequential identification failure.

Here, we use the term sequential identification failure to indicate the point in the parameter space at which all groups of parameters with weaker identification strength than the current group are unidentified. That is, $\psi^0_{k^-,n} = (\beta'_{k^-,n}, \beta^{0\prime}_{l_k,n}, \beta^{0\prime}_{k^+}, \zeta'_n, \pi'_{k^-,n})'$ where $\beta^0_{l_k,n}$ and $\beta^0_{k^+}$ are both 0, so $Q_{(i),n}(\psi^0_{(i),k^-,n})$ does not depend upon $\pi_{l_k,n}$ or $\pi_{k^+,n}$. This is similar to an argument used in the proof of the estimator limiting distribution for the non-mixed identification strength case in Andrews and Cheng (2012a,b), and this argument is additionally used in establishing the limiting distribution of the estimators here.

Use of this procedure in a sequential manner allows us to establish uniform consistency of each of the estimators of the semi-strongly identified parameters as detailed by part (b) of the lemma, and it also provides us with an improved rate of convergence at each step as given in part (c). These rates of convergence will be used to establish the limiting distribution of the estimators.

The asymptotic distribution of the estimators is characterized based on two possibilities: either (a) l_K is not empty in which case there are weakly identified parameters, or (b) l_K is empty in which case there are no weakly identified parameters, and all parameters are consistently estimable. In the former case, the asymptotic distribution is shown to be normal, but when we have weakly identified parameters, the limiting distribution of the estimators is non-standard.

Recall that when l_K is not empty, the last group, K, is weakly identified, so $\sqrt{n}\beta_{l_K} \to b_{l_K}$. For each i, let $\mathcal{G}_{(i)}(\pi_{(i),l_K};\gamma_0)$ be a zero mean Gaussian process with covariance kernel $\Omega_{(i)}(\pi_{(i),l_K}, \tilde{\pi}_{(i),l_K};\gamma_0)$, and define the processes

$$\tau_{(i)}(\pi_{(i),l_K};\gamma_0) = \Big[H_{(i),K}(\pi_{(i),l_K};\gamma_0)\Big]^{-1}\Big(\mathcal{K}_{(i),K}(\pi_{(i),l_K};\gamma_0)b_{(i),l_K} + \mathcal{G}_{(i)}(\pi_{(i),l_K};\gamma_0)\Big) \tag{3.2}$$

$$\chi_{(i)}(\pi_{(i),l_K};\gamma_0) = -\frac{1}{2}\tau_{(i)}(\pi_{(i),l_K};\gamma_0)'\Big[H_{(i),K}(\pi_{(i),l_K};\gamma_0)\Big]\tau_{(i)}(\pi_{(i),l_K};\gamma_0). \tag{3.3}$$

The process $\chi_{(i)}(\pi_{(i),l_K})$ appears as the limiting distribution of the normalized concentrated criterion function. Following a similar argument to that used in Lemma 3.4.2, the normalized and centered concentrated criterion function is minimized by $\hat{\pi}_{l_K}$, so if $\chi_{(i)}(\pi_{(i),l_K})$ is continuous and uniquely minimized by some $\pi^*_{l_K}$, then this will provide the limiting distribution for $\hat{\pi}_{l_K}$ by the argmax continuity theorem in van der Vaart and Wellner (1996). We state this assumption next.

In case (b) for which l_K is empty, the second derivative of the criterion function $\nabla^2_{\psi_{(i),K^-}} m_t(\theta_{(i)}) = \nabla^2_{\theta_{(i)}} m_t(\theta_{(i)})$, and its normalized limit becomes

$$H_{(i),K-1}(\pi_{(i),l_{K-1}}, \pi_{(i),K}; \gamma_0) = H_{(i),K-1}(\gamma_0).$$

Further, we show that the normalized first derivative

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} B(\beta_{(i),K^-,n})^{-1} \nabla_{\psi_{(i),k^-}} m_t(\theta_{(i)}) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} B(\beta_{(i),n})^{-1} \nabla_{\theta_{(i)}} m_t(\theta_{(i)}) \xrightarrow{d} \mathcal{G}_{(i),\theta}(\gamma_0).$$

where $\mathcal{G}_{(i),\theta}(\gamma_0) \sim N(0, \Omega_{(i),\theta}(\gamma_0))$.

In order to establish the joint limiting distribution, we require that the joint process $\{\chi_{(i)}(\pi_{(i),l_K}; \gamma_0) : 1 \leq i \leq \mathring{k}_w\}$ be uniquely minimized by some vector $[\pi^*_{(i),l_K}]_{i=1,\ldots,\mathring{k}_w}$, where $\mathring{k}_w$ is the number of parsimonious models with weakly identified parameters.

Assumption 21. *Each sample path of the joint process* $\{\chi_{(i)}(\pi_{(i),l_K}; \gamma_0) : 1 \leq i \leq \mathring{k}_w\}$ *is a.s. continuous and uniquely minimized by the vector* $[\pi^*_{(i),l_K}]_{i=1,\ldots,\mathring{k}_w}$ *with probability 1.*

Theorem 3.4.3. *Let* $\gamma_n \to \gamma_0$ *and suppose Assumptions 1-8 hold. Then*

$$\left\{ n^{1/2} B_{(i)}(\beta_{(i),n}) \begin{pmatrix} (\hat{\psi}_{(i),K^-} - \psi_{(i),K^-,n}) \\ \hat{\pi}_{(i),l_K} - \pi_{(i),l_K,0} \end{pmatrix} : 1 \leq i \leq \mathring{k} \right\} \xrightarrow{d} \left\{ \mathfrak{Z}_{(i)} : 1 \leq i \leq \mathring{k} \right\},$$

where $\mathfrak{Z}_{(i)}$ *are defined point-wise in* i *based on the cases:*

a) If $l_K \neq \emptyset$, where l_K indexes the weakly identified subvector of $\pi_{(i)}$, then

$$\mathfrak{Z}_{(i)} = \begin{pmatrix} \tau_{(i)}(\pi^*_{(i),l_K}(b, \gamma_0)) - S_{l_K} b_{(i),l_K} \\ ||\tau_{(i),\beta_K}(\pi^*_{(i),l_K}(b, \gamma_0))|| \left(\pi^*_{(i),l_K}(b, \gamma_0) - S_{\pi_{(i),l_K}} \pi_{l_K,n} \right) \end{pmatrix}$$

where S_{l_K} is the selection matrix that selects the columns corresponding to $\beta_{(i),l_k}$, and $S_{\pi_{(i),l_K}}$ selects the elements of the vector $\pi_{l_K,n}$ corresponding to $\pi_{(i),l_K,n}$.

b) if $l_k = \emptyset$, then no parameters are weakly identified, and

$$\mathfrak{Z}_{(i)} = H_{(i),K-1}(\gamma_0)^{-1}\mathcal{G}_{(i),\theta}(\gamma_0)$$

where $\mathcal{G}_{(i),\theta}(\gamma_0) \sim N(0, \Omega_{(i),\theta}(\gamma_0))$.

Note that we suppress the dependence of l_K upon i for convenience in notation, and recall that if l_K is empty, then no parameters are weakly identified, so all parameters and estimators indexed by l_K disappear and $\psi_{(i),K^-} = \theta_{(i)}$. If $l_{(i),K}$ is empty for every i, then the limiting distribution $\left\{\mathfrak{Z}_{(i)} : 1 \leq i \leq \mathring{k}\right\}$ is a Gaussian process with covariance kernel $E_{\gamma_0}[\mathfrak{Z}_{(i)}\mathfrak{Z}'_{(j)}]$. However, the presence of weakly identified parameters complicates this limiting distribution.

Note also that when $l_K \neq \emptyset$, we suppress the dependence of $\mathfrak{Z}_{(i)} = \mathfrak{Z}_{(i)}(\pi^*_{(i),l_K}(b, \gamma_0))$ upon $\pi^*_{(i),l_K}(b, \gamma_0)$ in the statement of Theorem 3.4.3, as doing so simplifies the presentation of the theorem. The reader should note, however, that in the presence of weak identification, the limiting distribution of the ith parsimonious estimator involves a Gaussian process evaluated at the random variable $\pi^*_{(i),l_K}(b, \gamma_0)$, which is itself a function of nuisance parameters. This differs from the case when there are no weakly identified parameters and the limiting distribution of the ith parsimonious estimator is a Gaussian random variable.

The point-wise in i asymptotic distribution of the parsimonious estimators are described in Theorem B.4.1 in the appendix. However, the max test combines estimators across parsimonious models, so it is necessary that we analyze the joint limiting distribution of the parsimonious estimators. Theorem 3.4.3 provides this joint asymptotic distribution. Observe that when there are both parsimonious models with weakly identified parameters and without weakly identified parameters, the joint limiting distribution consists of a combination of normal random variables and Gaussian functionals of the random variables $\pi^*_{(i),l_K}$. We know then from Cramér's Theorem that the joint limiting distribution is non-standard.

Corollary 3.4.4 (Asymptotic Distribution of Parsimonious Estimators)**.** *Suppose Assumptions 1-7*

and 9 hold. Then, under $\gamma_n \to \gamma_0$*,*

$$\left\{ \mathcal{N}_{(i),\lambda,n} \hat{\lambda}_{(i)} : \ 1 \leq i \leq \mathring{k} \right\} \xrightarrow{d} \left\{ S'_{(i),\lambda} \mathfrak{Z}_{(i)} : \ 1 \leq i \leq \mathring{k} \right\},$$

where $\mathcal{N}_{(i),\lambda,n} = S'_{(i),\lambda}(diag(n^{1/2} B_{(i)}(\beta_{(i),K^-,n}))', 1'_{d_{\pi_{(i),l_K}}})'$ *and* $S_{(i),\lambda}$ *is the selection matrix that selects the element corresponding to* $\lambda_{(i)}$*.*

For example, consider the additive nonlinear model

$$Y_t = \zeta' Z_t + \sum_{j=1}^{p} \beta_j g_j(X_{j,t}, \pi_j) + \varepsilon_t$$

with the null hypothesis $H_0 : \beta_j = 0 \ \forall j$. The parsimonious models are

$$Y_t = \zeta' Z_t + \beta_i g_j(X_{j,t}, \pi_i) + \nu_{(i),t}$$

for $i = 1, \ldots, \mathring{k}$ for some $\mathring{k} \geq p$. $\theta_{(i)} = (\zeta', \beta_i, \pi_i')'$, $\lambda_{(i)} = \beta_i$, and $S_{(i),\lambda} = (0'_{d_\zeta}, 1, 0'_{d_{\pi_{(i)}}})'$.

3.4.2 Linking the Unrestricted and Parsimonious Models

Since each parsimonious estimator minimizes a misspecified loss function, we cannot in general show that any parsimonious estimator consistently estimates the relevant components of the true parameter. Each parsimonious estimator is still a best predictor in some sense, and we can say that the estimator is consistent for a pseudo-true value that minimizes the population version of the misspecified parsimonious loss function. See White (1981) for a discussion of this issue. In particular, the pseudo-true value $\theta_{(i),n}$ is not in general equal to the relevant components of the true parameter $[\theta_n]_{(i)}$. Here, we show that our conditions are sufficient to equate these two values under the null hypothesis. We require one additional assumption that has not yet been stated.

Assumption 22. *1. if* $l_K = \emptyset$*, then* $E_{\gamma_0}(m_t(\theta; W_t))$ *is minimized uniquely by* $\theta = \theta_0 \in \Theta^*$*.*

2. *if* $l_K \neq \emptyset$*, then* $E_{\gamma_0}(m_t(\psi_{K^-}, \pi_K; W_t))$ *is minimized uniquely by* $\psi_{K^-} = \psi_{K^-,0} \in \Psi^*_{K^-}$ *for every* $\pi_K \in \Pi_K$*.*

This assumption is similar to Assumption 4; however, where Assumption 4 states that the parsimonious population criterion function is minimized uniquely by the pseudo-true parameter $\theta_{(i),0}$, this assumption states that the unrestricted population criterion function is minimized uniquely by the true parameter θ_0.

Recall that we partition $\theta = (\delta', \lambda', \tilde{\lambda}')'$ and $\theta_{(i)} = (\delta'_{(i)}, \lambda'_{(i)}, \tilde{\lambda}'_{(i)})'$.

Theorem 3.4.5. *Let Assumptions 4, 5, and 9 hold, and let λ be a subvector of ψ_{K^-}. Then $\lambda_0 = 0_{d_\lambda}$ if and only if $\lambda_{(i),0} = 0$ for every i, $1 \leq i \leq d_\lambda$.*

Together, Assumptions 4, 5, and 9 are sufficient for a similar assumption to Assumption 1 in Hill and Dennis (2018) which establishes Theorem 3.4.5 as Theorem 2.1 in Hill and Dennis (2018). In particular, by construction of our criterion function $Q_{(i),n}(\theta) \equiv Q_{(i),n}(\theta_{(i)}; W_t) = \frac{1}{n}\sum_{t=1}^{n} m_{(i),t}(\theta_{(i)}; W_t)$ and $m_{(i),t}(\theta_{(i)}; W_t) = m_t([\theta]_{(i)}; W_t)$ where $[\theta]_{(i)}$ is the restricted full parameter with $\lambda_j = 0$ for every $j \neq i$. It follows that

$$\nabla_{\psi_{(i),K^-}} m_{(i),t}(\theta_{(i)}) = \nabla_{\psi_{(i),K^-}} m_t([\theta]_{(i)})$$

for every i, providing the necessary link between the unrestricted and parsimonious models specifically at the point where the null hypothesis holds. The result in Theorem 3.4.5 then follows from the assumptions that both the population unrestricted and population parsimonious models are uniquely minimized by their respective true and pseudo-true values.

Note the importance of the imposition of the particular null hypothesis that $\lambda = 0$. It is this particular hypothesis that results in the parsimonious models used in estimation. An extension of this framework to allow a different null $\tilde{H}_0 : \lambda = \lambda_0$ for some $\lambda_0 \neq 0$ can be considered when the criterion function is based on a regression model. Consider the linear example $y_t = \delta' Z_t + \lambda' X_t + \varepsilon_t$. Testing $\tilde{H}_0$ will require construction and estimation of the sequential null imposed parsimonious models $y_t - \lambda'_{-i,0} X_{-i,t} \equiv y_{i,t} = \delta' Z_t + \lambda'_i X_{i,t} + \nu_{i,t}$.

This theorem provides the convenient implication that under the null hypothesis $H_0 : \lambda_0 = 0_{d_\lambda}$, $\delta_{(i),0} = \delta_0$ for every $i = 1, \ldots, d_\lambda$. Effectively, under the null hypothesis and when considering only the limiting values, there are no omitted variables, so there is no omitted variable bias. Since

we consider uniform inference over the parameter space, we allow for sequences $\lambda_n \to 0$ under the null hypothesis, so that omitted variable bias may still exist under the null hypothesis in finite samples. This theorem provides the result that under the null hypothesis there is no asymptotic omitted variable bias. Further, consistency of our test follows since under any alternative H_A : $\lambda_0 \neq 0_{d_\lambda}$, there is an i with $1 \leq i \leq d_\lambda$ such that $\lambda_{(i),0} \neq 0$ when λ is a subvector of ψ_{K^-}.

This implication helps us establish consistency of our test when the null hypothesis only involves parameters that are not weakly identified. The reader should note that when the null hypothesis involves weakly identified parameters, consistency cannot be guaranteed in general. Lemma B.6.1 illustrates this point. In order to describe the distribution of the estimators, we employ a more convenient normalization different from that used in Lemma B.6.1; however, the principle remains the same. In particular, this normalization, described in Theorem B.4.2, implies that when there are weakly identified parameters ($l_K \neq \emptyset$)

$$n^{1/2} B(\beta_{(i),n}) \begin{pmatrix} (\hat{\psi}_{(i),K^-} - \psi_{(i),K^-,n}) \\ \hat{\pi}_{(i),l_K} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \tau_{(i)}(\pi^*_{(i),l_K}) - S_{l_K} b_{(i),l_K} \\ ||\tau_{(i),\beta_K}(\pi^*_{(i),l_K})||\pi^*_{(i),l_K} \end{pmatrix}.$$

Ignoring the parsimonious model subscript i for a moment, we can see that the issue here arises because $n^{1/2}||\beta_{K,n}||$ converges to a finite constant rather than diverging to infinity. Since $\hat{\beta}_{K,n} \xrightarrow{p} 0$ at rate $n^{1/2}$, $n^{1/2}||\hat{\beta}_{K,n}|| \xrightarrow{d} ||\tau_{\beta_K}(\pi^*_{l_K})||$, a random variable.

However, since the inability to identify any π_j is driven by a particular $\beta_{j,0} = 0$ (by Assumption 3), the result of Theorem 3.4.5 implies that a subvector of $\pi_{(i)}$ is weakly identified if and only if its true model counterpart is weakly identified. Hence the distribution of the parsimonious estimator may be used for inference.

Here we must also point out another limitation of the test; namely, our max test does not accommodate complex hypothesis tests. This is an interesting issue in testing with mixed identification strength, since a test involving a complex hypothesis such as $\pi_1 + \pi_2 = 0$ will be dominated by the parameter with the weakest identification strength. Cheng (2015) addresses this issue in her Wald test with a rotation matrix that effectively alters a complex Wald test involving parameters

with more than one identification strength to a test involving only the parameters with the weakest identification strength.

3.5 Max Test Limit Theory and Inference

Recall that we partition our parameter vector $\theta = (\delta', \lambda', \tilde{\lambda}')'$ where δ is a vector of nuisance parameters, $\tilde{\lambda}$ is an additional vector of nuisance parameters that are tied to λ and are described later, and we are interested in testing the hypothesis $H_0 : \lambda_0 = 0_{d_\lambda}$ where the dimension of λ, d_λ, is potentially large.

We test the equivalent null hypothesis that $\lambda_{i,0} = 0$ for every i by estimating the restricted parameter $\theta_{(i)} = (\delta', \lambda_i', \tilde{\lambda}_i')'$ from the ith parsimoniously constructed model with loss function $Q_{(i),n}(\theta_{(i)}) = Q_n([\theta]_{(i)})$ where $[\theta]_{(i)} = (\delta', 0, \ldots, 0, \lambda_i, 0, \ldots, 0, \tilde{\lambda}_i, 0, \ldots, 0)'$. That is, the ith parsimonious model is constructed by restricting all elements $\lambda_j = 0$ for $j \neq i$ and $i = 1, \ldots, \mathring{k}$ where $\mathring{k} \geq d_\lambda$. The associated $\tilde{\lambda}_j$ elements do not appear when $\lambda_j = 0$, so we set them equal to zero without loss of generality. That is, for each i

$$\hat{\theta}_{(i)} = \underset{\theta_{(i)} \in \Theta_{(i)}}{\operatorname{argmin}} \; Q_{(i),n}(\theta_{(i)}).$$

We test the null hypothesis that a subvector λ of θ is the zero vector:

$$H_0 : \lambda_0 = 0_{d_\lambda} \qquad \text{vs.} \qquad H_A : \lambda_{i,0} \neq 0 \text{ for some } i.$$

To test this hypothesis, we collect the relevant $\hat{\lambda}_{(i)}$'s and form the test statistic

$$\hat{\mathcal{T}}_n = \max_{1 \leq i \leq \mathring{k}_n} \left| \mathcal{N}_{(i),\lambda,n} \mathcal{W}_{(i),n} \hat{\lambda}_{(i)} \right|$$

where $\mathcal{N}_{(i),\lambda,n}$ gives the appropriate standardization as described in section 3.4.1, $\mathcal{W}_{(i),n}$ is an additional weighting term that we assume is uniformly consistent for some constant $W_{(i)}$, and $\mathring{k}_n \to \mathring{k} \geq d_\lambda$ where $\mathring{k}$ is allowed to be ∞. Since each $\hat{\lambda}_{(i)}$ may be scaled differently, the $\mathcal{W}_{(i),n}$ can provide the appropriate scaling; we will use $\mathcal{W}_{(i),n}$ as the inverse of the standard error.

Theorem 3.4.3 paired with the CMT and a result from Hill and Dennis (2018) allow us to

establish the following result.

Theorem 3.5.1. *Let Assumptions 1-9 hold, and let* $\mathfrak{Z}_{(i)}$ *be as in Theorem 3.4.3. Then under* H_0*,*

$$\left|\hat{\mathcal{T}}_n - \max_{1\leq i\leq \mathring{k}_n} |S'_{(i),\lambda} W_{(i)} \mathfrak{Z}_{(i)}|\right| \xrightarrow{p} 0$$

for some non-unique $\mathring{k}_n = o(n)$*.*

Recall that, as discussed in section 3.4.2, any dependence across the parsimonious estimators $\hat{\lambda}_{(i)}$ due to omitted variables is asymptotically negligible under H_0. One strategy to derive the asymptotic null distribution of the maximum test statistic would be to appeal to extreme value theoretic results. This would allow us to pin down the exact limiting distribution by use of the extremal types theorem (Gnedenko, 1943; de Haan, 1976; Leadbetter et al., 1983). Standard arguments used for inference on maximum statistics rely on this method.[11] However, the asymptotic analysis of extremes of non-identically distributed Gaussian processes poses interesting statistical challenges that are not addressed here.

Instead, we provide an inference method based on the bootstrap. Our procedure does not require that we know the asymptotic distribution of the test statistic; valid inference only requires that our bootstrap estimator and test statistic converge to the same distribution.

Recent work for high dimensional statistics has focused on by-passing extreme value theory but has been limited by not allowing for dependence or residuals or by only allowing for Gaussian approximation (Chernozhukov et al., 2013, 2017; Zhang and Cheng, 2018; Zhang and Wu, 2017). Theory in Hill and Motegi (2018); Hill and Dennis (2018) allows for dependence under the null, residuals, and does not require Gaussianity. Here, we side-step the extreme value theory asymptotics by using the approach found in Hill and Dennis (2018) paired with the wild bootstrap Wu (1986); Liu (1988); Shao (2010, 2011a).

In practice, we do not know which parameters, if any, are weakly identified. Computation of

[11] see e.g. Xiao and Wu (2014) for a recent treatment involving the maximum of a sequence of covariances or Chernozhukov (2005) and Chernozhukov, Fernández-val, and Kaji (2017) for extreme value theory applied to quantile regression.

robust critical values will involve a null imposed information criteria selection (ICS) procedure. This involves computation of critical values under both the situation where we have a consistent estimator for $\pi_{(i)}$ and the situation for which we do not. Below, we first describe the algorithm used to perform the bootstrap assuming that the correct identification categories are known. Then we discuss data dependent critical value computation and the identification category selection procedures.

3.5.1 Inference

We provide two procedures for conducting inference. Both methods are based on the wild bootstrap (Wu, 1986; Liu, 1988). The wild bootstrap is a multiplier bootstrap. Wu (1986) and Liu (1988) detail the classic wild bootstrap for iid and non-iid sequences. Hansen (1996) allows for adapted martingale difference sequences, and Shao (2010, 2011a) allows for dependent sequences. Shao (2010) uses iid random draws as weights with a kernel function, but does not allow for a truncated kernel. Shao (2011a) uses a truncated kernel function. The wild bootstrap in convenient to use in many circumstances as it has been shown to allow for heteroskedasticity of unknown forms (Davidson and MacKinnon, 2010).

The first procedure is designed to work for any model satisfying the assumptions of Theorem 3.5.1. It involves calculation of the sample analogues of the key quantities used to construct the null imposed limiting distributions detailed in Theorem 3.4.3. These quantities are functions of the true parameter γ_0; hence they contain nuisance parameters since $\pi_{l_K,0}$ is not consistently estimable. The second procedure is easier to implement but only able to accommodate a restricted class of models, and we believe it to be valid only for strongly identified parameters. We provide only simulation evidence for the second procedure rather than proving its validity.

We prove the validity of the procedures presented here, but we do not expect the bootstrap to provide any second order improvements. In this sense, the bootstrap is meant to provide a convenient method for inference. See e.g. Moreira, Porter, and Suarez (2009) for a discussion of this issue. For both bootstrap procedures, we impose the null hypothesis following the advice of Davidson and MacKinnon (1999, 2010), who argue that doing so will provide higher power. Additionally, this has the benefit of reducing the dimension of the nuisance parameter space in

many cases.

3.5.2 Conditional Simulation Based Inference

This method is a variant of a wild bootstrap that falls in-between a traditional wild bootstrap and inference based on simulation. We call it conditional simulation based inference, as we simulate the limiting distribution of the test statistic conditional on the data. Recall the quantities defined in equations 3.2 and 3.3 and Theorems B.4.1 and 3.4.3.

$$\tau_{(i)}(\pi_{(i),l_K};\gamma_0) = \Big[H_{(i),K}(\pi_{(i),l_K};\gamma_0)\Big]^{-1}\Big(\mathcal{K}_{(i),K}(\pi_{(i),l_K};\gamma_0)b_{(i),l_K} + \mathcal{G}_{(i)}(\pi_{(i),l_K};\gamma_0)\Big)$$
$$\chi_{(i)}(\pi_{(i),l_K};\gamma_0) = -\frac{1}{2}\tau_{(i)}(\pi_{(i),l_K};\gamma_0)'\Big[H_{(i),K}(\pi_{(i),l_K};\gamma_0)\Big]\tau_{(i)}(\pi_{(i),l_K};\gamma_0)$$

The sample analogue of $\mathcal{G}_{(i)}(\pi_{(i),l_K};\gamma_0)$ is $\hat{\mathcal{G}}_{(i)}(\pi_{(i),l_K})$. Draw $z_t \sim N(0,1)$ and form the bootstrap sample analogue

$$\hat{\mathcal{G}}^{bs}_{(i)}(\pi_{(i),l_K}) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} z_t\Big\{m_{(i),t}(\hat{\psi}^0_{(i),K^-,n}(\pi_{(i),l_K}),\pi_{(i),l_K}) - \frac{1}{n}\sum_{t=1}^{n} m_{(i),t}(\hat{\psi}^0_{(i),K^-,n}(\pi_{(i),l_K}),\pi_{(i),l_K})\Big\}.$$

Use this to form the quantities

$$\hat{\tau}^{bs}_{(i)}(\pi_{(i),l_K};\gamma_0,b) = \Big[\hat{H}_{(i),K}(\pi_{(i),l_K})\Big]^{-1}\Big(\hat{\mathcal{K}}_{(i),K}(\pi_{(i),l_K};\gamma_0)b_{(i),l_K} + \hat{\mathcal{G}}^{bs}_{(i)}(\pi_{(i),l_K})\Big)$$
$$\hat{\chi}^{bs}_{(i)}(\pi_{(i),l_K};\gamma_0,b) = -\frac{1}{2}\hat{\tau}^{bs}_{(i)}(\pi_{(i),l_K};\gamma_0,b)'\Big[\hat{H}_{(i),K}(\pi_{(i),l_K})\Big]\hat{\tau}^{bs}_{(i)}(\pi_{(i),l_K};\gamma_0,b)$$

Next, compute

$$\pi^{*,bs}_{(i),l_K}(\gamma_0,b) = \underset{\pi_{(i),l_K}\in\Pi_{(i),l_K}}{\operatorname{argmin}} \hat{\chi}^{bs}_{(i)}(\pi_{(i),l_K};\gamma_0,b).$$

Denote by $\Rightarrow^p$ weak convergence in probability on the space of uniformly bounded functions, l_∞, as defined in Giné and Zinn (1990).

Lemma 3.5.2. *Let the assumptions of Theorem 3.5.1 hold.*

$$\left\{ \begin{pmatrix} \hat{\tau}^{bs}_{(i)}(\pi^{*,bs}_{(i),l_K}(\gamma_0, b); \gamma_0, b) \\ \pi^{*,bs}_{(i),l_K}(\gamma_0, b) \end{pmatrix} : 1 \leq i \leq \mathring{k} \right\} \Rightarrow^p \left\{ \mathring{\mathfrak{Z}}_{(i)}(\gamma_0, b) : \ 1 \leq i \leq \mathring{k} \right\},$$

where $\mathring{\mathfrak{Z}}_{(i)}(\gamma_0, b)$ *is an independent copy of the process described in Theorems B.4.1 and 3.4.3.*

Let $S_{(i),\lambda}$ be the selection matrix that selects the element corresponding to $\lambda_{(i)}$ as described in Corollary 3.4.4, and define

$$\mathring{\lambda}_{(i)}(\gamma_0, b) = S'_{(i),\lambda} \begin{pmatrix} \hat{\tau}^{bs}_{(i)}(\pi^{*,bs}_{(i),l_K}(\gamma_0, b); \gamma_0, b) \\ \pi^{*,bs}_{(i),l_K}(\gamma_0, b) \end{pmatrix}.$$

The following corollary is a direct result of the previous lemma.

Corollary 3.5.3. *Let the assumptions of Theorem 3.5.1 hold.*

$$\left\{ \mathring{\lambda}_{(i)}(\gamma_0, b) : \ 1 \leq i \leq \mathring{k} \right\} \Rightarrow^p \left\{ S'_{(i),\lambda} \mathring{\mathfrak{Z}}_{(i)}(\gamma_0, b) : \ 1 \leq i \leq \mathring{k} \right\}$$

It is easy to see from the above corollary that the procedure described above will simulate the distribution of the max statistic for fixed $\mathring{k}$. The next theorem utilizes a result in Hill and Dennis (2018) to extend this result to allow for increasing sequences $k_n = o(n)$.

Theorem 3.5.4. *Let the assumptions of Theorem 3.5.1 hold. For some non-unique sequence of positive integers* $\{k_n\}$, $k_n \to \infty$ *and* $k_n = o(n)$,

$$\sup_{c>0} \left| P(\max_{1 \leq h \leq k_n} |\mathring{\lambda}_{(i)}(\gamma_0, b)| \leq c | \mathcal{W}_n) - P(\max_{1 \leq h \leq k_n} |S'_{(i),\lambda} \mathring{\mathfrak{Z}}_{(i)}(\gamma_0, b)| \leq c) \right| \xrightarrow{p} 0$$

3.5.3 Residual Multiplier Bootstrap

We present this procedure for models of the form

$$Y_t = f(X_t, Z_t, \theta) + \sigma(X_t, Z_t, \theta)\varepsilon_t.$$

We first estimate the null imposed model $Q_n(\theta^{(0)}) = \frac{1}{n}\sum_{t=1}^{n} m_t(\theta^{(0)})$ with $\theta^{(0)} = [\delta^{(0)\prime}, 0'_{d_\lambda}]'$ to generate the estimates of the nuisance parameters $\hat{\delta}^{(0)}$. If elements of π_{l_K} are included in δ, then the estimates of these elements will not be consistent. Since $\hat{\pi}_{l_K}$ is not a consistent estimator of $\pi_{l_K,0}$, we must consider all vectors $\hat{\delta}^{(0)}(\pi_{l_K}) \in \{(\hat{\delta}_{K^-}(\pi_{l_K})', \pi'_{l_K})' : \pi_{l_K} \in \Pi_{l_K}\}$. Define $\hat{\theta}^{(0)}(\pi_{l_K}) = (\hat{\delta}_{K^-}(\pi_{l_K})', \pi'_{l_K}, 0'_{d_\lambda})'$. Use this estimator to construct the null imposed residuals $\tilde{\varepsilon}_t(\pi_{l_K}) = \varepsilon_t(\hat{\theta}^{(0)}(\pi_{l_K}))$. Imposing the null hypothesis allows the test to have power, and it has the potential to greatly reduce the dimension of the nuisance parameter space.

Draw a multiplier sequence $\{z_t\}_{t=1}^{n} \sim$ iid $N(0,1)$ and generate $\tilde{Y}_t^m(\pi_{l_K}) = f(X_t, Z_t, \hat{\theta}^{(0)}(\pi_{l_K})) + \sigma(X_t, Z_t, \hat{\theta}^{(0)}(\pi_{l_K}))\tilde{\varepsilon}_t(\pi_{l_K})z_t$, and let $\tilde{\mathcal{W}}_t^m(\pi_{l_K}) = (\tilde{Y}_t^m(\pi_{l_K}), X_t, Z_t)$. Construct and estimate the $\mathring{k}_n$ models via the parsimonious loss functions

$$\hat{\tilde{\theta}}_{(i)}^m(\pi_{l_K}) = \underset{\theta_{(i)} \in \Theta_{(i)}}{\operatorname{argmin}} \frac{1}{n}\sum_{t=1}^{n} m_{(i),t}(\theta_{(i)}; \tilde{\mathcal{W}}_t^m(\pi_{l_K})).$$

Collect the $\hat{\tilde{\lambda}}_{(i)}^m(\pi_{l_K})$ and form the bootstrapped statistic $\hat{\tilde{\mathcal{T}}}_n^m(\pi_{l_K}) = \max_{1 \le i \le \mathring{k}_n} \left|\mathcal{N}_{(i),\lambda,n}\mathcal{W}_{(i),n}\hat{\tilde{\lambda}}_{(i)}^m(\pi_{l_K})\right|$. Repeat this procedure $\mathcal{M}$ times to generate the sequence $\{\hat{\tilde{\mathcal{T}}}_n^m(\pi_{l_K})\}_{m=1}^{\mathcal{M}}$.

The π_{l_K} dependent α-level critical value is then $\hat{\tilde{\mathcal{T}}}_n^{[(1-\alpha)m]}(\pi_{l_K})$, and the associated p-value is $\hat{\tilde{p}}_{n,\mathcal{M}}(\pi_{l_K}) = \frac{1}{\mathcal{M}}\sum_{m=1}^{\mathcal{M}} I(\hat{\tilde{\mathcal{T}}}_n^m(\pi_{l_K}) > \hat{\mathcal{T}}_n)$. As this is a function of π_{l_K}, we consider two types of critical values - one for benchmarking and the other for use in practice. The infeasible α-level critical value is $\hat{\tilde{\mathcal{T}}}_n^{[(1-\alpha)m]}(\pi_{l_K,0})$, and the feasible α-level critical value is $\sup_{\pi_{l_K} \in \Pi_{l_K}} \hat{\tilde{\mathcal{T}}}_n^{[(1-\alpha)m]}(\pi_{l_K})$.

3.5.4 Robust Inference

In practice, we do not know if l_K is empty or not. We utilize a data dependent identification category selection procedure as described in Andrews and Cheng (2012a) and Cheng (2015) to select the elements from π that we believe to be not weakly identified. This Identification Category Selection procedure cannot fully determine the group specification, but does provide less conservative critical values than a procedure that does not rely on identification category selection (see

Cheng (2015) for details). Define the ICS statistic

$$ICS_{(i),j,n} = \left(n\hat{\beta}'_{(i),j}(\hat{\Sigma}_{(i),j})^{-1}\hat{\beta}_{(i),j}/d_{(i),\beta_j}\right)^{1/2}$$

where $\hat{\Sigma}_{(i),j}$ is the submatrix of $\hat{\Sigma}_{(i)}$ that corresponds to β_j. Note that $\hat{\Sigma}_{(i)} = \hat{H}^{-1}_{(i),K-1}\hat{\Omega}_{(i),\theta}\hat{H}^{-1}_{(i),K-1}$ is constructed assuming that l_K is empty.

Let $\{\kappa_{(i),j,n} : n \geq 1\}$ be a sequence of constants such that $\kappa_{(i),j,n} \to \infty$ and $\kappa_{(i),j,n}/n^{1/2} \to 0$ for every i and j. The weak identification group is selected as the set

$$\hat{l}_{(i),K} = \{j : ICS_{(i),j,n} \leq \kappa_{(i),j,n}\}.$$

The idea behind the identification category selection procedure is that the ICS statistic will diverge to ∞ whenever $\beta_{(i),j}$ is 'large enough.' Hence, this procedure forms a pre-test in which we reject the hypothesis of $\pi_{(i),j}$ being weakly identified whenever the ICS statistic is large. If the ICS statistic is small so that we fail to reject the weak identification hypothesis on $\pi_{(i),j}$, then we place $j \in \hat{l}_{(i),K}$. If the our null hypothesis involves $\beta_j = 0$, then we put $j \in \hat{l}_{(i),K}$ without selection.

3.6 Additional Examples

Here we discuss several examples that may be studied by utilizing this framework. The first example describes how this test can be used as a test for additional omitted non-linearity by using the additive non-linear model studied in Cheng (2015). This is related to the empirical study of non-linear mean reversion in exchange rate dynamics that has been used as a possible explanation for the Purchasing Power Parity Paradox. The second example describes the relationship between weakly identified ARMA(p,q) models, the common roots problem, and weak instruments in time series models. The third example discusses a nonlinear binary choice model. The fourth example discusses limited information maximum likelihood estimation (LIML) of linear instrumental variables models with many weak instruments.

3.6.1 Testing for Nonlinearity in Exchange Rate Dynamics

Purchasing Power Parity (PPP) embodies the idea that, when expressed in the same currency units, price levels should be equal across nations (Cassel, 1918). Variations in the real exchange rate can be thought of as deviations from PPP. A long literature has attempted to reconcile the high short-term volatility in real exchange rates with the slow rate at which convergence to PPP seems to occur. This has become known as the PPP puzzle (Rogoff, 1996).

The (log) real exchange rate can be expressed as $q_t = s_t - p_t + p_t*$, where s_t is the (log) nominal exchange rate, p_t is the logarithm of the domestic price level, and p_t* is the logarithm of the foreign price level. This formulation allows one to interpret the real exchange rate as a measure of deviation from Purchasing Power Parity. Taylor et al. (2001) and others note that studies of the effect of transaction costs on PPP suggest that exchange rate adjustments resemble a non-linear process in which the rate appears to be a unit root process within a band and a stationary process outside of that band. They model real exchange rate dynamics with a model that allows a smooth transition at the boundary of the band. In particular, they examine the STAR model (Granger and Teräsvirta, 1993)

$$q_t - \mu = \sum_{j=1}^{p} \beta_j (q_{t-j} - \mu) + \Big[\sum_{j=1}^{p} \beta_j^* (q_{t-j} - \mu) \Big] \Phi(\gamma; q_{t-d} - \mu) + \varepsilon_t$$

where $\{q_t\}$ is assumed stationary and ergodic with $\varepsilon_t \sim iid(0, \sigma^2)$ and the exponential transition function

$$\Phi(\gamma; q_{t-d} - \mu) = 1 - \exp(-\gamma^2 (q_{t-d} - \mu)^2).$$

Alongside the exponential transition function, the model is referred to as the ESTAR model. Similar models, including the Logistic (LSTAR) model with transition function

$$\Phi(\gamma; q_{t-d} - \mu) = \Big[1 - \exp(-\gamma (q_{t-d} - \mu)) \Big]^{-1}$$

have been used as specification tests for the estimated models. van Dijk et al. (2002) provide an extensive review of smooth transition models.

Two potential issues appear in this modeling exercise. First the unknown value of d must be selected. Second, given the non-linearity of the chosen model, parameter identification failure may result under some situations, and in particular parameter identification failure occurs under the null hypothesis when testing no omitted non-linearity. For the first point, Taylor et al. (2001) provide economic intuition in favor of smaller values of the parameter d, namely that we should not expected a long lag between a shock and the adjustment response from the exchange rate. The second issue is handled less satisfactorily, as the modeling procedure is based on a linearization of the non-linear model about the point of identification failure. This method addresses issues that arise from identification failure, but as recent research indicates, this may provide a poor approximation to the desired model (Kilic, 2016).

Further, Hill (2008) notes that the traditional method involving a truncated Taylor approximation simply "directs power toward low order polynomials" and is therefore not truly a test against smooth transition alternatives. He draws attention to the fact that treating d as a parameter to be estimated yields a non-standard limiting distribution, a fact that was ignored in the early literature. Importantly, he notes that a test that only considers a finite number of conditions (e.g. a small support for d or a finite-order polynomial approximation) can give rise to inconsistency. Francq, Horvath, and Zakoïan (2010) also examine non-standard tests that result due to the presence of nuisance parameters when testing for linearity against smooth transition autoregressive alternatives.

Taylor et al. (2001) follow a sequential modeling procedure similar to those suggested in Granger and Teräsvirta (1993), Terasvirta (1994), and Eitrheim and Teräsvirta (1996). Kilic (2016) follows the specification procedure in Teräsvirta (2004) and utilizes the diagnostic tests suggested by Eitrheim and Teräsvirta (1996) for the first differenced model

$$\Delta q_t = \left[\beta_0^* + \sum_{i=1}^{p} \beta_i^* \Delta q_{t-i}\right] \Phi(\gamma, \Delta q_{t-d}) + u_t.$$

This class of models fits into the class of additive nonlinear models

$$y_t = \sum_{j=1}^{p} \beta_j g_j(X_{j,t}, \pi_j) + Z_t'\zeta + \varepsilon_t.$$

In particular, there is no need for different models for each d, as the model corresponding to a particular d is just a restriction on a larger model:

$$\begin{aligned} y_t &= \sum_{j=1}^{s} \tilde{\beta}_j y_{t-j} + \sum_{d=1}^{r}\sum_{j=1}^{s} \tilde{\beta}^*_{j,d} y_{t-j}\Phi(\gamma; y_{t-d}) + \varepsilon_t \\ &= Z_t'\zeta + \sum_{j=1}^{p} \beta_j g_j(X_{j,t}, \pi_j) + \varepsilon_t \end{aligned}$$

where $p = rs$, $Z_t = (y_{t-1}, \ldots, y_{t-s})$, and $g_j(X_{j,t}, \pi_j) \equiv y_{t-j}\Phi(\gamma; y_{t-d})$. Letting $r \to \infty$, we can then form a test of no nonlinearity via the null hypothesis that $\beta = 0$ or a test of no omitted nonlinearity with the null hypothesis that a subset of β is the zero vector.

Typically in this literature, $\gamma = 0$ drives identification failure in β. This parameterization may lead to issues with inference under the framework presented here, since $\beta = 0$ would also induce the identification failure of γ. We are not aware of any study of such 'double identification failure.' For this setup, we will require that either $\gamma > 0$ or $\beta > 0$ so that only a single point of identification failure exists.

3.6.2 Weak Identification in Time Series

Here we describe the relationship between weakly identified ARMA(p,q) models, the common roots problem, and weak instruments in time series models.

Weak Identification and Common Roots

First, consider the ARMA(1,1) model $y_t = (\beta+\pi)y_{t-1}+\varepsilon_t-\pi\varepsilon_{t-1}$. Under commonly assumed conditions, one can show that when $\beta = 0$, the model reduces to $y_t = \varepsilon_t$.[12] Observe that the model can be rewritten as $\big(1 - (\beta - \pi)L\big)y_t = (1 - \pi L)\varepsilon_t$ where L is the lag operator. When $\beta = 0$,

[12] Write $\big(1 - (\beta - \pi)L\big)y_t = (1 - \pi L)\varepsilon_t$ and assume $\pi < 1$. Then the model can be written using a geometric sum $\big(1 - \beta\sum_{j=0}^{\infty} \pi^j L^{j+1}\big)y_t = \varepsilon_t$.

the roots of both the AR and MA polynomials are $1/\pi$. It is clear that if $\beta = 0$ then π is not identified; this is referred to as the 'common roots problem.' Researchers typically assume away this problem, as it leads to non-standard asymptotic analysis.

Further, Andrews and Ploberger (1996) note that ARMA models provide parsimonious representations of many different stationary time series, Poterba and Summers (1988) show that many mean-reverting financial time series can be represented by ARMA models, and Taylor (2005) shows that the ARMA model can be used to represent certain price-trend models. An issue arises with the assumption that the time series possesses no common roots in practice, as one does not know the data generating process that generated the data being analyzed. In particular, this is an issue for practitioners representing financial series with ARMA models, as with many such series, we expect there to be no correlation across time due to the forces associated with arbitrage. The no arbitrage condition manifests itself in the ARMA model as a common root, indicating that tests based on standard asymptotic analysis may be distorted. That is, inference based on standard asymptotics, which do not account for the distributions induced by weak or non-identified parameters, may tend to over-reject the null hypothesis that $\beta = 0$. Related issues specifically for the ARMA(1,1) model are studied by Andrews and Ploberger (1996); Andrews, Liu, and Ploberger (1998); Andrews and Cheng (2012a) and Dennis (2019).

In general, we can use the framework developed in this paper, in particular that developed section B.2, to analyze the ARMA(p,q) model, thereby extending current research. Write the ARMA(p,q) model in the form $\Phi(L)y_t = \Upsilon(L)\varepsilon_t$ where $\Phi(L) = \big(1-(\beta_1+\pi_1)L\big)\cdots\big(1-(\beta_p+\pi_p)L\big)$ and $\Upsilon(L) = (1-\pi_1)\cdots(1-\pi_q)$. For example, the ARMA(2,2) model can be written $\big(1-(\beta_1-\pi_1)L\big)\big(1-(\beta_2-\pi_2)L\big)y_t = (1-\pi_1 L)(1-\pi_2 L)\varepsilon_t$. Assume that $\varepsilon_t \sim$ iid $(0,\zeta)$ and $1-\beta_i-\pi_i < 1$ for each i. Quasi-maximum likelihood is used to estimate the model with

$$m_t(\theta) = \ln(\zeta) + \Big(\frac{\Phi(L)}{\Upsilon(L)}y_t/\zeta\Big)^2$$

Note that this example could be extended in a straight forward manner to account for analysis of ARMA models with conditional volatility.

This is related to the issue of studying a unit root ARMA process with an MA parameter close to -1 as studied by Schwert (1989), Davidson (2010) and the references there in. However, we do not study unit root processes or the distribution of the ADF test statistic.

Weak Instruments

The common roots problem is related to weak instruments in time series models. Consider the model $y_t = \alpha z_t + \varepsilon_t$ where we only observe z_t with error $x_t = z_t + \eta_t$. z_t is assumed to follow a time series process. For clarity of exposition, we demonstrate this section assuming z_t is ARMA(1,1): $z_t = (\beta + \pi)z_{t-1} + e_t - \pi e_{t-1}$, but more general models are allowed under the framework established in section B.2.

We let ε_t and e_t be correlated, but assume that ε_t and e_{t-1} are not correlated. The idea used to conduct inference on α is to use x_{t-1} as an instrument for z_t. When β is close to zero in a statistical sense,[13] the influence of z_{t-1} on z_t drops to zero. Hence, the correlation between x_{t-1} and z_t diminishes, inducing a weak instruments problem. In particular, we can show that $E[z_t z_{t-1}] = \sigma^2\left[\left(\frac{1-(\beta+\pi)(1-\beta)}{1-(\beta+\pi)}\right)(\beta+\pi) - \pi\right]$ which tends to 0 as $\beta \to 0$.

Common in the literature is to assume that z_t follows an AR(1) process. This amounts to the identification restriction that $\pi = 0$. The weak instruments issue described above still manifests near $\beta = 0$, as having both $\beta = 0$ and $\pi = 0$ is a special case of a common root. In general, however, π need not be zero, and how relaxing this assumption will affect inference on α is not clear and would be an interesting topic for study.

3.6.3 Nonlinear Binary Choice Model

Andrews and Cheng (2013) demonstrate that their framework is appropriate for analysis of the nonlinear binary choice model

$$y_i = 1(y_i^* > 0) \qquad \text{with} \qquad y_i^* = \beta g(X_i, \pi) + Z_i'\zeta - \varepsilon_i$$

where $g(X_i, \pi) \in \mathbb{R}$ is known up to the finite dimensional parameter π and estimation is carried out via maximum likelihood under some assumption on the specification of $P(y_i = 1|X_i, Z_i)$ such

[13] e.g. $\beta \leq b/\sqrt{n}$ for some small $b < \infty$, or technically, in terms of drifting sequences, $\sqrt{n}\beta_n \to b < \infty$.

as a probit or logit model. Their framework allows for vector β, but requires that all elements of π exhibit the same identification strength. Put another way, this requires that for a vector $\beta = \beta_n$ allowed to drift to zero, either all elements of β_n drift to zero at the same rate, or the identification strength of π must depend upon $\max_k |\beta_{k,n}|$. The latter case is not handled by their theory, and the former case seems to be a restrictive assumption.

The framework developed in section B.2 in this paper, however, is appropriate for analysis of the nonlinear binary choice model when the elements of π are allowed to exhibit different identification strengths. This is a relaxation of the assumption mentioned in the previous paragraph. In particular, the theory developed here is appropriate for models of the form

$$y_i = 1(y_i^* > 0) \qquad \text{with} \qquad y_i^* = \sum_{j=1}^{p} \beta_j g_j(X_{j,i}, \pi_j) + Z_i'\zeta - \varepsilon_i.$$

Observe that estimation and inference, allowing for mixed identification strength, for the model y_i^* is covered by the theory in Cheng (2015). However, her theory is only appropriate for the additive nonlinear model estimated by least squares; hence it does not apply to estimation and inference for the model given jointly by y_i and y_i^*, and in particular, y_i^* is usually not observed for this class of models.

3.6.4 Linear IV Model

Andrews and Cheng (2012a,b) demonstrate that the linear instrumental variable model

$$y_{1,i} = y_{2,i}\pi + u_i^*, \qquad y_{2,i} = Z_i'\beta + v_i^*$$

fits within their framework when estimated via limited information maximum likelihood (LIML). In particular, the reduced form equations $y_{1,i} - \pi \cdot Z_i'\beta + u_i$ and $y_{2,i} - Z_i'\beta + v_i$ with $u_i = u_i^* + \pi v_i^*$, $v_i = v_i^*$, and $(u_i, v_i) \sim N(0, Y)$ are estimated with the likelihood function

$$Q_n(\theta) = \log|Y| + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i(\beta, \pi)' Y^{-1} \varepsilon_i(\beta, \pi)$$

where $\varepsilon_i(\beta,\pi) = \begin{pmatrix} y_{1,i} - \pi \cdot Z_i'\beta \\ y_{2,i} - Z_i'\beta \end{pmatrix}$. Similarly to the discussion for the nonlinear binary choice model, their theory only accommodates a single endogenous covariate in this setup, as they do not allow for mixed identification strength in π. The theory developed in this paper, however, can be used to analyze models in this setup with more that one endogenous covariate.[14] Consider the structural model

$$y_i = x_{1,i}\pi_1 + x_{2,i}\pi_2 + u_i^*, \qquad x_{1,i} = Z_{1,i}'\beta_1 + v_i, \qquad x_{2,i} = Z_{2,i}'\beta_2 + \eta_i.$$

The reduced form equations are

$$y_i = Z_{1,i}'\beta_1\pi_1 + Z_{2,i}'\beta_2\pi_2 + u_i, \qquad x_{1,i} = Z_{1,i}'\beta_1 + v_i, \qquad x_{2,i} = Z_{2,i}'\beta_2 + \eta_i$$

where $u_i = v_i^*\pi_1 + \eta_i^*\pi_2 + u_i^*$, and similarly we can assume $(u_i, v_i, \eta_i) \sim N(0, Y)$. LIML estimation of instrumental variables models with weak instruments has been studied by Bound, Jaeger, and Baker (1996), Staiger and Stock (1997), Moreira (2003), Andrews, Moreira, and Stock (2006), Chao and Swanson (2007) and many others.

In addition to not allowing mixed identification strength (Andrews and Cheng, 2012a, 2013, 2014) and restricting the class of allowable models (Cheng, 2015), previous results for identification robust inference do not consider high dimensional parameters or max tests. In contrast, our theory allows for testing a large dimensional parameter by estimation of many parsimoniously constructed models and a test on the maximum of the sequence of estimators attained from the estimation.

Inference in models with many parameters is typically conducted with an imposed sparsity assumption by forcing a large number of the parameters to be equal to zero with a penalized estimator such as LASSO (Tibshirani, 1996) in a way that precludes inference on those parameters.

[14] Andrews and Stock (2007) note that the most important case in empirical applications involves only a single right hand side endogenous covariate; however, this does not mean that the ability to analyze a system with more than one endogenous covariate is not important.

As a result, valid inference can only be conducted on the remaining non-zero parameters in many cases. Recent work focusing on this inference issue has relied on 'desparsification' (van de Geer et al., 2014; Caner and Kock, 2018; Dezeure, Bühlmann, and Zhang, 2017) or 'debiasing' (Belloni et al., 2014b; Wooldridge and Zhu, ming) the LASSO estimator; however, using these procedures to conduct inference when some parameters are weakly identified has not been studied. In particular, one of the nice features of the LASSO is that it is a convex relaxation of a nonconvex problem; however, this convexity is not guaranteed when operating on nonlinear models.

Further, the LASSO sets exactly equal to zero any parameter that cannot be statistically distinguished from zero. Belloni et al. (2016), Leeb and Pötscher (2008) and Pötscher (2009) note that this can be problematic for conducting inference with approximately sparse models that include both variables with small but nonzero coefficients and strong predictors, since the LASSO will exclude the variables with small coefficients, which the authors note, can lead to omitted variable bias and irregular sampling behavior. Our approach differs in that we estimate a collection of parsimonious models by considering each parameter in turn and evaluating the maximum of the estimated values, thereby allowing inference on all parameters (Ghysels et al., 2016a; Hill and Dennis, 2018; Ghysels, Hill, and Motegi, ming).

In general, we may have a desire to test a large subset of our parameters based on economic reasoning or functional form. For example, Belloni et al. (2014b,a) perform a follow-up study regarding the effect of legalized abortion on crime (Donohue and Levitt, 2001, 2008; Foote and Goetz, 2008) in which they examine inference on treatment after selection amongst a high dimensional set of controls. They include a large set of controls that allows for flexible trends that vary with state-level characteristics. In particular, they alter the baseline model of Donohue and Levitt (2001) to include 284 variables[15] that allow for a "cubic trend for the level of the crime rate and abortion rate which is allowed to depend on observed state-level characteristics." The data set consists of only 600 observations, and they illustrate the poor performance of OLS due to the large

[15]"the levels, differences, initial level, initial difference, and within-state average of the eight state-specific time-varying observables, the initial level and initial difference of the abortion rate relevant for crime type, quadratics in each of the preceding variables, interactions of all the aforementioned variables with t and t^2, and the main effects t and t^2."

number of covariates relative to observations.

Their LASSO-double-selection method suggests that i) results based on a small set of intuitively selected controls differ from results obtained through formal variable selection and ii) accounting for nonlinear trends in the data affects the results, as well. Based on this discrepancy between results based on formal selection and intuitive selection, we can use the framework developed in this paper to examine whether the group of intuitively or economically relevant controls is relevant for the regression. Alternatively, we can use the max test to construct a test of the relevance of the controls added for fidelity, such as the group of all interactions of variables that are meant to allow for a more flexible functional form.

For simplicity of exposition, consider the model with one endogenous covariate

$$y_t = x_t\pi + Z_t'\omega + u_t^*, \qquad x_t = Z_t'\beta + v_t^*$$

where $\beta \in \mathbb{R}^{d_\beta}$ with $d_\beta = o(n)$ and t is used for the observation to avoid confusion with the parsimonious model index below. Here we wish to test the relevance of a potentially large subset of instruments, so the null hypothesis is $H_0 : (\beta_2', \omega_2')' = 0$ for some subvector β_2 of $\beta = (\beta_1', \beta_2')'$ and similarly for ω_2. The reduced form parsimonious models are

$$y_t = Z_{1,t}'(\beta_1\pi + \omega_1) + Z_{2,i,t}'(\beta_{2,i}\pi + \omega_{2,i}) + u_{i,t}, \qquad x_t = Z_{1,t}'\beta_1 + Z_{2,i,t}'\beta_{2,i} + v_{i,t}$$

In its simplest form, Z_1 will be empty ($\beta_1 = 0$), so each parsimonious model will have exactly one exogenous covariate, $Z_{2,i}$.

This is related to the literature that studies estimation and testing with many weak instruments (Bekker, 1994; Bekker and Kleibergen, 2003; Chao and Swanson, 2005; Chamberlain and Imbens, 2004; Andrews and Stock, 2007; Hansen, Hausman, and Newey, 2012; Hausman, Newey, Woutersen, Chao, and Swanson, 2012) and many others. In particular, Andrews and Stock (2007) examine the properties of certain tests and discuss the rate condition, $k^3/n \to 0$ needed for correct asymptotic size.

3.7 Monte Carlo Simulations

We consider the additive non-linear model

$$Y_t = \zeta' Z_t + \sum_{j=1}^{d_\beta} \beta_j g(X_{j,t}, \pi_j) + \varepsilon_t$$

where

$$g(X_{j,t}, \pi_j) = \left[1 - \exp(-\pi_{j,1}(X_{j,t} - \pi_{j,2}))\right]^{-1}.$$

For computational simplicity we fix $\pi_{j,1} = 10$ and only estimate $\pi_{j,2}$. We consider 3 data generating processes:

1) Independent regressors: $X_{j,t}$, $Z_{j,t}$, $\varepsilon_t \sim$ iid $N(0,1)$. Under this DGP, $Z_t \perp X_t$, $X_t \perp \varepsilon_t$, and $Z_t \perp \varepsilon_t$.

2) Block-wise Independent, Correlated Regressors: $X_t \sim N(0_{d_\beta}, \Sigma_x)$, $Z_t \sim N(0_{d_\beta}, \Sigma_z)$, but $Z_t \perp X_t$. $\varepsilon_t \sim$ iid $N(0,1)$.

3) Correlated Regressors: $(X_t', Z_t')' \sim N(0_{d_\beta + d_\zeta}, \Sigma)$. $\varepsilon_t \sim$ iid $N(0,1)$.

Without loss of generality, we set $d_\zeta = 2$ where the first element corresponds to a constant. For each DGP, we consider $n = 100, 500$ and the scenarios $\beta_1 \in \{0, b_1/\sqrt{n}, b_1\}$ for $b_1 = 1$. Additionally, we consider $d_\lambda \in \{1, 10, 20, 5\sqrt{n}\}$ and $k_n \in \{1, 10, 20, 5\sqrt{n}\}$ for $k_n \leq d_\lambda$.

We will test if a subvector of β is different from zero. Again without loss of generality, we test the subvector $\lambda = (\beta_2, \ldots, \beta_{d_\beta})'$. Hence the parsimonious models are constructed and estimated as

$$Y_t = \zeta' Z_t + \beta_1 g(X_{1,t}, \pi_1) + \lambda_{(i)} g(X_{(i),t}, \pi_{(i)}) + \nu_{(i),t}.$$

We consider the following hypotheses:

1) $H_0 : \lambda = 0$

2) Local Alternative $H_1 : \lambda_1 = b_2/\sqrt{n}$, $b_2 \in \{1, 2, 5, 10\}$, $\lambda_j = 0$ for every $j > 1$.

3) $H_2 : \lambda_1 = 1$, $\lambda_j = 0$ for every $j > 1$.

The parsimonious models are estimated via least squares:

$$Q_{(i),n}(\theta_{(i)}) = \frac{1}{n}\sum_{t=1}^{n} \nu_{(i),t}(\theta_{(i)})^2,$$

$$\nu_{(i),t}(\theta_{(i)}) = Y_t - \zeta' Z_t - \beta_1 g(X_{1,t}, \pi_1) - \lambda_{(i)} g(X_{(i),t}, \pi_{(i)}).$$

This gives the gradient and hessian of the criterion function:

$$\nabla_{\theta_{(i)}} m_{(i),t}(\theta_{(i)}) = 2\nu_{(i),t}(\theta_{(i)}) \nabla_{\theta_{(i)}} \nu_{(i),t}(\theta_{(i)})$$

and

$$\nabla^2_{\theta_{(i)}} m_{(i),t}(\theta_{(i)}) = 2\nu_{(i),t}(\theta_{(i)}) \nabla^2_{\theta_{(i)}} \nu_{(i),t}(\theta_{(i)}) + 2\nabla_{\theta_{(i)}} \nu_{(i),t}(\theta_{(i)}) \nabla_{\theta_{(i)}} \nu_{(i),t}(\theta_{(i)})'$$

where

$$\nabla_{\theta_{(i)}} \nu_{(i),t}(\theta_{(i)}) = -\begin{pmatrix} Z_t \\ g(X_{1,t}, \pi_1) \\ g(X_{(i),t}, \pi_{(i)}) \\ \beta_1 \frac{\partial}{\partial \pi_1} g(X_{1,t}, \pi_1) \\ \lambda_{(i)} \frac{\partial}{\partial \pi_{(i)}} g(X_{(i),t}, \pi_{(i)}) \end{pmatrix}$$

and

$$\nabla^2_{\theta_{(i)}} \nu_{(i),t}(\theta_{(i)}) =$$

$$
-\begin{pmatrix}
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{\partial}{\partial \pi_1} g(X_{1,t}, \pi_1) & 0 \\
0 & 0 & 0 & 0 & \frac{\partial}{\partial \pi_{(i)}} g(X_{(i),t}, \pi_{(i)}) \\
0 & \frac{\partial}{\partial \pi_1} g(X_{1,t}, \pi_1) & 0 & \beta_1 \frac{\partial}{\partial \pi_1} \frac{\partial}{\partial \pi_1'} g(X_{1,t}, \pi_1) & 0 \\
0 & 0 & \frac{\partial}{\partial \pi_{(i)}} g(X_{(i),t}, \pi_{(i)}) & 0 & \lambda_{(i)} \frac{\partial}{\partial \pi_{(i)}} \frac{\partial}{\partial \pi_{(i)}'} g(X_{(i),t}, \pi_{(i)})
\end{pmatrix}
$$

and $m_{(i),t}(\theta_{(i)}) = \nu_{(i),t}(\theta_{(i)})^2$, $g(X_{j,t}, \pi_j) = \Big[1 - \exp(-\pi_{j,1}(X_{j,t} - \pi_{j,2}))\Big]^{-1}$

$$
\frac{\partial}{\partial \pi_j} g(X_{j,t}, \pi_j) = \Big[1 - \exp(-\pi_{j,1}(X_{j,t} - \pi_{j,2}))\Big]^{-2} \Big[\exp(-\pi_{j,1}(X_{j,t} - \pi_{j,2})\Big] \pi_{j,1}
$$

and

$$
\begin{aligned}
\frac{\partial}{\partial \pi_j} \frac{\partial}{\partial \pi_j'} g(X_{j,t}, \pi_j) = 2\Big[1 - \exp(-\pi_{j,1}(X_{j,t} - \pi_{j,2}))\Big]^{-3} \Big[\exp(-\pi_{j,1}(X_{j,t} - \pi_{j,2})\Big]^2 \pi_{j,1}^2 \\
+ \Big[1 - \exp(-\pi_{j,1}(X_{j,t} - \pi_{j,2}))\Big]^{-2} \Big[\exp(-\pi_{j,1}(X_{j,t} - \pi_{j,2})\Big] \pi_{j,1}^2
\end{aligned}
$$

since we fix $\pi_{j,1} = 10$. Note that under the null hypothesis

$$
\frac{1}{n} \sum_{t=1}^{n} \nabla^2_{\theta_{(i)}} m_{(i),t}(\theta_{(i)}) = \frac{1}{n} \sum_{t=1}^{n} \nabla_{\theta_{(i)}} \nu_{(i),t}(\theta_{(i)}) \nabla_{\theta_{(i)}} \nu_{(i),t}(\theta_{(i)})' + o_{p,\pi_{(i),K}}(1).
$$

When l_K is not empty, we have

$$
\mathcal{K}_{(i),K}(\pi_{(i),l_K,0}; \gamma_0) = E_{\gamma_0} \nabla_{\theta_{(i)}} \nu_{(i),t}(\theta_{(i)}) \nabla_{\theta_{(i)}} \nu_{(i),t}(\psi_{(i),K^-}, \pi_{(i),l_K,0})' S'_{\beta_{l_K}}
$$

where, for example, $\nabla_{\theta_{(i)}} \nu_{(i),t}(\psi_{(i),K^-}, \pi_{(i),l_K,0})' S'_{\beta_{l_K}} = g(X_{(i),t}, \pi_{(i),0})$ when $l_{(i),K} = \{2\}$.

The following table shows rejection frequencies for the Wald, Max, and Max-t tests using various inference procedures as described. The standard inference method is labeled *standard*, the two inference procedures discussed in this paper are labeled *BS1* and *BS2*, and *Taylor* denotes the respective tests conducted on the model linearized with a first order Taylor expansion. Here, we present results only for the second data generating process with block-wise independent, correlated

regressors, as the others yield similar results.

Table 3.2: Max Test Simulations - Additive Nonlinear Model under the Null Hypothesis

	$k_{\lambda,n} = 1$						$k_{\lambda,n} = 20$					
b_1	0	1	2	5	10	14	0	1	2	5	10	14
Wald Test Standard	0.11	0.12	0.12	0.13	0.12	0.12	0.83	0.84	0.83	0.83	0.84	0.84
Max Test Standard	0.10	0.10	0.10	0.10	0.10	0.10	0.12	0.11	0.11	0.11	0.11	0.11
Max t-Test Standard	0.11	0.12	0.11	0.11	0.11	0.11	0.19	0.19	0.19	0.19	0.20	0.20
Wald Test BS1	0.06	0.06	0.06	0.06	0.06	0.06	0.68	0.68	0.67	0.68	0.69	0.69
Max Test BS1	0.04	0.04	0.04	0.04	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.03
Max t-Test BS1	0.06	0.06	0.06	0.06	0.06	0.06	0.13	0.12	0.13	0.13	0.13	0.13
Wald Test BS2	0.06	0.06	0.06	0.06	0.06	0.06	0.25	0.26	0.26	0.26	0.25	0.25
Max Test BS2	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.04	0.04
Max t-Test BS2	0.06	0.06	0.06	0.06	0.06	0.06	0.08	0.08	0.08	0.08	0.08	0.09
Wald Test Taylor	0.09	0.09	0.09	0.09	0.09	0.09	0.52	0.52	0.52	0.52	0.52	0.52
Max Test Taylor	0.60	0.60	0.60	0.60	0.60	0.60	0.99	0.99	0.99	0.99	0.99	0.99
Max t-Test Taylor	0.09	0.09	0.09	0.09	0.09	0.09	0.16	0.16	0.16	0.16	0.16	0.16

Rejection Frequencies, Experiment: 1, DGP: 2, Hyp: Null, $n = 200$, $J = 10000$, $\alpha = 0.05$

The Wald tests listed as BS1 and BS2 are the bootstrapped variants of Cheng's (2015) Wald test, and the comparison between these tests and the standard Wald test for the columns corresponding to $k_{\lambda,n} = 1$ tell the same story that Cheng (2015) tells in her paper. That is, the standard Wald test exhibits size distortions when weak identification is present, and accounting for weak identification adjusts the size of the test. The Max test variants also illustrate this same story.

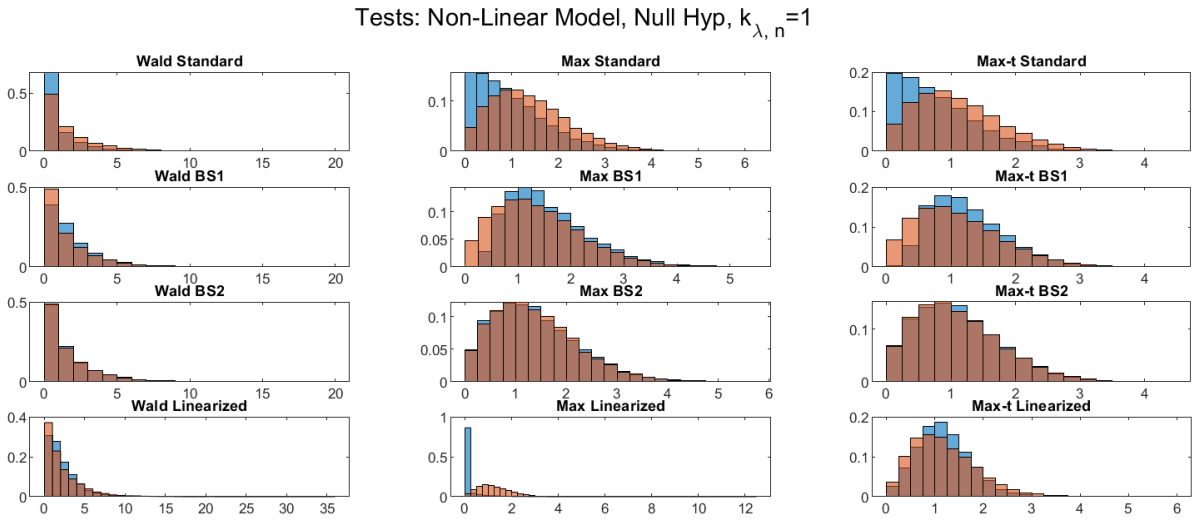
This table further illustrates the effect of the combination of weak identification and a parameter of large dimension on inference. When a large dimensional parameter is introduced, the Wald test begins to exhibit considerable size distortions. This is true even for the Wald tests that account for weak identification. The Max test, however, is able to accommodate testing the large dimensional parameter with a much lower, if any, size distortion.

Since weak identification is an issue with nonlinear models, we linearize these models with a first order Taylor expansion and test the corresponding parameters of interest in the linearized model. The rows, labeled Taylor, at the bottom of the table give the results for these tests. In the low dimensional model, the rejection frequencies indicate that the linearized tests do not alleviate the size distortions induced by weak identification to the same degree as the tests that accommodate

weak identification using the true model. We do not explore reasons for this, but we suspect that the first order expansion does not provide an adequate approximation to the true model. It would be interesting to examine if a higher order Taylor expansion paired with the Max test would provide an adequate work around for this issue,[16] but this is beyond the scope of this paper, so we leave it for future research.

The histograms shown below provide the simulated distribution of the test statistics for the case $\beta_1 = 1$. These histograms illustrate the story told above. In particular, it is evident that the standard tests are not able replicate the tail behavior of the test statistic when weak identification is present and the parameter dimension is large. The max test variants, however, are able to provide a much closer approximation to the tail behavior. Further, the final two tables demonstrate that the tests have non-trivial power against the local alternative design and power approaching one against the alternative hypothesis design.

Figure 3.1: Empirical Distribution of the Max Test



[16]Thanks to Eric Ghysels and Valentin Verdier for this suggestion.

Figure 3.2: Empirical Distribution of the Max Test

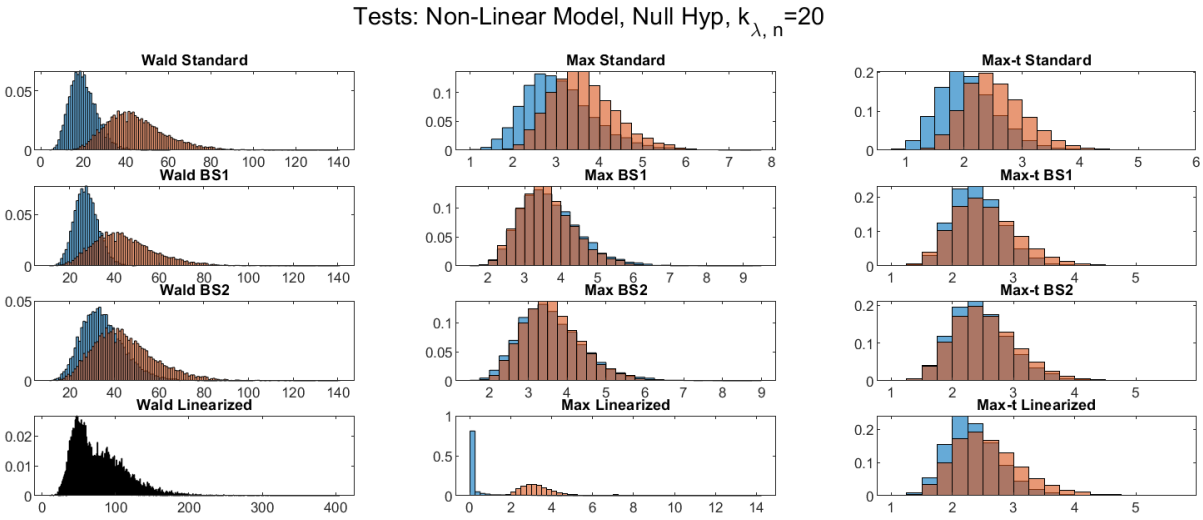


Table 3.3: Max Test Simulations - Additive Nonlinear Model under the Local Alternative Hypothesis

	$k_{\lambda,n}=1$						$k_{\lambda,n}=20$					
b_1	0	1	2	5	10	14	0	1	2	5	10	14
Wald Test Standard	0.21	0.22	0.20	0.22	0.21	0.21	0.91	0.91	0.90	0.91	0.91	0.91
Max Test Standard	0.19	0.20	0.20	0.20	0.19	0.19	0.25	0.25	0.25	0.26	0.25	0.25
Max t-Test Standard	0.20	0.20	0.21	0.21	0.20	0.21	0.50	0.50	0.51	0.50	0.50	0.50
Wald Test BS1	0.12	0.12	0.12	0.12	0.12	0.12	0.81	0.80	0.79	0.80	0.80	0.80
Max Test BS1	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.10
Max t-Test BS1	0.12	0.12	0.12	0.12	0.12	0.12	0.42	0.42	0.42	0.43	0.43	0.43
Wald Test BS2	0.13	0.13	0.13	0.13	0.13	0.13	0.39	0.39	0.39	0.39	0.39	0.39
Max Test BS2	0.11	0.11	0.12	0.12	0.12	0.11	0.11	0.10	0.11	0.11	0.11	0.11
Max t-Test BS2	0.13	0.13	0.13	0.13	0.13	0.13	0.34	0.33	0.33	0.33	0.33	0.33
Wald Test Taylor	0.15	0.15	0.15	0.15	0.15	0.15	0.61	0.61	0.61	0.61	0.61	0.61
Max Test Taylor	0.71	0.71	0.71	0.71	0.71	0.71	1.00	1.00	1.00	1.00	1.00	1.00
Max t-Test Taylor	0.15	0.15	0.15	0.15	0.15	0.15	0.37	0.37	0.37	0.37	0.37	0.37

Rejection Frequencies, Experiment: 1, DGP: 2, Hyp: Local Alternative, $n = 200$, $J = 10000$, $\alpha = 0.05$.

Table 3.4: Max Test Simulations - Additive Nonlinear Model under the Alternative Hypothesis

	$k_{\lambda,n} = 1$						$k_{\lambda,n} = 20$					
b_1	0	1	2	5	10	14	0	1	2	5	10	14
Wald Test Standard	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Max Test Standard	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Max t-Test Standard	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Wald Test BS1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Max Test BS1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Max t-Test BS1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Wald Test BS2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Max Test BS2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Max t-Test BS2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Wald Test Taylor	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Max Test Taylor	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Max t-Test Taylor	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Rejection Frequencies, Experiment: 1, DGP: 2, Hyp: Alternative, $n = 200$, $J = 10000$, $\alpha = 0.05$.

Table 3.5: Max Test Simulations - Additive Nonlinear Model

	Hyp: Null		Hyp: Local Alt	
b_1	0	22	0	22
Wald Test Standard	0.99	0.99	1.00	1.00
Max Test Standard	0.12	0.11	0.37	0.37
Max t-Test Standard	0.17	0.17	0.70	0.70
Wald Test BS1	0.90	0.90	0.96	0.96
Max Test BS1	0.04	0.04	0.17	0.17
Max t-Test BS1	0.09	0.09	0.59	0.59
Wald Test BS2	0.43	0.43	0.62	0.62
Max Test BS2	0.04	0.04	0.19	0.19
Max t-Test BS2	0.07	0.08	0.55	0.55
Wald Test Taylor	0.78	0.78	0.86	0.86
Max Test Taylor	0.40	0.40	0.76	0.76
Max t-Test Taylor	0.12	0.12	0.58	0.58

Rejection Frequencies, Experiment: 1, DGP: 2, $n = 500$, $J = 10000$, $\alpha = 0.05$, $k_{\lambda,n} = 50$.

3.8 Conclusion

Traditional Inference is distorted in the presence of large dimensional parameters and parameter identification failure. Previous work addresses these issues in isolation, but some economic questions require considering both of these issues jointly. We provide a testing framework that accommodates a large dimensional parameter when some of the parameter elements may be weakly

identified. The procedure is based on the maximum estimate in absolute value taken from a sequence of parameters that are estimated from carefully constructed sub-models and is implemented with a Gaussian multiplier bootstrap. Each sub-model is constructed by including one element from the parameter of interest. The test statistic is then formed from the maximum value of the estimates of the parameters of interest across these sub-models.

Simulations indicate that tests ignoring identification failure tend to over-reject the null hypothesis when the dimension of the parameter being tested is large, while the testing procedure that is designed to accommodate identification failure tends to control these rejection frequencies. Further, this testing procedure is able to reproduce existing results for the Wald test under weak identification when the dimension of the parameter being tested is small. Additionally, the Wald test of Cheng (2015), though designed to accommodate weakly identified parameters, tends to over-reject the null hypothesis when the dimension of the parameter being tested is large. However, the testing procedure presented here, based on the maximum estimated value, tends to better control empirical size in the presence of weak identification when the dimension of the parameter being tested is large.

APPENDIX A

APPENDIX FOR TESTING WHITE NOISE WHEN SOME PARAMETERS MAY BE WEAKLY IDENTIFIED

A.1 Appendix: Proofs of Main Results

A.1.1 Appendix: Proof of Lemma 2.3.1

Lemma. *2.3.1. Let Assumptions 3 - 11 hold. For some non-unique sequence of positive integers* $\{\mathcal{L}_n\}$ *with* $\mathcal{L}_n \to \infty$ *and* $\mathcal{L}_n = o(n)$,

(a) under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $||b|| < \infty$,

$$\left| \max_{1\leq h\leq \mathcal{L}_n} \sup_{\pi\in\Pi} (\sqrt{n}|\hat{\rho}_n(h;\pi) - \rho(h)|) - \max_{1\leq h\leq \mathcal{L}_n} \sup_{\pi\in\Pi} (|\mathcal{Z}_n^{\psi}(h,\pi)|) \right|$$
$$\leq \max_{1\leq h\leq \mathcal{L}_n} \sup_{\pi\in\Pi} (|\sqrt{n}(\hat{\rho}_n(h;\pi) - \rho(h)) - \mathcal{Z}_n^{\psi}(h,\pi)|) \xrightarrow{p} 0.$$

(b) under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,

$$\left| \max_{1\leq h\leq \mathcal{L}_n} (\sqrt{n}|\hat{\rho}_n(h) - \rho(h)|) - \max_{1\leq h\leq \mathcal{L}_n} (|\mathcal{Z}_n^{\theta}(h)|) \right| \leq \max_{1\leq h\leq \mathcal{L}_n} (|\sqrt{n}(\hat{\rho}_n(h) - \rho(h)) - \mathcal{Z}_n^{\theta}(h)|) \xrightarrow{p} 0.$$

Proof. Recall

$$r_t^{\theta}(h) = \frac{\varepsilon_t\varepsilon_{t-h} - E[\varepsilon_t\varepsilon_{t-h}] - \mathcal{D}^{\theta}(h)' J^{-1}(\gamma_0) m_t^{\theta}}{E[\varepsilon_t^2]}$$
$$r_t^{\psi,n}(h,\pi) = \frac{\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi) - E[\varepsilon_t\varepsilon_{t-h}] - \mathcal{D}(h,\pi)' H_n^{-1}(\pi;\gamma_0) m_t^{\psi}(\psi_{0,n},\pi)}{E[\varepsilon_t^2]}.$$

Define under strong and weak identification, respectively, $z_t^{\theta}(h) = r_t^{\theta}(h) - \rho(h) r_t^{\theta}(0)$ and $z_t^{\psi,n}(h,\pi) = r_t^{\psi,n}(h,\pi) - \rho(h) r_t^{\psi,n}(0,\pi)$.

Define $\mathcal{Z}_n^{\theta}(h) = \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t^{\theta}(h)$ and $\mathcal{Z}_n^{\psi}(h,\pi) = \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t^{\psi,n}(h,\pi)$.

Claim (a). We will prove for each h

$$\mathcal{X}_n(h) \equiv \sup_{\pi\in\Pi} |\sqrt{n}(\hat{\rho}_n(h;\pi) - \rho(h)) - \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t^{\psi,n}(h,\pi)| \xrightarrow{p} 0. \tag{A.1}$$

The claim will then follow from Lemma A.2 in Hill and Motegi (2018).

First observe that by Lemma A.2.4(a), for $h \geq 0$,

$$\begin{aligned}
&\sqrt{n}\Big(\frac{1}{n}\sum_{t=1+h}^{n}[\varepsilon_t(\hat{\psi}_n(\pi),\pi)\varepsilon_{t-h}(\hat{\psi}_n(\pi),\pi)] - E(\varepsilon_t\varepsilon_{t-h})\Big) \\
&\quad = \sqrt{n}\Big(\frac{1}{n}\sum_{t=1+h}^{n}\Big[\varepsilon_t\varepsilon_{t-h} - E(\varepsilon_t\varepsilon_{t-h})\Big]\Big) \\
&\qquad\qquad + \Big(H_n^{-1}(\psi_{0,n},\pi)\frac{1}{\sqrt{n}}\sum_{t=1}^{n} m_t^{\psi}(\psi_{0,n},\pi)\Big)'\mathcal{D}_n(h,\pi) \\
&\qquad\qquad + \sqrt{n}E_{\gamma_n}\Big[\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi) - \varepsilon_t\varepsilon_{t-h}\Big] + o_{p\pi}(1)
\end{aligned}$$

Next, define $\hat{R}_n(h,\hat{\psi}_n(\pi),\pi) = \frac{1}{n}\sum_{t=1+h}^{n}[\varepsilon_t(\hat{\psi}_n(\pi),\pi)\varepsilon_{t-h}(\hat{\psi}_n(\pi),\pi)]$ and $R(h) = E(\varepsilon_t\varepsilon_{t-h})$, and observe

$$\begin{aligned}
&\sqrt{n}(\hat{\rho}_n(h;\pi) - \rho(h)) \\
&\quad = \sqrt{n}(\frac{\hat{R}_n(h,\hat{\psi}_n(\pi),\pi)}{\hat{R}_n(0)} - \frac{R(h)}{R(0)}) \\
&\quad = \frac{\sqrt{n}(\hat{R}_n(h,\hat{\psi}_n(\pi),\pi) - R(h))}{\hat{R}_n(0,\hat{\psi}_n(\pi),\pi)} - \frac{R(h)}{\hat{R}_n(0,\hat{\psi}_n(\pi),\pi)R(0)}\sqrt{n}(\hat{R}_n(0,\hat{\psi}_n(\pi),\pi) - R(0)) \\
&\quad = \frac{\sqrt{n}(\hat{R}_n(h,\hat{\psi}_n(\pi),\pi) - R(h))}{R(0)}(1 + o_{p,\pi}(1)) \\
&\qquad\qquad - \frac{R(h)}{R(0)^2}\sqrt{n}(\hat{R}_n(0,\hat{\psi}_n(\pi),\pi) - R(0))(1 + o_{p,\pi}(1)) \\
&\quad = \Big(\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} r_t^{\psi,n}(h,\pi)\Big)(1 + o_{p,\pi}(1)) - \rho(h)\Big(\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} r_t^{\psi,n}(0,\pi)\Big)(1 + o_{p,\pi}(1)) \\
&\quad = \Big(\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}\big[r_t^{\psi,n}(h,\pi) - \rho(h)r_t^{\psi,n}(0,\pi)\big]\Big)(1 + o_{p,\pi}(1)) \\
&\quad = \frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}\big[r_t^{\psi,n}(h,\pi) - \rho(h)r_t^{\psi,n}(0,\pi)\big] + o_{p,\pi}(1)
\end{aligned}$$

where the last equality follows from Theorem 17.5 in Davidson (1994) and Theorem 1.6 in McLeish (1975).

Claim (b). We will prove for each h

$$\mathcal{X}_n(h) \equiv |\sqrt{n}(\hat{\rho}_n(h) - \rho(h)) - \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t^{\theta}(h)| \xrightarrow{p} 0. \tag{A.2}$$

The claim will then follow from Lemma A.2 in Hill and Motegi (2018).

First observe that by Lemma A.2.4(b), for $h \geq 0$,

$$\begin{aligned}
&\sqrt{n}\Big(\frac{1}{n}\sum_{t=1+h}^{n}[\varepsilon_t(\hat{\theta}_n)\varepsilon_{t-h}(\hat{\theta}_n)] - E(\varepsilon_t\varepsilon_{t-h})\Big) \\
&\quad = \Big(\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}\Big[\varepsilon_t(\theta_n)\varepsilon_{t-h}(\theta_n) - E(\varepsilon_t\varepsilon_{t-h})\Big]\Big) \\
&\qquad\qquad + \Big(J_n^{-1}(\theta_n)\frac{1}{\sqrt{n}}\sum_{t=1}^{n} m_t^{\theta}(\theta_n)\Big)' B^{-1}(\beta_n)\mathcal{D}_n^{\theta}(h) + o_p(1) \\
&\quad = \frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}\Big(\varepsilon_t\varepsilon_{t-h} - E[\varepsilon_t\varepsilon_{t-h}] + \big[J_n^{-1}(\gamma_0)m_t^{\theta}(\theta_n)\big]' B^{-1}(\beta_n)\mathcal{D}_n^{\theta}(h)\Big) + o_p(1)
\end{aligned}$$

Next, define $\hat{R}_n(h) = \frac{1}{n}\sum_{t=1+h}^{n}[\varepsilon_t(\hat{\theta}_n)\varepsilon_{t-h}(\hat{\theta}_n)]$ and $R(h) = E(\varepsilon_t\varepsilon_{t-h})$, and observe

$$\begin{aligned}
&\sqrt{n}(\hat{\rho}_n(h) - \rho(h)) \\
&\quad = \sqrt{n}(\frac{\hat{R}_n(h)}{\hat{R}_n(0)} - \frac{R(h)}{R(0)}) \\
&\quad = \frac{\sqrt{n}(\hat{R}_n(h) - R(h))}{\hat{R}_n(0)} - \frac{R(h)}{\hat{R}_n(0)R(0)}\sqrt{n}(\hat{R}_n(0) - R(0)) \\
&\quad = \frac{\sqrt{n}(\hat{R}_n(h) - R(h))}{R(0)}(1 + o_p(1)) - \frac{R(h)}{R(0)^2}\sqrt{n}(\hat{R}_n(0) - R(0))(1 + o_p(1)) \\
&\quad = \Big(\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} r_t^{\theta}(h)\Big)(1 + o_p(1)) - \rho(h)\Big(\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} r_t^{\theta}(0)\Big)(1 + o_p(1)) \\
&\quad = \Big(\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}\big[r_t^{\theta}(h) - \rho(h)r_t^{\theta}(0)\big]\Big)(1 + o_p(1)) \\
&\quad = \frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}\big[r_t^{\theta}(h) - \rho(h)r_t^{\theta}(0)\big] + o_p(1)
\end{aligned}$$

where the last equality follows from Theorem 17.5 in Davidson (1994) and Theorem 1.6 in

McLeish (1975).

□

A.1.2 Appendix: Lemma A.1.2

Lemma A.1.2. *(a) Let Assumptions 3, 7, 8, 1, 4(i), 5, 9, and 11(i) hold, and let* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $||b|| < \infty$. *Let* $\{\mathcal{Z}^{\psi}(h,\pi) : h \in \mathbb{N}, \pi \in \Pi\}$ *be a Gaussian process with mean* $\lim_{n\to\infty} z_s^{2,\psi,n}(h,\pi) < \infty$ *and variance* $\lim_{n\to\infty} \frac{1}{n}\sum_{s,t=1}^{n} E[z_s^{1,\psi,n}(h,\pi) z_t^{1,\psi,n}(h,\pi)] < \infty$ *and covariance kernel* $\lim_{n\to\infty} \frac{1}{n}\sum_{s,t=1}^{n} E[z_s^{1,\psi,n}(h,\pi)\ z_t^{1,\psi,n}(\tilde{h},\tilde{\pi})]$. *Then for some non-unique sequence of positive integers* $\{\mathcal{L}_n\}$ *with* $\mathcal{L}_n \to \infty$ *and* $\mathcal{L}_n = o(n)$,

$$\Big| \max_{1\le h\le \mathcal{L}_n} |\mathcal{Z}_n^{\psi}(h,\hat{\pi}_n)| - \max_{1\le h\le \mathcal{L}_n} |\mathcal{Z}^{\psi}(h,\pi^*(b,\gamma_0))| \Big| \le \max_{1\le h\le \mathcal{L}_n} |\mathcal{Z}_n^{\psi}(h,\hat{\pi}_n) - \mathcal{Z}^{\psi}(h,\pi^*(b,\gamma_0))| \xrightarrow{p} 0.$$

(b) Let Assumptions 3, 7, 8, 2, 4(ii), 6, 10, and 11(ii) hold, and let $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$. *Let* $\{\mathcal{Z}^{\theta}(h) : h \in \mathbb{N}\}$ *be a zero mean Gaussian process with variance* $\lim_{n\to\infty} \frac{1}{n}\sum_{s,t=1}^{n} E[z_s^{\theta}(h) z_t^{\theta}(h)] < \infty$ *and covariance kernel* $\lim_{n\to\infty} \frac{1}{n}\sum_{s,t=1}^{n} E[z_s^{\theta}(h) z_t^{\theta}(\tilde{h})]$. *Then for some non-unique sequence of positive integers* $\{\mathcal{L}_n\}$ *with* $\mathcal{L}_n \to \infty$ *and* $\mathcal{L}_n = o(n)$,

$$\Big| \max_{1\le h\le \mathcal{L}_n} |\mathcal{Z}_n^{\theta}(h)| - \max_{1\le h\le \mathcal{L}_n} |\mathcal{Z}^{\theta}(h)| \Big| \le \max_{1\le h\le \mathcal{L}_n} |\mathcal{Z}_n^{\theta}(h) - \mathcal{Z}^{\theta}(h)| \xrightarrow{p} 0.$$

Proof. The proof of claim (b) follows similarly to the proof of Lemma 2.2 in Hill and Motegi (2018). For part (a), recall that $\pi^*(b,\gamma_0)$ in (a) is a random variable, so the proof of claim (a) requires more steps. First, we must prove weak convergence; then, we must show joint convergence of $\hat{\pi}_n$ and $\mathcal{Z}_n^{\psi}(h,\pi)$.

(a) We prove for each $\mathcal{L} \in \mathbb{N}$

$$\{\mathcal{Z}_n^{\psi}(h,\pi) :\ 1 \le h \le \mathcal{L},\ \pi \in \Pi\} \Rightarrow \{\mathcal{Z}^{\psi}(h,\pi) :\ 1 \le h \le \mathcal{L},\ \pi \in \Pi\} \tag{A.3}$$

$$\{\mathcal{Z}_n^{\psi}(h,\hat{\pi}_n) :\ 1 \le h \le \mathcal{L}\} \xrightarrow{d} \{\mathcal{Z}^{\psi}(h,\pi^*) :\ 1 \le h \le \mathcal{L}\} \tag{A.4}$$

In the first step, we establish weak convergence over h and π; this involves the finite dimensional convergence and stochastic equicontinuity of $\mathcal{Z}_n^{\psi}(h,\pi)$. The second step then will follow from the joint convergence of $\hat{\psi}_n$ and $\hat{\pi}_n$. We first split $\mathcal{Z}_n^{\psi}$ into a mean zero component and a component that converges in probability uniformly over Π.

Recall that $z_t^{\psi,n}(h,\pi) = r_t^{\psi,n}(h,\pi) - \rho(h) r_t^{\psi,n}(0,\pi)$ and

$$r_t^{\psi,n}(h,\pi) = \frac{\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi) - E[\varepsilon_t\varepsilon_{t-h}] - \mathcal{D}(h,\pi)' H_n^{-1}(\pi;\gamma_0) m_t^{\psi}(\psi_{0,n},\pi)}{E[\varepsilon_t^2]}.$$

Observe that by the mean value theorem, for some $\tilde{\gamma}_n$ such that $||\tilde{\gamma}_n - \gamma_n|| \leq ||\gamma_{0,n} - \gamma_n||$,

$$\begin{aligned}
m_t^{\psi}(\psi_{0,n},\pi) &= m_t^{\psi}(\psi_{0,n},\pi) - E_{\gamma_n}[m_t^{\psi}(\psi_{0,n},\pi)] + E_{\gamma_n}[m_t^{\psi}(\psi_{0,n},\pi)] \\
&= m_t^{\psi}(\psi_{0,n},\pi) - E_{\gamma_n}[m_t^{\psi}(\psi_{0,n},\pi)] + E_{\gamma_{0,n}}[m_t^{\psi}(\psi_{0,n},\pi)] \\
&\qquad + \beta_n \frac{\partial}{\partial \tilde{\beta}} E_{\tilde{\gamma}_n}[m_t^{\psi}(\psi_{0,n},\pi)] \\
&= m_t^{\psi}(\psi_{0,n},\pi) - E_{\gamma_n}[m_t^{\psi}(\psi_{0,n},\pi)] + \beta_n \frac{\partial}{\partial \tilde{\beta}} E_{\tilde{\gamma}_n}[m_t^{\psi}(\psi_{0,n},\pi)].
\end{aligned}$$

Further, add and subtract $\varepsilon_t\varepsilon_{t-h}$, and observe

$$\begin{aligned}
r_t^{\psi,n}(h,\pi) &= \frac{\varepsilon_t\varepsilon_{t-h} - E[\varepsilon_t\varepsilon_{t-h}]}{E[\varepsilon_t^2]} - \frac{\mathcal{D}(h,\pi)' H^{-1}(\pi;\gamma_0)\big(m_t^{\psi}(\psi_{0,n},\pi) - E_{\gamma_n}[m_t^{\psi}(\psi_{0,n},\pi)]\big)}{E[\varepsilon_t^2]} \\
&\qquad + \frac{\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi) - \varepsilon_t\varepsilon_{t-h}}{E[\varepsilon_t^2]} \\
&\qquad - \frac{\mathcal{D}(h,\pi)' H^{-1}(\pi;\gamma_0)\big(\beta_n \frac{\partial}{\partial\tilde{\beta}} E_{\tilde{\gamma}_n}[m_t^{\psi}(\psi_{0,n},\pi)]\big)}{E[\varepsilon_t^2]} \\
&= r_t^{1,\psi,n}(h,\pi) + r_t^{2,\psi,n}(h,\pi)
\end{aligned}$$

where $r_t^{1,\psi,n}(h,\pi) = \frac{\varepsilon_t\varepsilon_{t-h} - E[\varepsilon_t\varepsilon_{t-h}]}{E[\varepsilon_t^2]} - \frac{\mathcal{D}(h,\pi)' H^{-1}(\pi;\gamma_0)\left(m_t^{\psi}(\psi_{0,n},\pi) - E_{\gamma_n}[m_t^{\psi}(\psi_{0,n},\pi)]\right)}{E[\varepsilon_t^2]}$ and $r_t^{2,\psi,n}(h,\pi) = \frac{\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi) - \varepsilon_t\varepsilon_{t-h}}{E[\varepsilon_t^2]} - \frac{\mathcal{D}(h,\pi)' H^{-1}(\pi;\gamma_0)\left(\beta_n \frac{\partial}{\partial\tilde{\beta}} E_{\tilde{\gamma}_n}[m_t^{\psi}(\psi_{0,n},\pi)]\right)}{E[\varepsilon_t^2]}$.

For $i = 1, 2$, define $z_t^{i,\psi,n}(h,\pi) = r_t^{i,\psi,n}(h,\pi) - \rho(h) r_t^{i,\psi,n}(0,\pi)$, and observe that $z_t^{\psi,n}(h,\pi) = z_t^{1,\psi,n}(h,\pi) + z_t^{2,\psi,n}(h,\pi)$.

Now define $\mathcal{Z}_n^{i,\psi}(h,\pi) = \frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t^{i,\psi,n}(h,\pi)$ for $i = 1, 2$. We show that $\mathcal{Z}_n^{1,\psi}(h,\pi)$ converges weakly to a Gaussian process and $\mathcal{Z}_n^{2,\psi}(h,\pi)$ converges uniformly in probability to a mean component.

Let $\mathcal{L}, K \in \mathbb{N}$ be arbitrary, and take $[\lambda_h]_{h=1}^{\mathcal{L}} = \lambda \in \mathbb{R}^{\mathcal{L}}$ and $a \in \mathbb{R}^K$ with $\lambda'\lambda = 1$ and $a'a = 1$. Take $\{\pi_1, \ldots, \pi_K\} \in \Pi^{\otimes K}$, and observe

$$\begin{aligned}
\sum_{h=1}^{\mathcal{L}}\sum_{k=1}^{K} \lambda_h a_k \mathcal{Z}_n^{1,\psi}(h,\pi_k) &= \frac{1}{\sqrt{n}}\sum_{h=1}^{\mathcal{L}}\sum_{k=1}^{K}\lambda_h a_k \sum_{t=1+h}^{n} z_t^{1,\psi,n}(h,\pi_k) \\
&= \frac{1}{\sqrt{n}}\sum_{h=1}^{\mathcal{L}}\sum_{k=1}^{K}\lambda_h a_k \sum_{t=1}^{n} z_t^{1,\psi,n}(h,\pi_k)1(1+h \leq t \leq n) \\
&= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\sum_{h=1}^{\mathcal{L}}\sum_{k=1}^{K}\lambda_h a_k z_t^{1,\psi,n}(h,\pi_k)1(1+h \leq t \leq n) \\
&= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\lambda \otimes a)' z_{t,\mathcal{L},K}^{1,\psi,n}
\end{aligned}$$

where $z_{t,\mathcal{L},K}^{1,\psi,n} = [z_t^{1,\psi,n}(h,\pi_k)1(1+h \leq t \leq n)]_{\substack{h=1,\ldots,\mathcal{L} \\ k=1,\ldots,K}}$. Next, define the quantity $\sigma^2(\lambda,a) = E\big(\sum_{h=1}^{\mathcal{L}}\sum_{k=1}^{K}\lambda_h a_k \mathcal{Z}_n^{1,\psi}(h,\pi_k)\big)^2$. We must show $\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\lambda\otimes a)' z_{t,\mathcal{L},K}^{1,\psi,n} \xrightarrow{d} N(0, \lim_{n\to\infty}\sigma^2(\lambda,a))$. Then finite dimensional convergence follows from the Cramér-Wold theorem.

Next, by Theorems 17.8 and 17.9 in Davidson (1994), $(\lambda\otimes a)' z_{t,\mathcal{L},K}^{1,\psi,n}$ is mean zero, stationary, L_p-bounded for some $p > 2$, and L_2-NED with size $-1/2$ on an α-mixing base with decay rate $O(h^{-p/(p-2)-\iota})$. Thus, $\sigma_n^2(\lambda,a) = O(1)$ by McLeish (1975), and $\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\lambda\otimes a)' z_{t,\mathcal{L},K}^{1,\psi,n} \xrightarrow{d} N(0, \lim_{n\to\infty}\sigma^2(\lambda,a))$ by Theorem 2 in de Jong (1997). This established finite dimensional convergence.

Next, we show uniform convergence in probability of $\mathcal{Z}_n^{2,\psi}(h,\pi_k)$. Then we show stochastic equicontinuity of $\mathcal{Z}_n^{\psi}(h,\pi_k)$.

Recall $\mathcal{Z}_n^{2,\psi}(h,\pi) = \frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t^{2,\psi,n}(h,\pi)$, $z_t^{2,\psi,n}(h,\pi) = r_t^{2,\psi,n}(h,\pi) - \rho(h) r_t^{2,\psi,n}(0,\pi)$, and $r_t^{2,\psi,n}(h,\pi) = \frac{\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi) - \varepsilon_t\varepsilon_{t-h}}{E[\varepsilon_t^2]} - \frac{\mathcal{D}(h,\pi)' H^{-1}(\pi;\gamma_0)\left(\beta_n \frac{\partial}{\partial\tilde{\beta}} E_{\tilde{\gamma}_n}[m_t^{\psi}(\psi_{0,n},\pi)]\right)}{E[\varepsilon_t^2]}$. It is sufficient then,

to show uniform convergence in probability of $\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} r_t^{2,\psi,n}(h,\pi)$ for all $h \geq 0$. Observe

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} r_t^{2,\psi,n}(h,\pi) \\
&= \frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}\left(\frac{\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h}}{E[\varepsilon_t^2]}\right. \\
&\qquad\left. - \frac{\mathcal{D}(h,\pi)'H^{-1}(\pi;\gamma_0)\left(\beta_n\frac{\partial}{\partial\tilde{\beta}}E_{\tilde{\gamma}_n}[m_t^{\psi}(\psi_{0,n},\pi)]\right)}{E[\varepsilon_t^2]}\right) \\
&= \frac{\sqrt{n}}{n}\sum_{t=1+h}^{n}\frac{\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h}}{E[\varepsilon_t^2]} \\
&\qquad - \frac{\mathcal{D}(h,\pi)'H^{-1}(\pi;\gamma_0)\left(\sqrt{n}\beta_n\frac{1}{n}\sum_{t=1+h}^{n}\frac{\partial}{\partial\tilde{\beta}}E_{\tilde{\gamma}_n}[m_t^{\psi}(\psi_{0,n},\pi)]\right)}{E[\varepsilon_t^2]}
\end{aligned}
$$

Consider the first term and $h \geq 1$ ($h = 0$ follows similarly).

$$
\begin{aligned}
&\frac{\sqrt{n}}{n}\sum_{t=1+h}^{n}(\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h}) \\
&= \frac{\sqrt{n}}{n}\sum_{t=1}^{n}(\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h}) - \frac{\sqrt{n}}{n}\sum_{t=1}^{h}(\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h}) \\
&\xrightarrow{p} \lim_{n\to\infty}\sqrt{n}E_{\gamma_n}[\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h}]
\end{aligned}
$$

since $\varepsilon_t(\psi_{0,n},\pi)$ does not depend on π and $\limsup_{n\to\infty}\sqrt{n}E_{\gamma_n}[\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h}] = O(1)$ from Assumption 9 and Assumption 8.

Next, observe that for the second term

$$
\begin{aligned}
&\mathcal{D}(h,\pi)'H^{-1}(\pi;\gamma_0)\sqrt{n}\beta_n\frac{1}{n}\sum_{t=1+h}^{n}\frac{\partial}{\partial\tilde{\beta}}E_{\tilde{\gamma}_n}[m_t^{\psi}(\psi_{0,n},\pi)] \\
&= \mathcal{D}(h,\pi)'H^{-1}(\pi;\gamma_0)\sqrt{n}\beta_n K_n(\psi_{0,n},\pi;\tilde{\gamma}_n) + o(1).
\end{aligned}
$$

Then $\sqrt{n}\beta_n \to b$ and $K_n(\psi_{0,n},\pi;\tilde{\gamma}_n) \xrightarrow{p} K_n(\psi_0,\pi;\gamma_0)$ uniformly on Π (Assumption 1). This establishes uniform convergence in probability of $\mathcal{Z}_n^{2,\psi}(h,\pi_k)$ on Π.

To establish stochastic equicontinuity of $\mathcal{Z}_n^{\psi}(h,\pi)$, first observe that $\{1,\ldots,\mathcal{L}\}$ is compact.

Next, recall $\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)$ does not depend on π under Assumption 3. Further, $\hat{H}_n(\pi;\gamma_n)$, $\hat{K}_n(\pi;\tilde{\gamma}_n)$, and $\hat{D}_n(h,\pi)$ each converge uniformly in probability to the respective limits $H(\pi;\gamma_0)$, $K(\pi;\gamma_0)$, and $\mathcal{D}(h,\pi)$. Thus, in order to establish stochastic equicontinuity of $\mathcal{Z}_n^{\psi}(h,\pi)$, we only need stochastic equicontinuity of $m_t^{\psi}(\psi_{0,n},\pi)$, which is ensured by Assumption 4, and to invoke probability sub-additivity. This establishes A.3.

Now in order to show A.4, we only need to show joint convergence of $\mathcal{Z}_n^{\psi}(h,\pi)$ and $\hat{\pi}_n$. The latter joint convergence occurs because $\hat{\pi}_n$ can be written as a continuous function of $H_n(\pi;\gamma_n)$, $K_n(\psi_{0,n},\pi;\tilde{\gamma}_n)$, and $G_n(\psi_{0,n},\pi;\gamma_n)$.[1]

Finally, A.4 implies $\mathcal{Z}_n^{\psi}(h,\hat{\pi}_n)-\mathcal{Z}^{\psi}(h,\pi^*(b,\gamma_0))=o_p(1)$ for each h. The result follows from (Hill and Motegi, 2018, Lemma A.2).

(b) Recall $z_t^{\theta}(h)=r_t^{\theta}(h)-\rho(h)r_t^{\theta}(0)$ where

$$r_t^{\theta}(h)=\frac{\varepsilon_t\varepsilon_{t-h}-E[\varepsilon_t\varepsilon_{t-h}]-(\mathcal{D}^{\theta}(h))'J^{-1}(\gamma_0)m_t^{\theta}}{E[\varepsilon_t^2]}$$

We prove for each $\mathcal{L}\in\mathbb{N}$

$$\{\mathcal{Z}_n^{\theta}(h):\ 1\leq h\leq\mathcal{L}\}\xrightarrow{d}\{\mathcal{Z}^{\theta}(h):\ 1\leq h\leq\mathcal{L}\} \tag{A.5}$$

where $\{\mathcal{Z}^{\theta}(h):h\in\mathbb{N}\}$ is a zero mean Gaussian process with variance $\lim_{n\to\infty}\frac{1}{n}\sum_{s,t=1}^{n}E[z_s^{\theta}(h)z_t^{\theta}(h)]<\infty$ and covariance kernel $\lim_{n\to\infty}\frac{1}{n}\sum_{s,t=1}^{n}E[z_s^{\theta}(h)z_t^{\theta}(\tilde{h})]$. Let $\mathcal{L},K\in\mathbb{N}$ be arbitrary, and take $[\lambda_h]_{h=1}^{\mathcal{L}}=\lambda\in\mathbb{R}^{\mathcal{L}}$ with $\lambda'\lambda=1$. Observe

$$\begin{aligned}\sum_{h=1}^{\mathcal{L}}\lambda_h\mathcal{Z}_n^{\theta}(h)&=\frac{1}{\sqrt{n}}\sum_{h=1}^{\mathcal{L}}\lambda_h\sum_{t=1+h}^{n}z_t^{\theta}(h)\\&=\frac{1}{\sqrt{n}}\sum_{h=1}^{\mathcal{L}}\lambda_h\sum_{t=1}^{n}z_t^{\theta}(h)1(1+h\leq t\leq n)\\&=\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\sum_{h=1}^{\mathcal{L}}\lambda_h z_t^{\theta}(h)1(1+h\leq t\leq n)=\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\lambda' z_{t,\mathcal{L}}^{\theta}\end{aligned}$$

[1] See Andrews and Cheng (2012b), proof of theorem 3.1, page 25.

where $z^{\theta}_{t,\mathcal{L}} = [z^{\theta}_t(h)1(1+h \leq t \leq n)]_{h=1,\ldots,\mathcal{L}}$. Define $\sigma^2(\lambda) = E\big(\sum_{h=1}^{\mathcal{L}} \lambda_h \mathcal{Z}^{\theta}_n(h)\big)^2$. We show $\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \lambda' z^{\theta}_{t,\mathcal{L}} \xrightarrow{d} N(0, \lim_{n\to\infty}\sigma^2(\lambda))$. Then A.5 follows from the Cramér-Wold theorem.

Next, by Theorems 17.8 and 17.9 in Davidson (1994), $(\lambda \otimes a)' z^{\psi,n}_{t,\mathcal{L},K}$ is mean zero, stationary, L_p-bounded for some $p > 2$, and L_2-NED with size $-1/2$ on an α-mixing base with decay rate $O(h^{-p/(p-2)-\iota})$. Thus, $\sigma^2_n(\lambda) = O(1)$ by McLeish (1975), and $\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \lambda' z^{\psi,n}_{t,\mathcal{L},K} \xrightarrow{d} N(0, \lim_{n\to\infty}\sigma^2(\lambda))$ by Theorem 2 in de Jong (1997).

Finally, A.5 implies $\mathcal{Z}^{\theta}_n(h) - \mathcal{Z}^{\theta}(h) = o_p(1)$ for each h, so the result follows from (Hill and Motegi, 2018, Lemma A.2).

□

A.1.3 Proof of Theorem 2.4.1

Theorem. *2.4.1. Let Assumptions 1 - 11 hold, and let the number of bootstrap samples $M_n \to \infty$.*

(a) Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$, there is a non-unique sequence of positive integers $\{\mathcal{L}_n\}$ with $\mathcal{L}_n \to \infty$ and $\mathcal{L}_n = o(n)$ such that $|\hat{c}^{(w)}_{1-\alpha,n} - c_{n,1-\alpha}| \xrightarrow{p} 0$.

(b) Under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, there is a non-unique sequence of positive integers $\{\mathcal{L}_n\}$ with $\mathcal{L}_n \to \infty$ and $\mathcal{L}_n = o(n)$ such that $|\hat{c}^{(s)}_{1-\alpha,n} - c_{n,1-\alpha}| \xrightarrow{p} 0$.

Moreover, under the alternative hypothesis, $P(\hat{T}_n > \hat{c}^{(k)}_{1-\alpha,n}) \to 1$ for $k = w, s$.

Proof. Since it is considerably shorter, we first prove the claim for case (b), strong identification. The proof follows the proof of Theorem 2.5 in Hill and Motegi (2018) very closely. We rely on the notion of weak convergence in probability, written $\Rightarrow^p$, on the space of bounded functions, l_∞, as defined in Giné and Zinn (1990).

(b) Strong Identification. Let $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$. Define the sample $\mathcal{W}_n \equiv \{m_t, x_t, y_t\}_{t=1}^{n}$.

We prove the following two steps:

$$\{\sqrt{n}\hat{\rho}^{(s)}_n(h) : \ 1 \leq h \leq \mathcal{L}\} \Rightarrow^p \{\overset{\circ}{\mathcal{Z}}(h) : \ 1 \leq h \leq \mathcal{L}\} \tag{A.6}$$

for each $\mathcal{L} \in \mathbb{N}$, where $\{\overset{\circ}{\mathcal{Z}}(h) : h \in \mathbb{N}\}$ is an independent copy of $\{\mathcal{Z}^{\theta}(h) : h \in \mathbb{N}\}$, the zero mean Gaussian process in Lemma 3.2. For the process $\{\overset{\circ}{\mathcal{Z}}(h) : \ 1 \leq h \leq \mathcal{L}\}$ and some sequence

of positive integers $\{\mathcal{L}_n\}$, $\mathcal{L}_n \to \infty$ and $\mathcal{L}_n = o(n)$,

$$\sup_{c>0} \left| P(\max_{1\leq h\leq \mathcal{L}_n} |\sqrt{n}\hat{\rho}_n^{(s)}(h)| \leq c|\mathcal{X}_n) - P(\max_{1\leq h\leq \mathcal{L}_n} |\overset{\circ}{\mathcal{Z}}(h)| \leq c) \right| \xrightarrow{p} 0 \tag{A.7}$$

Let $\{z_t\}_{t=1}^n$ be a draw of the auxiliary variables, and recall

$$\hat{\mathcal{E}}_{t,h}(\theta) = \varepsilon_t(\theta)\varepsilon_{t-h}(\theta) - (B(\hat{\beta}_n)^{-1}\hat{\mathcal{D}}_n^{\theta}(h,\theta))'(\hat{J}_n(\hat{\theta}_n))^{-1}m_t^{\theta}(\theta)$$

$$\hat{\rho}_n^{(s)}(h) = \frac{1}{n^{-1}\sum_{t=1}^n \varepsilon_t^2(\hat{\theta}_n)} \times \left\{ \frac{1}{n}\sum_{t=1+h}^n z_t\Big(\hat{\mathcal{E}}_{t,h}(\hat{\theta}_n) - \frac{1}{n}\sum_{t=1+h}^n \hat{\mathcal{E}}_{t,h}(\hat{\theta}_n)\Big)\right\}.$$

Define

$$\rho_n^*(h) = \frac{1}{E(\varepsilon_t^2)} \times \left\{ \frac{1}{n}\sum_{t=1+h}^n z_t\Big(\mathcal{E}_{t,h} - E(\mathcal{E}_{t,h})\Big)\right\}$$

$$\mathcal{E}_{t,h} = \varepsilon_t\varepsilon_{t-h} - \mathcal{D}^{\theta}(h)'J^{-1}m_t^{\theta}.$$

We prove A.6 with the following two steps:

$$\{\sqrt{n}\rho_n^*(h) : 1 \leq h \leq \mathcal{L}\} \Rightarrow^p \{\overset{\circ}{\mathcal{Z}}(h) : \ 1 \leq h \leq \mathcal{L}\} \tag{A.8}$$

$$\sqrt{n}|\hat{\rho}_n^{(s)}(h) - \rho_n^*(h)| \xrightarrow{p} 0 \text{ for each } h \tag{A.9}$$

where $\{\overset{\circ}{\mathcal{Z}}(h) : h \in \mathbb{N}\}$ is an independent copy of $\{\mathcal{Z}^{\theta}(h) : h \in \mathbb{N}\}$.

In the general case, Shao (2011a) requires the sub-auxiliary variables $\{\xi_t\}_{t=1}^{n/b_n}$, which are used to construct the auxiliary variables z_t, to be iid and satisfy $E(\xi_t) = 0$, $E(\xi_t^2) = 1$, and $E(\xi_t^4) < \infty$. Following Hill and Motegi (2018), we shorten the proof by letting ξ_t be iid $N(0,1)$ random variables, which eliminates the extra steps needed to show asymptotic convergence in conditional distribution.

In order to prove A.8, we prove weak convergence in the sense of Hoffmann-Jorgensen (1984, 1991). This requires a totally bounded pseudo metric space, finite dimensional convergence, and stochastic equicontinuity. The proof of this step follows exactly the proof of Lemma A.3, step 1 in

Hill and Motegi (2018).

Observe that $\{1, \ldots, \mathcal{L}\}$ is compact, so the space $\{1, \ldots, \mathcal{L}\}$ with the sup-norm is totally bounded. The distributions governing $\{\sqrt{n}\rho_n^*(h) : n \geq 1\}$ are stochastically equicontinuous on $\{1, \ldots, \mathcal{L}\}$ because the latter is discrete and bounded. Finally, we prove finite dimensional distributions in the following.

We operate conditionally on the sample $\mathcal{W}_n$. Write

$$\rho_n^*(h) = \frac{1}{E(\varepsilon_t^2)} \times \frac{1}{n/b_n} \sum_{s=1}^{n/b_n} \xi_s \left\{ \frac{1}{b_n} \sum_{t=(s-1)b_n+1+h}^{sb_n} \Big(\mathcal{E}_{t,h} - E(\mathcal{E}_{t,h}) \Big) \right\}$$

By joint Gaussianity and independence of ξ_s, $\{\sqrt{n}\rho_n^*(h) : 1 \leq h \leq \mathcal{L}\}$ is a zero mean Gaussian process with covariance function

$$\begin{aligned}
&nE(\rho_n^*(h)\rho_n^*(\tilde{h})|\mathcal{W}_n) \\
&= \frac{1}{\big(E(\varepsilon_t^2)\big)^2} \times \frac{1}{n} \sum_{s=1}^{n/b_n} \left\{ \sum_{t=(s-1)b_n+1+h}^{sb_n} \Big(\mathcal{E}_{t,h} - E(\mathcal{E}_{t,h}) \Big) \right\} \left\{ \sum_{t=(s-1)b_n+1+\tilde{h}}^{sb_n} \Big(\mathcal{E}_{t,\tilde{h}} - E(\mathcal{E}_{t,\tilde{h}}) \Big) \right\}
\end{aligned}$$

for each $\mathcal{L} \in \mathbb{N}$. Observe

$$\begin{aligned}
&\lim_{n\to\infty} E\big[nE(\rho_n^*(h)\rho_n^*(\tilde{h})|\mathcal{W}_n)\big] \\
&= \frac{1}{\big(E(\varepsilon_t^2)\big)^2} \times \lim_{n\to\infty} \frac{1}{n} \sum_{s=1}^{n/b_n} \sum_{t=(s-1)b_n+1+h}^{sb_n} \sum_{u=(s-1)b_n+1+h}^{sb_n} E\Big[\Big(\mathcal{E}_{t,h} - E(\mathcal{E}_{t,h}) \Big) \Big(\mathcal{E}_{u,\tilde{h}} - E(\mathcal{E}_{u,\tilde{h}}) \Big) \Big] \\
&= \lim_{n\to\infty} \frac{1}{n} E\left[\sum_{t=1}^{n} \Big(\frac{\mathcal{E}_{t,h} - E(\mathcal{E}_{t,h})}{E(\varepsilon_t^2)} \Big) \sum_{t=1}^{n} \Big(\frac{\mathcal{E}_{t,\tilde{h}} - E(\mathcal{E}_{t,\tilde{h}})}{E(\varepsilon_t^2)} \Big) \right] \\
&= E[\mathcal{Z}^\theta(h)\mathcal{Z}^\theta(\tilde{h})]
\end{aligned}$$

where the final equality follows from the definition of $\mathcal{Z}(h)$ in Lemma A.1.2.

Let $\mathcal{W}$ be the set of samples such that

$$nE(\rho_n^*(h)\rho_n^*(\tilde{h})|\mathcal{W}_n) \xrightarrow{p} E[\mathcal{Z}^\theta(h)\mathcal{Z}^\theta(\tilde{h})].$$

We will show that $P(\mathcal{W}_n \in \mathcal{W}) = 1$. This will show that the finite dimensional distributions of $\{\sqrt{n}\rho_n^*(h) : 1 \leq h \leq \mathcal{L}\}$ converge to the zero mean Gaussian process $\{\overset{\circ}{\mathcal{Z}}(h) : 1 \leq h \leq \mathcal{L}\}$ with covariance function $E[\mathcal{Z}^\theta(h)\mathcal{Z}^\theta(\tilde{h})]$, where the independence of the ξ_s and Gaussianity imply that $\{\overset{\circ}{\mathcal{Z}}(h) : 1 \leq h \leq \mathcal{L}\}$ is an independent copy of $\{\mathcal{Z}^\theta(h) : 1 \leq h \leq \mathcal{L}\}$.

The following step in this argument follows verbatim from the proof of Lemma A.3(a), step 1, in Hill and Motegi (2018), which utilizes arguments presented in de Jong (1997), specifically Theorem 2. Let $\{l_n\}$ be a sequence of integers with $l_n \in \{1, \ldots, b_n\}$ such that $l_n \to \infty$ and $l_n = o(b_n)$. Define

$$\mathcal{R}(h) = -\sum_{t=1}^{h} \left[\mathcal{E}_{t,h} - E(\mathcal{E}_{t,h})\right]$$
$$\mathcal{U}_{n,s}(h) = \sum_{t=(s-1)b_n+1}^{(s-1)b_n+l_n} \left[\mathcal{E}_{t,h} - E(\mathcal{E}_{t,h})\right]$$
$$\mathcal{Y}_{n,s}(h) = \sum_{t=(s-1)b_n+l_n+1}^{sb_n} \left[\mathcal{E}_{t,h} - E(\mathcal{E}_{t,h})\right]$$

Observe that for $h < l_n$, $\sum_{t=(s-1)b_n+1+h}^{sb_n} \left[\mathcal{E}_{t,h} - E(\mathcal{E}_{t,h})\right] = \mathcal{Y}_{n,s}(h) + \mathcal{U}_{n,s}(h) + \mathcal{R}(h)$ by construction. This implies

$$\begin{aligned}
&\frac{1}{n}\sum_{s=1}^{n/b_n} \left\{ \sum_{t=(s-1)b_n+1+h}^{sb_n} \left(\mathcal{E}_{t,h} - E(\mathcal{E}_{t,h})\right) \right\} \left\{ \sum_{t=(s-1)b_n+1+\tilde{h}}^{sb_n} \left(\mathcal{E}_{t,\tilde{h}} - E(\mathcal{E}_{t,\tilde{h}})\right) \right\} \\
&= \frac{1}{n}\sum_{s=1}^{n/b_n} \left\{ \mathcal{Y}_{n,s}(h) + \mathcal{U}_{n,s}(h) + \mathcal{R}(h) \right\} \left\{ \mathcal{Y}_{n,s}(\tilde{h}) + \mathcal{U}_{n,s}(\tilde{h}) + \mathcal{R}(\tilde{h})) \right\} \\
&= \frac{1}{n}\sum_{s=1}^{n/b_n} \mathcal{Y}_{n,s}(h)\mathcal{Y}_{n,s}(\tilde{h}) + \frac{1}{n}\sum_{s=1}^{n/b_n} \mathcal{U}_{n,s}(h)\mathcal{U}_{n,s}(\tilde{h}) + \frac{1}{n}\sum_{s=1}^{n/b_n} \mathcal{R}_{n,s}(h)\mathcal{R}_{n,s}(\tilde{h}) \\
&+ \frac{1}{n}\sum_{s=1}^{n/b_n} \mathcal{Y}_{n,s}(h)\mathcal{U}_{n,s}(\tilde{h}) + \frac{1}{n}\sum_{s=1}^{n/b_n} \mathcal{Y}_{n,s}(h)\mathcal{R}_{n,s}(\tilde{h}) + \frac{1}{n}\sum_{s=1}^{n/b_n} \mathcal{U}_{n,s}(h)\mathcal{Y}_{n,s}(\tilde{h}) \\
&+ \frac{1}{n}\sum_{s=1}^{n/b_n} \mathcal{U}_{n,s}(h)\mathcal{R}_{n,s}(\tilde{h}) + \frac{1}{n}\sum_{s=1}^{n/b_n} \mathcal{R}_{n,s}(h)\mathcal{Y}_{n,s}(\tilde{h}) + \frac{1}{n}\sum_{s=1}^{n/b_n} \mathcal{R}_{n,s}(h)\mathcal{U}_{n,s}(\tilde{h}).
\end{aligned}$$

We prove

$$\frac{1}{n}\sum_{s=1}^{n/b_n}\left\{\sum_{t=(s-1)b_n+1+h}^{sb_n}\left(\mathcal{E}_{t,h}-E(\mathcal{E}_{t,h})\right)\right\}\left\{\sum_{t=(s-1)b_n+1+\tilde{h}}^{sb_n}\left(\mathcal{E}_{t,\tilde{h}}-E(\mathcal{E}_{t,\tilde{h}})\right)\right\}$$
$$=\frac{1}{n}\sum_{s=1}^{n/b_n}\mathcal{Y}_{n,s}(h)\mathcal{Y}_{n,s}(\tilde{h})+o_p(1). \tag{A.10}$$

Observe that $\frac{1}{n}\sum_{s=1}^{n/b_n}\mathcal{R}_{n,s}(h)\mathcal{R}_{n,s}(\tilde{h})=\frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\frac{1}{b_n}\mathcal{R}_{n,s}(h)\mathcal{R}_{n,s}(\tilde{h})=\frac{1}{b_n}\mathcal{R}_{n,s}(h)\mathcal{R}_{n,s}(\tilde{h})$. Under Assumptions 7, 4, and 8, $\mathcal{E}_{t,h}$ is stationary, ergodic, and L_2-bounded. Therefore

$$E\left|\left|\frac{1}{b_n}\mathcal{R}_{n,s}(h)\mathcal{R}_{n,s}(\tilde{h})\right|\right|\leq K/b_n\to 0.$$

Next, the NED properties and moment bounds of ε_t and m_t^θ in Assumptions 7, 4, and 8 imply that $\mathcal{E}_{t,h}=\varepsilon_t\varepsilon_{t-h}-\mathcal{D}^\theta(h)'J^{-1}m_t^\theta$ is stationary, L_p-bounded for some $p>2$, and L_2-NED on an α-mixing base with decay rate $O(h^{-p/(p-2)})$. Then $||(1/\sqrt{b_n})\mathcal{Y}_{n,1}(h)||_2$ and $||(1/\sqrt{l_n})\mathcal{U}_{n,1}(h)||_2$ are $O(1)$ by McLeish (1975), Theorem 1.6. Observe

$$\begin{aligned}
\left|\left|\frac{1}{n}\sum_{s=1}^{n/b_n}\mathcal{Y}_{n,s}(h)\mathcal{U}_{n,s}(\tilde{h})\right|\right|_1 &= \left|\left|\frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\frac{l_n}{b_n}\mathcal{Y}_{n,s}(h)\frac{1}{l_n}\mathcal{U}_{n,s}(\tilde{h})\right|\right|_1 \\
&= \left|\left|\frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\left(\frac{l_n}{b_n}\right)^{1/2}\frac{1}{\sqrt{b_n}}\mathcal{Y}_{n,s}(h)\frac{1}{\sqrt{l_n}}\mathcal{U}_{n,s}(\tilde{h})\right|\right|_1 \\
&\leq \frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\left|\left|\left(\frac{l_n}{b_n}\right)^{1/2}\frac{1}{\sqrt{b_n}}\mathcal{Y}_{n,s}(h)\frac{1}{\sqrt{l_n}}\mathcal{U}_{n,s}(\tilde{h})\right|\right|_1 \\
&\leq \frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\left(\frac{l_n}{b_n}\right)^{1/2}\left|\left|\frac{1}{\sqrt{b_n}}\mathcal{Y}_{n,s}(h)\right|\right|_2\left|\left|\frac{1}{\sqrt{l_n}}\mathcal{U}_{n,s}(\tilde{h})\right|\right|_2 \\
&= \left(\frac{l_n}{b_n}\right)^{1/2}\left|\left|\frac{1}{\sqrt{b_n}}\mathcal{Y}_{n,1}(h)\right|\right|_2\left|\left|\frac{1}{\sqrt{l_n}}\mathcal{U}_{n,1}(\tilde{h})\right|\right|_2 \\
&= O\left(\left(\frac{l_n}{b_n}\right)^{1/2}\right)=o(1)
\end{aligned}$$

by stationarity, Minkowski's inequality, and the Cauchy-Schwartz inequality. The remaining terms

are shown to be $o(1)$ in a similar fashion. This proves A.10.

Finally, by the NED property, we see by the proof of de Jong's (1997) Theorem 2 that

$$\frac{1}{n}\sum_{s=1}^{n/b_n}\mathcal{Y}_{n,s}(h)\mathcal{Y}_{n,s}(\tilde{h}) \xrightarrow{p} \lim_{n\to\infty}\frac{1}{n}E\left[\left\{\sum_{t=1}^{n}\left(\mathcal{E}_{t,h}-E(\mathcal{E}_{t,h})\right)\right\}\left\{\sum_{t=1}^{n}\left(\mathcal{E}_{t,\tilde{h}}-E(\mathcal{E}_{t,\tilde{h}})\right)\right\}\right].$$

Combine this with A.10 to see that

$$nE(\rho_n^*(h)\rho_n^*(\tilde{h})|\mathcal{W}_n) \xrightarrow{p} \lim_{n\to\infty}E[nE(\rho_n^*(h)\rho_n^*(\tilde{h})|\mathcal{W}_n)] = E[\mathcal{Z}^\theta(h)\mathcal{Z}^\theta(\tilde{h})],$$

so that $P(\mathcal{W}_n \in \mathcal{W}) = 1$ as desired.

Now consider A.9. Recall $\hat{\mathcal{E}}_{t,h}(\hat{\theta}_n) = \varepsilon_t(\hat{\theta}_n)\varepsilon_{t-h}(\hat{\theta}_n) - (B(\hat{\beta}_n)^{-1}\hat{\mathcal{D}}_n^\theta(h,\hat{\theta}_n))'(\hat{J}_n(\hat{\theta}_n))^{-1}m_t^\theta(\hat{\theta}_n)$ and $\mathcal{E}_{t,h} = \varepsilon_t\varepsilon_{t-h} - \mathcal{D}^\theta(h)'J^{-1}m_t^\theta$.

Observe that by the construction of z_t for ξ_s iid $N(0,1)$,

$$\begin{aligned}E\Big[\Big(\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}z_t\big(\mathcal{E}_{t,h}-E(\mathcal{E}_{t,h})\big)\Big)^2\Big] &= E\Big[\Big(\frac{1}{\sqrt{n}}\sum_{s=1}^{n/b_n}\xi_s\sum_{t=(s-1)b_n+1}^{sb_n}\big(\mathcal{E}_{t,h}-E(\mathcal{E}_{t,h})\big)\Big)^2\Big]\\ &= E\Big[\Big(\frac{1}{\sqrt{b_n}}\sum_{t=1}^{b_n}\big(\mathcal{E}_{t,h}-E(\mathcal{E}_{t,h})\big)\Big)^2\Big].\end{aligned}$$

Then Under Assumptions ..., $\mathcal{E}_{t,h} - E(\mathcal{E}_{t,h})$ is zero mean, stationary, L_p-bounded for some $p > 2$, and L_2-NED with size $-1/2$ on an α-mixing base with decay rate $O(h^{-p/(p-2)-\iota})$ by Theorems 17.8 and 17.9 in Davidson (1994). Hence the term above is $O(1)$, so that $\frac{1}{\sqrt{b_n}}\sum_{t=1}^{b_n}\big(\mathcal{E}_{t,h} - E(\mathcal{E}_{t,h})\big) = O_p(1)$. Further, Lemma A.2.3 implies that $\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2(\hat{\theta}_n) \xrightarrow{p} E(\varepsilon_t^2)$.

We will next prove

$$\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}z_t\Big(\hat{\mathcal{E}}_{t,h}(\hat{\theta}_n) - \frac{1}{n}\sum_{s=1+h}^{n}\hat{\mathcal{E}}_{t,h}(\hat{\theta}_n)\Big) = \frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}z_t\big(\mathcal{E}_{t,h}-E(\mathcal{E}_{t,h})\big) + o_p(1) \qquad \text{(A.11)}$$

which coupled with the previous arguments give

$$\sqrt{n}\hat{\rho}_n^{(s)}(h) = \frac{1}{n^{-1}\sum_{t=1}^{n}\varepsilon_t^2(\hat{\theta}_n)} \times \left\{\frac{1}{n}\sum_{t=1+h}^{n}z_t\Big(\hat{\mathcal{E}}_{t,h}(\hat{\theta}_n) - \frac{1}{n}\sum_{t=1+h}^{n}\hat{\mathcal{E}}_{t,h}(\hat{\theta}_n)\Big)\right\}$$

$$= \frac{1}{E(\varepsilon_t^2)} \times \left\{ \frac{1}{n} \sum_{t=1+h}^{n} z_t \Big(\mathcal{E}_{t,h} - E(\mathcal{E}_{t,h}) \Big) \right\} + o_p(1)$$

In order to prove A.11, we must show the following steps:

$$\frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t \varepsilon_t(\hat{\theta}_n) \varepsilon_{t-h}(\hat{\theta}_n) = \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t \varepsilon_t \varepsilon_{t-h} + o_p(1) \tag{A.12}$$

$$(B^{-1}(\hat{\beta}_n) \hat{\mathcal{D}}_n^{\theta}(h, \hat{\theta}_n))' \hat{J}_n^{-1}(\hat{\theta}_n) \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t m_t^{\theta}(\hat{\theta}_n) = \mathcal{D}^{\theta}(h)' J^{-1} \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t m_t^{\theta} + o_p(1) \tag{A.13}$$

$$\frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t \frac{1}{n} \sum_{s=1+h}^{n} \varepsilon_s(\hat{\theta}_n) \varepsilon_{s-h}(\hat{\theta}_n) = \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t E(\varepsilon_t \varepsilon_{t-h}) + o_p(1) \tag{A.14}$$

$$\frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t (B^{-1}(\hat{\beta}_n) \hat{\mathcal{D}}_n^{\theta}(h, \hat{\theta}_n))' \hat{J}_n^{-1}(\hat{\theta}_n) \frac{1}{n} \sum_{s=1+h}^{n} m_s^{\theta}(\hat{\theta}_n) = o_p(1) \tag{A.15}$$

Since z_t is a mean zero Gaussian random variable that is independent of the sample, the proof of Lemma A.2.4 applies to show

$$\frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t \varepsilon_t(\hat{\theta}_n) \varepsilon_{t-h}(\hat{\theta}_n)$$
$$= \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t \varepsilon_t \varepsilon_{t-h} - \sqrt{n}(\hat{\theta}_n - \theta_n)' \frac{1}{n} \sum_{t=1+h}^{n} z_t [d_{\theta,t} \varepsilon_{t-h} + d_{\theta,t-h} \varepsilon_t] + o_p(1)$$

Next, observe

$$\frac{1}{n} \sum_{t=1+h}^{n} z_t [d_{\theta,t} \varepsilon_{t-h}] = \frac{1}{n} \sum_{t=1}^{n} z_t [d_{\theta,t} \varepsilon_{t-h}] + o_p(1)$$
$$= \frac{1}{n} \sum_{s=1}^{n/b_n} \xi_s \sum_{t=(s-1)b_n+1}^{sb_n} [d_{\theta,t} \varepsilon_{t-h}] + o_p(1)$$

by the moment bounds and the construction of z_t. Recall ξ_s is iid, independent of the sample, and

has zero mean and unit variance. Then stationarity, and the moment bounds imply

$$E\Big[\Big(\frac{1}{n}\sum_{s=1}^{n/b_n}\xi_s\sum_{t=(s-1)b_n+1}^{sb_n}[d_{\theta,t}\varepsilon_{t-h}]\Big)^2\Big] = \frac{b_n}{n}E\Big[\Big(\frac{1}{b_n}\sum_{t=1}^{b_n}[d_{\theta,t}\varepsilon_{t-h}]\Big)^2\Big]$$
$$\leq \frac{b_n}{n}E\Big[[d_{\theta,t}\varepsilon_{t-h}]^2\Big] = o(1)$$

because $b_n = o(n)$. Now apply Chebyshev's inequality to see

$$P(\frac{1}{n}\sum_{t=1+h}^{n} z_t[d_{\theta,t}\varepsilon_{t-h}] > \eta) \leq \frac{b_n}{n}E\Big[[d_{\theta,t}\varepsilon_{t-h}]^2\Big]/\eta^2 \to 0,$$

so that $\frac{1}{n}\sum_{t=1+h}^{n} z_t[d_{\theta,t}\varepsilon_{t-h}] \xrightarrow{p} 0$. Then $\sqrt{n}(\hat{\theta}_n - \theta_n) = O_p(1)$ implies that

$$\sqrt{n}(\hat{\theta}_n - \theta_n)'\frac{1}{n}\sum_{t=1+h}^{n} z_t[d_{\theta,t}\varepsilon_{t-h} + d_{\theta,t-h}\varepsilon_t] = o_p(1),$$

so A.12 holds.

Next, consider A.13, and write

$$(B^{-1}(\hat{\beta}_n)\hat{\mathcal{D}}_n^{\theta}(h,\hat{\theta}_n))'\hat{J}_n^{-1}(\hat{\theta}_n)\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t m_t^{\theta}(\hat{\theta}_n)$$
$$= \mathcal{D}^{\theta}(h)'J^{-1}\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t m_t^{\theta}$$
$$+ \mathcal{D}^{\theta}(h)'J^{-1}\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t\big(m_t^{\theta}(\hat{\theta}_n) - m_t^{\theta}\big)$$
$$+ \Big((B^{-1}(\hat{\beta}_n)\hat{\mathcal{D}}_n^{\theta}(h,\hat{\theta}_n))'\hat{J}_n^{-1}(\hat{\theta}_n) - \mathcal{D}^{\theta}(h)'J^{-1}\Big)\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t m_t^{\theta}$$
$$+ \Big((B^{-1}(\hat{\beta}_n)\hat{\mathcal{D}}_n^{\theta}(h,\hat{\theta}_n))'\hat{J}_n^{-1}(\hat{\theta}_n) - \mathcal{D}^{\theta}(h)'J^{-1}\Big)\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t\big(m_t^{\theta}(\hat{\theta}_n) - m_t^{\theta}\big).$$

Recall that $B^{-1}(\hat{\beta}_n)\hat{\mathcal{D}}_n^{\theta}(h,\hat{\theta}_n) \xrightarrow{p} \mathcal{D}^{\theta}(h)$ by Lemma A.2.5, and $\hat{J}_n^{-1}(\hat{\theta}_n) \xrightarrow{p} J^{-1}$ by assumption; hence $(B^{-1}(\hat{\beta}_n)\hat{\mathcal{D}}_n^{\theta}(h,\hat{\theta}_n))'\hat{J}_n^{-1}(\hat{\theta}_n) - \mathcal{D}^{\theta}(h)'J^{-1} = o_p(1)$. Next, observe that $\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t m_t^{\theta} = O_p(1)$ by the moment bounds, the NED property, McLeish (1975), and Chebyshev's inequality.

Finally, observe

$$||\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t\big(m_t^\theta(\hat{\theta}_n) - m_t^\theta\big)|| \leq \sup_{\theta^*\in\Theta} ||\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t \frac{\partial}{\partial\theta} m_t^\theta(\theta^*)|| \times ||\hat{\theta}_n - \theta_n||$$

by the mean value theorem. Since $||\hat{\theta}_n - \theta_n|| = O_p(1/\sqrt{n})$, it remains to show that $\sup_{\theta^*\in\Theta} ||\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t \frac{\partial}{\partial\theta} m_t^\theta(\theta^*)|| = o_p(\sqrt{n})$, which is shown in Lemma A.2.1 in the Appendix.

Next, we prove A.14. First, observe

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} z_t = \frac{\sqrt{b_n}}{\sqrt{n/b_n}}\sum_{t=1}^{n/b_n} \xi_t = O_p(\sqrt{b_n}).$$

Now from the proof of Lemma A.2.3, we have that $\frac{1}{n}\sum_{s=1+h}^{n} \varepsilon_s(\hat{\theta}_n)\varepsilon_{s-h}(\hat{\theta}_n) = \frac{1}{n}\sum_{s=1+h}^{n} \varepsilon_s\varepsilon_{s-h} + O_p(1/\sqrt{n})$. Hence

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{t=1}^{n} z_t \frac{1}{n}\sum_{s=1+h}^{n} \varepsilon_s(\hat{\theta}_n)\varepsilon_{s-h}(\hat{\theta}_n) &= \frac{\sqrt{b_n}}{\sqrt{n/b_n}}\sum_{t=1}^{n/b_n} \xi_t\Big(\frac{1}{n}\sum_{s=1+h}^{n} \varepsilon_s\varepsilon_{s-h} + O_p(1/\sqrt{n})\Big) \\
&= \frac{\sqrt{b_n}}{\sqrt{n/b_n}}\sum_{t=1}^{n/b_n} \xi_t \frac{1}{n}\sum_{s=1+h}^{n} \varepsilon_s\varepsilon_{s-h} + O_p(\sqrt{b_n}/\sqrt{n}) \\
&= \frac{1}{\sqrt{n}}\sum_{t=1}^{n} z_t \frac{1}{n}\sum_{s=1+h}^{n} \varepsilon_s\varepsilon_{s-h} + O_p(1/\sqrt{n/b_n})
\end{aligned}$$

Now, $\varepsilon_t\varepsilon_{t-h} - E(\varepsilon_t\varepsilon_{t-h})$ is zero mean, stationary, L_p-bounded for some $p > 2$, and L_2-NED with size $-1/2$ on an α-mixing base with decay rate $O(h^{-p/(p-2)-\iota})$ by Theorems 17.8 and 17.9 in Davidson (1994). Then by Theorem 1.6 in McLeish (1975), $E((\frac{1}{\sqrt{n}}\sum_{s=1+h}^{n}(\varepsilon_s\varepsilon_{s-h} - E(\varepsilon_s\varepsilon_{s-h})))^2) = O(1)$. Thus $\frac{1}{\sqrt{n}}\sum_{s=1+h}^{n}(\varepsilon_s\varepsilon_{s-h} - E(\varepsilon_s\varepsilon_{s-h})) = O_p(1)$, so

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{t=1}^{n} z_t \frac{1}{n}\sum_{s=1+h}^{n} (\varepsilon_s\varepsilon_{s-h} - E(\varepsilon_s\varepsilon_{s-h})) \\
&\qquad = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} z_t \times O_p(1/\sqrt{n}) = O_p(1/\sqrt{n/b_n}) = o_p(1).
\end{aligned}$$

Finally, we prove A.15. Recall $\frac{1}{\sqrt{n}}\sum_{t=1}^{n} z_t = O_p(\sqrt{b_n})$ and $(B^{-1}(\hat{\beta}_n)\hat{\mathcal{D}}_n^\theta(h,\hat{\theta}_n))'\hat{J}_n^{-1}(\hat{\theta}_n) \xrightarrow{p}$

$\mathcal{D}^\theta(h)'J^{-1}$. It is therefore sufficient to show $\frac{1}{n}\sum_{t=1}^n m_t^\theta(\hat{\theta}_n) = o_p(1/\sqrt{b_n})$. Observe that

$$||\frac{1}{n}\sum_{t=1+h}^n m_t^\theta(\hat{\theta}_n) - \frac{1}{n}\sum_{t=1+h}^n m_t^\theta|| \leq \sup_{\theta^*\in\Theta}||\frac{1}{n}\sum_{t=1+h}^n \frac{\partial}{\partial\theta}m_t^\theta(\theta^*)|| \times ||\hat{\theta}_n - \theta_n||$$

by the mean value theorem. Now, $||\hat{\theta}_n - \theta_n|| = O_p(1/\sqrt{n})$, and $\{m_t\}$ is zero mean, stationary, L_p-bounded for some $p > 2$, and L_2-NED with size $-1/2$ on an α-mixing base with decay rate $O(h^{-p/(p-2)-\iota})$, so $E[(\frac{1}{\sqrt{n}}\sum_{t=1}^n m_t^\theta)^2] = O(1)$ by McLeish (1975), Theorem 1.6. Hence, $\frac{1}{\sqrt{n}}\sum_{t=1}^n m_t^\theta = O_p(1)$, and $\frac{1}{n}\sum_{t=1+h}^n m_t^\theta = O_p(1/\sqrt{n})$. Since, $b_n = o(n)$, we have that $O(\frac{1}{\sqrt{n}}) = O(\frac{\sqrt{b_n}}{\sqrt{n}}\frac{1}{\sqrt{b_n}}) = o(\frac{1}{\sqrt{b_n}})$. Thus $\frac{1}{n}\sum_{t=1+h}^n m_t^\theta = o_p(1/\sqrt{b_n})$. We need only show that $\sup_{\theta^*\in\Theta}||\frac{1}{n}\sum_{t=1+h}^n \frac{\partial}{\partial\theta}m_t^\theta(\theta^*)|| = O_p(1)$, which follows from Lemma A.2.1.

This completes the proof of A.6, so that

$$\{\sqrt{n}\hat{\rho}_n^{(s)}(h) : 1 \leq h \leq \mathcal{L}\} \Rightarrow^p \{\overset{\circ}{\mathcal{Z}}(h) : 1 \leq h \leq \mathcal{L}\}$$

for each $\mathcal{L} \in \mathbb{N}$, where $\{\overset{\circ}{\mathcal{Z}}(h) : h \in \mathbb{N}\}$ is an independent copy of $\{\mathcal{Z}^\theta(h) : h \in \mathbb{N}\}$, the zero mean Gaussian process in Lemma 3.2.

Now we prove A.7: For the process $\{\overset{\circ}{\mathcal{Z}}(h) : 1 \leq h \leq \mathcal{L}\}$ and some sequence of positive integers $\{\mathcal{L}_n\}$, $\mathcal{L}_n \to \infty$ and $\mathcal{L}_n = o(n)$,

$$\mathcal{A}_{\mathcal{L},n} \equiv \sup_{c>0}\left|P(\max_{1\leq h\leq\mathcal{L}_n}|\sqrt{n}\hat{\rho}_n^{(s)}(h)| \leq c|\mathcal{W}_n) - P(\max_{1\leq h\leq\mathcal{L}_n}|\overset{\circ}{\mathcal{Z}}(h)| \leq c)\right| \xrightarrow{p} 0.$$

Pair A.6 with the continuous mapping theorem to yield

$$\{\max_{1\leq h\leq\mathcal{L}_n}|\sqrt{n}\hat{\rho}_n^{(s)}(h)| : 1 \leq h \leq \mathcal{L}\} \Rightarrow^p \{\max_{1\leq h\leq\mathcal{L}_n}|\overset{\circ}{\mathcal{Z}}(h)| : 1 \leq h \leq \mathcal{L}\}$$

for each $\mathcal{L} \in \mathbb{N}$. This implies

$$\sup_{c>0}\left|P(\max_{1\leq h\leq\mathcal{L}}|\sqrt{n}\hat{\rho}_n^{(s)}(h)| \leq c|\mathcal{W}_n) - P(\max_{1\leq h\leq\mathcal{L}}|\overset{\circ}{\mathcal{Z}}(h)| \leq c)\right| \xrightarrow{p} 0.$$

(see Giné and Zinn (1990), page 862.) In order to prove A.7, $\mathcal{A}_{\mathcal{L}_n,n} \xrightarrow{p} 0$ for some $\{\mathcal{L}_n : n \geq 1\}$, $\mathcal{L}_n \to \infty$ with $\mathcal{L}_n = o(n)$, we follow the proof of Lemma A.2(a) in Hill and Motegi (2018).

The dominated convergence theorem implies

$$\lim_{n\to\infty} \int_0^1 P(\mathcal{A}_{\mathcal{L},n} > \eta)d\eta = \int_0^1 \lim_{n\to\infty} P(\mathcal{A}_{\mathcal{L},n} > \eta)d\eta = 0,$$

so by Lemma A.1 in Hill and Motegi (2018), we have $\int_0^1 P(\mathcal{A}_{\mathcal{L}_n,n} > \eta)d\eta = E(\mathcal{A}_{\mathcal{L}_n,n}) \to 0$ for some non-unique sequence of integers $\{\mathcal{L}_n : n \geq 1\}$, $\mathcal{L}_n \to \infty$ with $\mathcal{L}_n = o(n)$. Then by Markov's inequality, $P(\mathcal{A}_{\mathcal{L}_n,n} > \eta) \leq E(\mathcal{A}_{\mathcal{L}_n,n})/\eta \to 0$ for all $\eta > 0$. Hence $\mathcal{A}_{\mathcal{L}_n,n} \xrightarrow{p} 0$ for some non-unique sequence of integers $\{\mathcal{L}_n : n \geq 1\}$, $\mathcal{L}_n \to \infty$ with $\mathcal{L}_n = o(n)$.

Step 3. Finally, we show the consistency of the critical values. Define the quantile functions $\hat{F}_n^{-1}(u|\cdot) = \inf\{c \geq 0 : \ P(\hat{\mathcal{T}}_n^{(s)} \leq c|\cdot) \geq u\}$, $F_n^{-1}(u) = \inf\{c \geq 0 : \ P(\hat{\mathcal{T}}_n \leq c) \geq u\}$.

Operate conditionally on the sample $\mathcal{W}_n$. From A.7, $\{\hat{\mathcal{T}}_{n,j}^{(s)}\}_{j=1}^M$ is a sequence of iid draws from $\max_{1\leq h\leq \mathcal{L}_n} |\ \mathring{\mathcal{Z}}\ (h)|$ asymptotically with probability approaching one with respect to the sample $\mathcal{W}_n$. Thus under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $\hat{\mathcal{T}}_n$ and $\hat{\mathcal{T}}_n^{(s)}$ have the same limits under H_0. Hence, under H_0,

$$\sup_{c>0} \left| P(\hat{\mathcal{T}}_n^{(s)} \leq c|\mathcal{W}_n) - P(\hat{\mathcal{T}}_n \leq c) \right| \xrightarrow{p} 0.$$

Therefore $\sup_{u\in[0,1]} \left| \hat{F}_n^{-1}(u|\mathcal{W}_n) - F_n^{-1}(u) \right| \xrightarrow{p} 0$. Further, by independence and letting $M_n \to \infty$, the bootstrapped critical value $\hat{c}_{n,1-\alpha,M_n}^{(s)} = \hat{\mathcal{T}}_{n,[(1-\alpha)\cdot M_n]}^{(s)}$ is a central order statistic (see e.g. Galambos (1987)) of a conditionally iid random variable, so $\left| \hat{c}_{n,1-\alpha,M_n}^{(s)} - \hat{F}_n^{-1}(1-\alpha|\mathcal{W}_n) \right| \xrightarrow{p} 0$. Combining these statements yields $\left| \hat{c}_{n,1-\alpha,M_n}^{(s)} - F_n^{-1}(1-\alpha) \right| \xrightarrow{p} 0$. Since $c_{n,1-\alpha} = F_n^{-1}(1-\alpha)$, the proof is complete.

(a) Weak Identification. We next prove the claim under weak identification. The proof under this case will proceed similarly to the proof under strong identification; however, the following will require more steps due to the inconsistency of $\hat{\pi}_n$ for π_0 and the required bootstrap step for calculating the bootstrapped π^*, and the joint convergence of $\hat{\pi}_n$ with the other variables.

Let $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$. We will prove the following two steps:

$$\{\sqrt{n}\hat{\rho}_n^{(w)}(h) : \ 1 \leq h \leq \mathcal{L}\} \Rightarrow^p \{\overset{\circ}{\mathcal{Z}}\ (h, \pi^*(b, \gamma_0)) : \ 1 \leq h \leq \mathcal{L}\} \tag{A.16}$$

for each $\mathcal{L} \in \mathbb{N}$, where $\{\overset{\circ}{\mathcal{Z}}\ (h, \pi) : h \in \mathbb{N}, \ \pi \in \Pi\}$ is an independent copy of $\{\mathcal{Z}^{\psi}(h, \pi) : h \in \mathbb{N}, \ \pi \in \Pi\}$, the Gaussian process in Lemma 3.2(a). Second, for the process $\{\overset{\circ}{\mathcal{Z}}\ (h, \pi) : \ 1 \leq h \leq \mathcal{L}, \ \pi \in \Pi\}$ and some sequence of positive integers $\{\mathcal{L}_n\}$, $\mathcal{L}_n \to \infty$ and $\mathcal{L}_n = o(n)$,

$$\sup_{c>0} \left| P(\max_{1 \leq h \leq \mathcal{L}_n} |\sqrt{n}\hat{\rho}_n^{(w)}(h)| \leq c | \mathcal{W}_n) - P(\max_{1 \leq h \leq \mathcal{L}_n} | \overset{\circ}{\mathcal{Z}}\ (h, \pi^*(b, \gamma_0))| \leq c) \right| \xrightarrow{p} 0. \tag{A.17}$$

We prove A.16 in the following several steps: First, we prove

$$\{\hat{G}_n^{(bs)}(\pi) : \ \pi \in \Pi\} \Rightarrow^p \{G(\pi; \gamma_0) : \ \pi \in \Pi\} \tag{A.18}$$

where $G(\pi; \gamma_0)$ is the mean zero Gaussian process, with covariance kernel $\Omega(\pi, \tilde{\pi}; \gamma_0)$, that is the weak limit of $G_n(\cdot)$ under weak identification. Together with uniform convergence in probability of $H_n(\hat{\psi}_{0,n}, \pi)$ to $H(\pi; \gamma_0)$ and $K_n(\tilde{\psi}_n, \pi; \tilde{\gamma}_n)$ to $K(\psi_0, \pi; \gamma_0)$, this step will imply $\{\xi_n^{(bs)}(\pi; \gamma_0, b) : \pi \in \Pi\} \Rightarrow^p \{\xi(\pi; b, \gamma_0) : \pi \in \Pi\}$. Then the argmax continuity theorem (cf van der Vaart and Wellner (1996), Lemma 3.2.1 and Andrews and Cheng (2012b), Theorem 9.10.) will yield

$$\pi_{(bs)}^*(\gamma_0, b) \xrightarrow{d} \pi^*(\gamma_0, b). \tag{A.19}$$

Next, we will prove the weak convergence result

$$\{\sqrt{n}\hat{\rho}_n^{(w)}(h, \pi) : \ 1 \leq h \leq \mathcal{L}, \ \pi \in \Pi\} \Rightarrow^p \{\overset{\circ}{\mathcal{Z}}\ (h, \pi) : \ 1 \leq h \leq \mathcal{L}, \ \pi \in \Pi\} \tag{A.20}$$

where $\hat{\rho}_n^{(w)}(h, \pi)$ will be defined precisely. The proof of A.20 will follow similarly to the proof of A.6 under strong identification. Finally, we will prove joint weak convergence

$$\{\sqrt{n}\hat{\rho}_n^{(w)}(h, \pi), \pi_{(bs)}^*(b, \gamma_0) : 1 \leq h \leq \mathcal{L}, \ \pi \in \Pi\}$$

$$\Rightarrow^p \{\overset{\circ}{\mathcal{Z}}(h,\pi), \pi^*(b,\gamma_0) : \ 1 \leq h \leq \mathcal{L}, \ \pi \in \Pi\} \tag{A.21}$$

which will follow simply from the construction of $\hat{\rho}_n^{(w)}(h,\pi)$, the fact that $\pi^*_{(bs)}(b,\gamma_0)$ is a continuous function of $\hat{G}_n^{(bs)}(\pi)$, $K_n(\hat{\psi}_{0,n},\pi;\gamma_0)$, and $H_n(\psi_{0,n},\pi)$, and the continuous mapping theorem. The result A.16 will follow.

We now begin the proof of A.16. Let $\{\gamma_n\} \in \Gamma(\gamma_0,0,b)$ with $||b|| < \infty$, and operate conditionally on the sample $\mathcal{W}_n \equiv \{m_t, x_t, y_t\}_{t=1}^n$. First, we prove A.18:

$$\{\hat{G}_n^{(bs)}(\pi) : \ \pi \in \Pi\} \Rightarrow^p \{G(\pi;\gamma_0) : \ \pi \in \Pi\}$$

where $G(\pi;\gamma_0)$ is the mean zero Gaussian process, with covariance kernel $\Omega(\pi,\tilde{\pi};\gamma_0) = E[G(\pi;\gamma_0)\ G(\tilde{\pi};\gamma_0)']$, that is the weak limit of $G_n(\cdot)$ under weak identification. We must prove convergence in finite dimensional distributions and establish stochastic equicontinuity (see e.g. Giné and Zinn (1990), Andrews (1994), or Pollard (1990)).

Recall $\hat{G}_n^{(bs)}(\pi) = \frac{1}{\sqrt{n}}\sum_{t=1}^n z_t\big(m_t^\psi(\psi_{0,n},\pi) - \frac{1}{n}\sum_{t=1}^n m_t^\psi(\psi_{0,n},\pi)\big)$. We prove convergence in finite dimensional distributions with an argument in Hansen (1996). By construction of z_t, we can write

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{t=1}^n z_t\big(m_t^\psi(\psi_{0,n},\pi) - \frac{1}{n}\sum_{t=1}^n m_t^\psi(\psi_{0,n},\pi)\big) \\
&\qquad = \frac{b_n}{\sqrt{n}}\sum_{s=1}^{n/b_n}\xi_s\Big(\frac{1}{b_n}\sum_{t=(s-1)b_n+1}^{sb_n}\big(m_t^\psi(\psi_{0,n},\pi) - \frac{1}{n}\sum_{t=1}^n m_t^\psi(\psi_{0,n},\pi)\big)\Big) \\
&\qquad = \sqrt{n}\frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\xi_s\Big(\frac{1}{b_n}\sum_{t=(s-1)b_n+1}^{sb_n}\big(m_t^\psi(\psi_{0,n},\pi) - \frac{1}{n}\sum_{t=1}^n m_t^\psi(\psi_{0,n},\pi)\big)\Big)
\end{aligned}$$

Then since ξ_t is distributed $N(0,1)$ and independent of the sample, $\hat{G}_n^{(bs)}(\pi)$ is normally distributed with mean zero and covariance kernel

$$
\begin{aligned}
&E\Big(\hat{G}_n^{(bs)}(\pi)\hat{G}_n^{(bs)}(\tilde{\pi})'|\mathcal{W}_n\Big) \\
&= E\Bigg\{\Bigg[\frac{1}{\sqrt{n/b_n}}\sum_{s=1}^{n/b_n}\xi_s\Big(\frac{1}{\sqrt{b_n}}\sum_{t=(s-1)b_n+1}^{sb_n}\big(m_t^{\psi}(\psi_{0,n},\pi)-\frac{1}{n}\sum_{t=1}^{n}m_t^{\psi}(\psi_{0,n},\pi)\big)\Big)\Bigg] \\
&\quad\times\Bigg[\frac{1}{\sqrt{n/b_n}}\sum_{s=1}^{n/b_n}\xi_s\Big(\frac{1}{\sqrt{b_n}}\sum_{t=(s-1)b_n+1}^{sb_n}\big(m_t^{\psi}(\psi_{0,n},\tilde{\pi})-\frac{1}{n}\sum_{t=1}^{n}m_t^{\psi}(\psi_{0,n},\tilde{\pi})\big)\Big)\Bigg]'\Bigg|\mathcal{W}_n\Bigg\} \\
&= \Bigg[\frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\Big(\frac{1}{\sqrt{b_n}}\sum_{t=(s-1)b_n+1}^{sb_n}\big(m_t^{\psi}(\psi_{0,n},\pi)-\frac{1}{n}\sum_{t=1}^{n}m_t^{\psi}(\psi_{0,n},\pi)\big)\Big) \\
&\quad\times\Big(\frac{1}{\sqrt{b_n}}\sum_{u=(s-1)b_n+1}^{sb_n}\big(m_u^{\psi}(\psi_{0,n},\tilde{\pi})-\frac{1}{n}\sum_{t=1}^{n}m_u^{\psi}(\psi_{0,n},\tilde{\pi})\big)\Big)'\Bigg] \\
&= \hat{\Omega}_n(\pi,\tilde{\pi})
\end{aligned}
$$

where $\hat{\Omega}_n(\pi,\tilde{\pi})$ is defined implicitly. Let $\mathcal{W}$ be the set of samples such that

$$
\sup_{\pi,\tilde{\pi}\in\Pi\times\Pi}||E\Big(\hat{G}_n^{(bs)}(\pi)\hat{G}_n^{(bs)}(\tilde{\pi})'|\mathcal{W}_n\Big)-\Omega(\pi,\tilde{\pi};\gamma_0)||\xrightarrow{p}0.
$$

We must show that $\sup_{\pi,\tilde{\pi}\in\Pi\times\Pi}||\hat{\Omega}_n(\pi,\tilde{\pi})-\Omega(\pi,\tilde{\pi};\gamma_0)||\xrightarrow{p}0$ in order to prove that $P(\mathcal{W}_n\in\mathcal{W})=1$. This follows from stationarity, ergodicity and the moment bounds in Assumption 4. Thus $\hat{G}_n^{(bs)}(\pi)$ converges in finite dimensional distributions to a zero mean Gaussian process with covariance kernel $\Omega(\pi,\tilde{\pi};\gamma_0)$.

Observe that under $\{\gamma_n\}\in\Gamma(\gamma_0,0,b)$ with $||b||<\infty$ and H_0, $G_n(\pi)$ has the same limit by Assumption 1. Since Gaussian processes are characterized by their first two moments, the finite dimensional distributions of $\hat{G}_n^{(bs)}(\pi)$ and $G_n(\pi)$ converge to the same limit.

Next, we show stochastic equicontinuity. Let $r\in k_\psi$ be such that $r'r=1$. The mean value

theorem yields

$$r'\Big(m_t^{\psi}(\psi_{0,n},\pi)-m_t^{\psi}(\psi_{0,n},\tilde{\pi})\Big)\leq \sup_{\mathring{\pi}\in\Pi}||r'\frac{\partial}{\partial\pi}m_t^{\psi}(\psi_{0,n},\mathring{\pi})||\times||\tilde{\pi}-\pi||.$$

Next, use the construction of z_t and the fact that z_t is independent of the data and Chebychev's inequality, and observe the following:

$$\begin{aligned}
\mathcal{P}_n(\eta) &= P\Big(\sup_{\pi,\tilde{\pi}\in\Pi:||\tilde{\pi}-\pi||\leq\delta}\Big|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}z_t r'\Big(m_t^{\psi}(\psi_{0,n},\pi)-m_t^{\psi}(\psi_{0,n},\tilde{\pi})\Big)\Big|>\eta|\mathcal{W}_n\Big)\\
&\leq \frac{1}{\eta^2}E\Big[\sup_{\pi,\tilde{\pi}\in\Pi:||\tilde{\pi}-\pi||\leq\delta}\Big(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}z_t r'\Big(m_t^{\psi}(\psi_{0,n},\pi)-m_t^{\psi}(\psi_{0,n},\tilde{\pi})\Big)\Big)^2|\mathcal{W}_n\Big]\\
&= \frac{1}{\eta^2}E\Big[\sup_{\pi,\tilde{\pi}\in\Pi:||\tilde{\pi}-\pi||\leq\delta}\Big(\frac{1}{\sqrt{n/b_n}}\sum_{s=1}^{n/b_n}\xi_s\frac{1}{\sqrt{b_n}}\sum_{t=(s-1)b_n+1}^{sb_n}r'\Big(m_t^{\psi}(\psi_{0,n},\pi)\\
&\qquad -m_t^{\psi}(\psi_{0,n},\tilde{\pi})\Big)\Big)^2|\mathcal{W}_n\Big]\\
&= \frac{1}{\eta^2}\frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\sup_{\substack{\pi,\tilde{\pi}\in\Pi:\\||\tilde{\pi}-\pi||\leq\delta}}\Big(\frac{1}{\sqrt{b_n}}\sum_{t=(s-1)b_n+1}^{sb_n}r'\Big(m_t^{\psi}(\psi_{0,n},\pi)-m_t^{\psi}(\psi_{0,n},\tilde{\pi})\Big)\Big)\\
&\qquad\times\Big(\frac{1}{\sqrt{b_n}}\sum_{u=(s-1)b_n+1}^{sb_n}r'\Big(m_t^{\psi}(\psi_{0,n},\pi)-m_t^{\psi}(\psi_{0,n},\tilde{\pi})\Big)\Big)'\\
&\leq \frac{\delta^2}{\eta^2}\frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\sup_{\mathring{\pi}\in\Pi}\Big(\frac{1}{\sqrt{b_n}}\sum_{t=(s-1)b_n+1}^{sb_n}||r'\frac{\partial}{\partial\pi}m_t^{\psi}(\psi_{0,n},\mathring{\pi})||\Big)\\
&\qquad\times\Big(\frac{1}{\sqrt{b_n}}\sum_{u=(s-1)b_n+1}^{sb_n}||r'\frac{\partial}{\partial\pi}m_t^{\psi}(\psi_{0,n},\mathring{\pi})||\Big)'\\
&= \frac{\delta^2}{\eta^2}\frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\sup_{\mathring{\pi}\in\Pi}\Big(\frac{1}{\sqrt{b_n}}\sum_{t=(s-1)b_n+1}^{sb_n}||r'\frac{\partial}{\partial\pi}m_t^{\psi}(\psi_{0,n},\mathring{\pi})||\Big)^2\\
&= \frac{\delta^2}{\eta^2}C_n
\end{aligned}$$

Now observe that

$$E\left[\frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\sup_{\mathring{\pi}\in\Pi}\Big(\frac{1}{\sqrt{b_n}}\sum_{t=(s-1)b_n+1}^{sb_n}||r'\frac{\partial}{\partial\pi}m_t^{\psi}(\psi_{0,n},\mathring{\pi})||\Big)^2\right]$$

$$= E\left[\sup_{\mathring{\pi}\in\Pi}\left(\frac{1}{\sqrt{b_n}}\sum_{t=1}^{b_n}||r'\frac{\partial}{\partial\pi}m_t^{\psi}(\psi_{0,n},\mathring{\pi})||\right)^2\right]$$
$$= O(1)$$

by Assumption 4. Hence stationarity and ergodicity imply that $C_n \xrightarrow{p} C$ for a finite non-negative constant C. Take $\delta > 0$ such that $0 < \delta \leq (\varepsilon\eta^2/C)^{1/2}$ to see that for every $(\varepsilon, \eta) > 0$, there is a $\delta > 0$ such that $\lim_{n\to\infty}\mathcal{P}_n(\eta) < \varepsilon$ with probability approaching one with respect to the sample $\mathcal{W}_n$.

Next, we prove A.19. Recall $\sup_{\pi\in\Pi}||H_n(\hat{\psi}_{0,n},\pi) - H(\pi;\gamma_0)|| \xrightarrow{p} 0$ and $\sup_{\pi\in\Pi}||K_n(\tilde{\psi}_n,\pi;\tilde{\gamma}_n) - K(\psi_0,\pi;\gamma_0)|| \xrightarrow{p} 0$ for every pair of sequences $\tilde{\psi}_n \to \psi_0$ and $\tilde{\gamma}_n \to \gamma_0$. This paired with A.18 implies $\{\xi_n^{(bs)}(\pi;\gamma_0,b) : \pi\in\Pi\} \Rightarrow^p \{\xi(\pi;b,\gamma_0) : \pi\in\Pi\}$. The argmax continuity theorem (cf van der Vaart and Wellner (1996), Lemma 3.2.1 and Andrews and Cheng (2012b), Theorem 9.10.) then yields $\pi^*_{(bs)}(\gamma_0,b) \xrightarrow{d} \pi^*(\gamma_0,b)$.

Next in order to prove A.20, we show the following two intermediate steps:

$$\{\sqrt{n}\rho_n^*(h,\pi) : 1\leq h\leq\mathcal{L},\ \pi\in\Pi\} \Rightarrow^p \{\mathring{\mathcal{Z}}(h,\pi) :\ 1\leq h\leq\mathcal{L},\ \pi\in\Pi\} \tag{A.22}$$

$$\sqrt{n}\sup_{\pi\in\Pi}|\hat{\rho}_n^{(w)}(h,\pi) - \rho_n^*(h,\pi)| \xrightarrow{p} 0 \text{ for each } h \tag{A.23}$$

where $\{\mathring{\mathcal{Z}}(h,\pi) : h\in\mathbb{N},\ \pi\in\Pi\}$ is an independent copy of $\{\mathcal{Z}^{\psi}(h,\pi) : h\in\mathbb{N},\ \pi\in\Pi\}$.

Recall

$$\begin{aligned}\hat{\mathcal{E}}_{t,h}(\psi,\pi) &= \varepsilon_t(\psi,\pi)\varepsilon_{t-h}(\psi,\pi)\\ &\quad - \hat{\mathcal{D}}_n(h,\pi)'H_n^{-1}(\psi,\pi)\Big(m_t^{\psi}(\psi,\pi) - \frac{1}{n}\sum_{t=1}^{n}m_t^{\psi}(\psi,\pi)\Big)\\ &\quad - \frac{1}{n}\sum_{t=1+h}^{n}[\varepsilon_t(\psi,\pi)\varepsilon_{t-h}(\psi,\pi) - \varepsilon_t\varepsilon_{t-h}]\end{aligned}$$

$$\begin{aligned}\hat{\rho}_n^{(w)}(h,\pi;\gamma_n,b) &= \frac{1}{n^{-1}\sum_{t=1}^{n}\varepsilon_t^2(\hat{\theta}_n)}\times\left\{\frac{1}{n}\sum_{t=1+h}^{n}z_t\Big(\hat{\mathcal{E}}_{t,h}(\hat{\psi}_{0,n},\pi) - \frac{1}{n}\sum_{t=1+h}^{n}\hat{\mathcal{E}}_{t,h}(\hat{\psi}_{0,n},\pi)\Big)\right.\\ &\quad \left. + \frac{1}{n}\sum_{t=1+h}^{n}[\varepsilon_t(\hat{\psi}_{0,n},\pi)\varepsilon_{t-h}(\hat{\psi}_{0,n},\pi) - \varepsilon_t\varepsilon_{t-h}]\right.\end{aligned}$$

$$- \hat{\mathcal{D}}_n(h,\pi)' H_n^{-1}(\hat{\psi}_{0,n},\pi) \times \frac{1}{n}\sum_{t=1}^{n}(b/\sqrt{n})\frac{\partial}{\partial \tilde{\beta}_n} E_{\tilde{\gamma}_n}(m_t^{\psi}(\tilde{\psi_{0,n}},\pi))\Bigg\}.$$

Define

$$\mathcal{E}_{t,h}(\pi) = \varepsilon_t\varepsilon_{t-h} - \mathcal{D}^{\psi}(h,\pi)' H^{-1}(\pi;\gamma_0)\Big(m_t^{\psi}(\psi_{0,n},\pi) - E_{\gamma_n}[m_t^{\psi}(\psi_{0,n},\pi)]\Big)$$

$$\rho_n^*(h,\pi) = \frac{1}{E(\varepsilon_t^2)} \times \Bigg\{\frac{1}{n}\sum_{t=1+h}^{n} z_t\Big(\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi))\Big)$$
$$+ E_{\gamma_n}[\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi) - \varepsilon_t\varepsilon_{t-h}]$$
$$- \mathcal{D}^{\psi}(h,\pi)' H^{-1}(\pi;\gamma_0)(b/\sqrt{n})K(\psi_0,\pi;\gamma_0)\Bigg\}.$$

Further, separate $\rho_n^*(h,\pi)$ into a mean a conditional Gaussian components $\rho_n^*(h,\pi) = \rho_n^{1,*}(h,\pi) + \rho_n^{2,*}(h,\pi)$ where $\rho_n^{1,*}(h,\pi) = \frac{1}{E(\varepsilon_t^2)} \times \Big\{\frac{1}{n}\sum_{t=1+h}^{n} z_t\Big(\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi))\Big)\Big\}$ and $\rho_n^{2,*}(h,\pi) = \frac{1}{E(\varepsilon_t^2)} \times \Big\{E_{\gamma_n}[\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi) - \varepsilon_t\varepsilon_{t-h}] - \mathcal{D}^{\psi}(h,\pi)' H^{-1}(\pi;\gamma_0)(b/\sqrt{n})K(\psi_0,\pi;\gamma_0)\Big\}$.

Here, we again shorten the proof by letting ξ_t be iid $N(0,1)$ random variables, which eliminates the extra steps needed to show asymptotic convergence in conditional distribution. In order to prove A.22, we prove weak convergence in the sense of Hoffmann-Jorgensen (1984, 1991). This requires a totally bounded pseudo metric space, finite dimensional convergence, and stochastic equicontinuity. The proof of this step closely follows the proof of Lemma A.3, step 1 in Hill and Motegi (2018); however, it must be augmented to account for the convergence over Π.

Observe that $\{1,\ldots,\mathcal{L}\} \times \Pi$ is compact, so this space with the sup-norm is totally bounded. In order to prove stochastic equicontinuity, first note that $\{1,\ldots,\mathcal{L}\}$ is discrete and bounded. Next, recall the construction of $\rho_n^*(h,\pi)$, that $m_t^{\psi}(\psi_{0,n},\pi)$ is stochastically equicontinuous, and invoke probability sub-additivity. Finally, we establish convergence of the finite dimensional distributions with the following argument.

We operate conditionally on the sample $\mathcal{W}_n$. Write

$$\rho_n^{1,*}(h,\pi) = \frac{1}{E(\varepsilon_t^2)} \times \frac{1}{n/b_n} \sum_{s=1}^{n/b_n} \xi_s \left\{ \frac{1}{b_n} \sum_{t=(s-1)b_n+1+h}^{sb_n} \Big(\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi)) \Big) \right\}$$

By joint Gaussianity and independence of ξ_s, $\{\sqrt{n}\rho_n^{1,*}(h) : 1 \leq h \leq \mathcal{L}\}$ is a zero mean Gaussian process with covariance function

$$\begin{aligned}
&nE(\rho_n^{1,*}(h,\pi)\rho_n^{1,*}(\tilde{h},\tilde{\pi})'|\mathcal{W}_n) \\
&= \frac{1}{\big(E(\varepsilon_t^2)\big)^2} \times \frac{1}{n} \sum_{s=1}^{n/b_n} \left\{ \sum_{t=(s-1)b_n+1+h}^{sb_n} \Big(\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi)) \Big) \right\} \\
&\qquad \times \left\{ \sum_{t=(s-1)b_n+1+\tilde{h}}^{sb_n} \Big(\mathcal{E}_{t,\tilde{h}}(\tilde{\pi}) - E(\mathcal{E}_{t,\tilde{h}}(\tilde{\pi})) \Big) \right\}'
\end{aligned}$$

for each $\mathcal{L} \in \mathbb{N}$ and $\pi, \tilde{\pi} \in \Pi$. Observe

$$\begin{aligned}
&\lim_{n\to\infty} E\big[nE(\rho_n^{1,*}(h,\pi)\rho_n^{1,*}(\tilde{h},\tilde{\pi})'|\mathcal{W}_n)\big] \\
&= \frac{1}{\big(E(\varepsilon_t^2)\big)^2} \times \lim_{n\to\infty} \frac{1}{n} \sum_{s=1}^{n/b_n} \sum_{t=(s-1)b_n+1+h}^{sb_n} \sum_{u=(s-1)b_n+1+h}^{sb_n} E\Big[\Big(\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi)) \Big) \\
&\qquad \times \Big(\mathcal{E}_{u,\tilde{h}}(\tilde{\pi}) - E(\mathcal{E}_{u,\tilde{h}}(\tilde{\pi})) \Big)' \Big] \\
&= \lim_{n\to\infty} \frac{1}{n} E\left[\sum_{t=1}^{n} \Big(\frac{\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi))}{E(\varepsilon_t^2)} \Big) \sum_{t=1}^{n} \Big(\frac{\mathcal{E}_{t,\tilde{h}}(\tilde{\pi}) - E(\mathcal{E}_{t,\tilde{h}}(\tilde{\pi}))}{E(\varepsilon_t^2)} \Big)' \right] \\
&= E[\mathcal{Z}^{1,\psi}(h,\pi)\mathcal{Z}^{1,\psi}(\tilde{h},\tilde{\pi})]
\end{aligned}$$

where the final equality follows from the definition of $\mathcal{Z}^{1,\psi}(h,\pi)$ in Lemma A.1.2.

Let $\mathcal{W}$ be the set of samples such that

$$nE(\rho_n^{1,*}(h,\pi)\rho_n^{1,*}(\tilde{h},\tilde{\pi})|\mathcal{W}_n) \xrightarrow{p} E[\mathcal{Z}^{1,\psi}(h,\pi)\mathcal{Z}^{1,\psi}(\tilde{h},\tilde{\pi})].$$

We will show that $P(\mathcal{W}_n \in \mathcal{W}) = 1$. This argument is similar to the corresponding step in the

proof under strong identification, with modifications being necessary to accommodate the uniform convergence over Π.

Let $\{l_n\}$ be a sequence of integers with $l_n \in \{1, \ldots, b_n\}$ such that $l_n \to \infty$ and $l_n = o(b_n)$. Define

$$\mathcal{R}(h,\pi) = -\sum_{t=1}^{h} \left[\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi))\right]$$
$$\mathcal{U}_{n,s}(h,\pi) = \sum_{t=(s-1)b_n+1}^{(s-1)b_n+l_n} \left[\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi))\right]$$
$$\mathcal{Y}_{n,s}(h,\pi) = \sum_{t=(s-1)b_n+l_n+1}^{sb_n} \left[\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi))\right]$$

Observe that for $h < l_n$, $\sum_{t=(s-1)b_n+1+h}^{sb_n} \left[\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi))\right] = \mathcal{Y}_{n,s}(h,\pi) + \mathcal{U}_{n,s}(h,\pi) + \mathcal{R}(h,\pi)$ by construction. This implies

$$\begin{aligned}
&\frac{1}{n}\sum_{s=1}^{n/b_n}\left\{\sum_{t=(s-1)b_n+1+h}^{sb_n}\left(\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi))\right)\right\}\left\{\sum_{t=(s-1)b_n+1+\tilde{h}}^{sb_n}\left(\mathcal{E}_{t,\tilde{h}}(\tilde{\pi}) - E(\mathcal{E}_{t,\tilde{h}}(\tilde{\pi}))\right)\right\} \\
&= \frac{1}{n}\sum_{s=1}^{n/b_n}\left\{\mathcal{Y}_{n,s}(h,\pi) + \mathcal{U}_{n,s}(h,\pi) + \mathcal{R}(h,\pi)\right\}\left\{\mathcal{Y}_{n,s}(\tilde{h},\tilde{\pi}) + \mathcal{U}_{n,s}(\tilde{h},\tilde{\pi}) + \mathcal{R}(\tilde{h},\tilde{\pi}))\right\} \\
&= \frac{1}{n}\sum_{s=1}^{n/b_n}\mathcal{Y}_{n,s}(h,\pi)\mathcal{Y}_{n,s}(\tilde{h},\tilde{\pi}) + \frac{1}{n}\sum_{s=1}^{n/b_n}\mathcal{U}_{n,s}(h,\pi)\mathcal{U}_{n,s}(\tilde{h},\tilde{\pi}) + \frac{1}{n}\sum_{s=1}^{n/b_n}\mathcal{R}_{n,s}(h,\pi)\mathcal{R}_{n,s}(\tilde{h},\tilde{\pi}) \\
&+ \frac{1}{n}\sum_{s=1}^{n/b_n}\mathcal{Y}_{n,s}(h,\pi)\mathcal{U}_{n,s}(\tilde{h},\tilde{\pi}) + \frac{1}{n}\sum_{s=1}^{n/b_n}\mathcal{Y}_{n,s}(h,\pi)\mathcal{R}_{n,s}(\tilde{h},\tilde{\pi}) + \frac{1}{n}\sum_{s=1}^{n/b_n}\mathcal{U}_{n,s}(h,\pi)\mathcal{Y}_{n,s}(\tilde{h},\tilde{\pi}) \\
&+ \frac{1}{n}\sum_{s=1}^{n/b_n}\mathcal{U}_{n,s}(h,\pi)\mathcal{R}_{n,s}(\tilde{h},\tilde{\pi}) + \frac{1}{n}\sum_{s=1}^{n/b_n}\mathcal{R}_{n,s}(h,\pi)\mathcal{Y}_{n,s}(\tilde{h},\tilde{\pi}) + \frac{1}{n}\sum_{s=1}^{n/b_n}\mathcal{R}_{n,s}(h,\pi)\mathcal{U}_{n,s}(\tilde{h},\tilde{\pi}).
\end{aligned}$$

We prove

$$\begin{aligned}
&\frac{1}{n}\sum_{s=1}^{n/b_n}\left\{\sum_{t=(s-1)b_n+1+h}^{sb_n}\left(\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi))\right)\right\}\left\{\sum_{t=(s-1)b_n+1+\tilde{h}}^{sb_n}\left(\mathcal{E}_{t,\tilde{h}}(\tilde{\pi}) - E(\mathcal{E}_{t,\tilde{h}}(\tilde{\pi}))\right)\right\} \\
&= \frac{1}{n}\sum_{s=1}^{n/b_n}\mathcal{Y}_{n,s}(h,\pi)\mathcal{Y}_{n,s}(\tilde{h},\tilde{\pi}) + o_p(1). \qquad \text{(A.24)}
\end{aligned}$$

for every $\pi, \tilde{\pi} \in \Pi$. Stochastic equicontinuity follows from probability sub-additivity and because $\mathcal{D}^{\psi}(h,\pi)$, $H^{-1}(\pi;\gamma_0)$, and $m_t^{\psi}(\psi_{0,n},\pi)$ are each stochastically equicontinuous under Assumptions 9, 1, and 4 respectively.

Observe that

$$\begin{aligned}\frac{1}{n}\sum_{s=1}^{n/b_n}\mathcal{R}_{n,s}(h,\pi)\mathcal{R}_{n,s}(\tilde{h},\tilde{\pi}) &= \frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\frac{1}{b_n}\mathcal{R}_{n,s}(h,\pi)\mathcal{R}_{n,s}(\tilde{h},\tilde{\pi}) \\ &= \frac{1}{b_n}\mathcal{R}_{n,s}(h,\pi)\mathcal{R}_{n,s}(\tilde{h},\tilde{\pi}).\end{aligned}$$

Under Assumptions 7, 4, and 8, $\mathcal{E}_{t,h}(\pi)$ is stationary, ergodic, and L_2-bounded uniformly in π. Therefore

$$E\left|\left|\frac{1}{b_n}\mathcal{R}_{n,s}(h,\pi)\mathcal{R}_{n,s}(\tilde{h},\tilde{\pi})\right|\right| \leq K/b_n \to 0.$$

Next, the NED properties and moment bounds of ε_t and $m_t^{\psi}(\pi)$ in Assumptions 7, 4, and 8 imply that $\mathcal{E}_{t,h}(\pi) = \varepsilon_t\varepsilon_{t-h} - \mathcal{D}^{\psi}(h,\pi)'H^{-1}(\pi;\gamma_0)\left(m_t^{\psi}(\psi_{0,n},\pi) - E_{\gamma_n}[m_t^{\psi}(\psi_{0,n},\pi)]\right)$ is stationary, L_p-bounded for some $p>2$, and L_2-NED on an α-mixing base with decay rate $O(h^{-p/(p-2)})$. Then $||(1/\sqrt{b_n})\mathcal{Y}_{n,1}(h,\pi)||_2$ and $||(1/\sqrt{l_n})\mathcal{U}_{n,1}(h,\pi)||_2$ are $O(1)$ by Theorem 17.5 in Davidson (1994) and Theorem 1.6 in McLeish (1975). Observe

$$\begin{aligned}\left|\left|\frac{1}{n}\sum_{s=1}^{n/b_n}\mathcal{Y}_{n,s}(h)\mathcal{U}_{n,s}(\tilde{h},\tilde{\pi})\right|\right|_1 &= \left|\left|\frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\frac{l_n}{b_n}\mathcal{Y}_{n,s}(h,\pi)\frac{1}{l_n}\mathcal{U}_{n,s}(\tilde{h},\tilde{\pi})\right|\right|_1 \\ &= \left|\left|\frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\left(\frac{l_n}{b_n}\right)^{1/2}\frac{1}{\sqrt{b_n}}\mathcal{Y}_{n,s}(h,\pi)\frac{1}{\sqrt{l_n}}\mathcal{U}_{n,s}(\tilde{h},\tilde{\pi})\right|\right|_1 \\ &\leq \frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\left|\left|\left(\frac{l_n}{b_n}\right)^{1/2}\frac{1}{\sqrt{b_n}}\mathcal{Y}_{n,s}(h,\pi)\frac{1}{\sqrt{l_n}}\mathcal{U}_{n,s}(\tilde{h},\tilde{\pi})\right|\right|_1 \\ &\leq \frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\left(\frac{l_n}{b_n}\right)^{1/2}\left|\left|\frac{1}{\sqrt{b_n}}\mathcal{Y}_{n,s}(h,\pi)\right|\right|_2\left|\left|\frac{1}{\sqrt{l_n}}\mathcal{U}_{n,s}(\tilde{h},\tilde{\pi})\right|\right|_2\end{aligned}$$

$$= \left(\frac{l_n}{b_n}\right)^{1/2} \left\| \frac{1}{\sqrt{b_n}} \mathcal{Y}_{n,1}(h,\pi) \right\|_2 \left\| \frac{1}{\sqrt{l_n}} \mathcal{U}_{n,1}(\tilde{h},\tilde{\pi}) \right\|_2$$
$$= O\left(\left(\frac{l_n}{b_n}\right)^{1/2} \right) = o(1)$$

by stationarity, Minkowski's inequality, and the Cauchy-Schwartz inequality. The remaining terms are shown to be $o(1)$ in a similar fashion. This proves A.24.

Finally, by the NED property, we see by the proof of de Jong's (1997) Theorem 2 that

$$\frac{1}{n} \sum_{s=1}^{n/b_n} \mathcal{Y}_{n,s}(h,\pi) \mathcal{Y}_{n,s}(\tilde{h},\tilde{\pi})$$
$$\xrightarrow{p} \lim_{n\to\infty} \frac{1}{n} E\left[\left\{ \sum_{t=1}^{n} \left(\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi)) \right) \right\} \left\{ \sum_{t=1}^{n} \left(\mathcal{E}_{t,\tilde{h}}(\tilde{\pi}) - E(\mathcal{E}_{t,\tilde{h}}(\tilde{\pi})) \right) \right\} \right].$$

Combine this with A.24 to see that

$$nE(\rho_n^{1,*}(h,\pi)\rho_n^{1,*}(\tilde{h},\tilde{\pi})|\mathcal{W}_n)$$
$$\xrightarrow{p} \lim_{n\to\infty} E[nE(\rho_n^{1,*}(h,\pi)\rho_n^{1,*}(\tilde{h},\tilde{\pi})|\mathcal{W}_n)] = E[\mathcal{Z}^{1,\psi}(h,\pi)\mathcal{Z}^{1,\psi}(\tilde{h},\tilde{\pi})],$$

so that $P(\mathcal{W}_n \in \mathcal{W}) = 1$ as desired.

Now consider A.23. Recall $\hat{\mathcal{E}}_{t,h}(\hat{\psi}_{0,n},\pi) = \varepsilon_t(\hat{\psi}_{0,n},\pi)\varepsilon_{t-h}(\hat{\psi}_{0,n},\pi) - \hat{\mathcal{D}}_n(h,\pi)' H_n^{-1}(\hat{\psi}_{0,n},\pi)$ $\left(m_t(\hat{\psi}_{0,n},\pi) - \frac{1}{n}\sum_{t=1}^{n} m_t(\hat{\psi}_{0,n},\pi) \right) - \frac{1}{n}\sum_{t=1+h}^{n} [\varepsilon_t(\hat{\psi}_{0,n},\pi)\varepsilon_{t-h}(\hat{\psi}_{0,n},\pi) - \varepsilon_t\varepsilon_{t-h}]$ and $\mathcal{E}_{t,h}(\pi) = \varepsilon_t\varepsilon_{t-h} - \mathcal{D}^{\psi}(h,\pi)' H^{-1}(\pi;\gamma_0)$ $\left(m_t(\psi_{0,n},\pi) - E_{\gamma_n}[m_t(\psi_{0,n},\pi)] \right)$.

Observe that by the construction of z_t for ξ_s iid $N(0,1)$,

$$E\left[\left(\frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t \left(\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi)) \right) \right)^2 \right]$$
$$= E\left[\left(\frac{1}{\sqrt{n}} \sum_{s=1}^{n/b_n} \xi_s \sum_{t=(s-1)b_n+1}^{sb_n} \left(\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi)) \right) \right)^2 \right]$$
$$= E\left[\left(\frac{1}{\sqrt{b_n}} \sum_{t=1}^{b_n} \left(\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi)) \right) \right)^2 \right].$$

Then under Assumptions 4, 8, and 9, $\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi))$ is zero mean, stationary, L_p-bounded for some $p > 2$, and L_2-NED with size $-1/2$ on an α-mixing base with decay rate $O(h^{-p/(p-2)-\iota})$ by Theorems 17.8 and 17.9 in Davidson (1994). Hence the term above is $O_\pi(1)$, so that $\frac{1}{\sqrt{b_n}} \sum_{t=1}^{b_n} \left(\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi))\right) = O_{p,\pi}(1)$.

We will next prove

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t \Big(\hat{\mathcal{E}}_{t,h}(\hat{\psi}_{0,n}, \pi) - \frac{1}{n} \sum_{s=1+h}^{n} \hat{\mathcal{E}}_{t,h}(\hat{\psi}_{0,n}, \pi) \Big) \\
&\qquad = \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t \big(\mathcal{E}_{t,h}(\pi) - E_{\gamma_n}(\mathcal{E}_{t,h}(\pi)) \big) + o_{p,\pi}(1)
\end{aligned} \tag{A.25}$$

which coupled with the previous arguments, uniform convergence of $\hat{\mathcal{D}}_n$ and K_n, and Assumption 9(v) give

$$\begin{aligned}
&\sqrt{n}\hat{\rho}_n^{(w)}(h, \pi) \\
&= \frac{1}{n^{-1} \sum_{t=1}^{n} \varepsilon_t^2(\hat{\theta}_n)} \times \Bigg\{ \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t \Big(\hat{\mathcal{E}}_{t,h}(\hat{\psi}_{0,n}, \pi) - \frac{1}{n} \sum_{t=1+h}^{n} \hat{\mathcal{E}}_{t,h}(\hat{\psi}_{0,n}, \pi) \Big) \\
&\qquad + \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} [\varepsilon_t(\hat{\psi}_{0,n}, \pi)\varepsilon_{t-h}(\hat{\psi}_{0,n}, \pi) - \varepsilon_t\varepsilon_{t-h}] \\
&\qquad - \hat{\mathcal{D}}_n(h, \pi)' H_n^{-1}(\hat{\psi}_{0,n}, \pi) b K_n(\tilde{\psi}_{0,n}, \pi) \Bigg\} \\
&= \frac{1}{E(\varepsilon_t^2)} \times \Bigg\{ \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t \Big(\mathcal{E}_{t,h}(\pi) - E(\mathcal{E}_{t,h}(\pi)) \Big) \\
&\qquad + \sqrt{n} E_{\gamma_n}[\varepsilon_t(\psi_{0,n}, \pi)\varepsilon_{t-h}(\psi_{0,n}, \pi) - \varepsilon_t\varepsilon_{t-h}] \\
&\qquad - \mathcal{D}^{\psi}(h, \pi)' H^{-1}(\pi; \gamma_0) b K(\psi_0, \pi; \gamma_0) \Bigg\} + o_{p,\pi}(1)
\end{aligned}$$

In order to prove A.25, we must show the following steps:

$$\frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t(\varepsilon_t(\hat{\psi}_{0,n}, \pi)\varepsilon_{t-h}(\hat{\psi}_{0,n}, \pi) - \varepsilon_t\varepsilon_{t-h})$$

$$= \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t E_{\gamma_n}(\varepsilon_t(\hat{\psi}_{0,n}, \pi)\varepsilon_{t-h}(\hat{\psi}_{0,n}, \pi) - \varepsilon_t\varepsilon_{t-h}) + o_p(1) \tag{A.26}$$

$$\hat{\mathcal{D}}_n(h,\pi)' H_n^{-1}(\hat{\psi}_{0,n}, \pi) \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t \Big(m_t^{\psi}(\hat{\psi}_{0,n}, \pi) - \frac{1}{n} \sum_{s=1}^{n} m_s^{\psi}(\hat{\psi}_{0,n}, \pi) \Big)$$

$$= \mathcal{D}^{\psi}(h,\pi)' H^{-1}(\pi; \gamma_0) \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t \Big(m_t^{\psi}(\psi_{0,n}, \pi) - E_{\gamma_n}[m_s^{\psi}(\psi_{0,n}, \pi)] \Big)$$

$$+ o_{p,\pi}(1) \tag{A.27}$$

$$\frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t \frac{1}{n} \sum_{s=1+h}^{n} \varepsilon_s \varepsilon_{s-h} = \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t E[\varepsilon_s \varepsilon_{s-h}] + o_{p,\pi}(1) \tag{A.28}$$

$$\frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t \frac{1}{n} \sum_{s=1+h}^{n} (\varepsilon_s(\hat{\psi}_{0,n}, \pi)\varepsilon_{s-h}(\hat{\psi}_{0,n}, \pi) - \varepsilon_s\varepsilon_{s-h}) = o_{p,\pi}(1) \tag{A.29}$$

$$\frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t \hat{\mathcal{D}}_n(h,\pi)' H_n^{-1}(\hat{\psi}_{0,n}, \pi) \frac{1}{n} \sum_{s=1+h}^{n} \Big(m_s^{\psi}(\hat{\psi}_{0,n}, \pi) - \frac{1}{n} \sum_{u=1}^{n} m_u^{\psi}(\hat{\psi}_{0,n}, \pi) \Big)$$

$$= \mathcal{D}^{\psi}(h,\pi)' H^{-1}(\pi; \gamma_0) \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t E_{\gamma_n}[m_s^{\psi}(\psi_{0,n}, \pi) - E_{\gamma_n}[m_s^{\psi}(\psi_{0,n}, \pi)]]$$

$$+ o_{p,\pi}(1)$$

$$= o_{p,\pi}(1) \tag{A.30}$$

Consider A.26. Since z_t is a mean zero Gaussian random variable that is independent of the sample, the proof of Lemma A.2.4 applies to show

$$\frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t(\varepsilon_t(\hat{\psi}_{0,n}, \pi)\varepsilon_{t-h}(\hat{\psi}_{0,n}, \pi) - \varepsilon_t\varepsilon_{t-h})$$

$$= \frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t(\varepsilon_t(\psi_{0,n}, \pi)\varepsilon_{t-h}(\psi_{0,n}, \pi) - \varepsilon_t\varepsilon_{t-h})$$

$$+ \sqrt{n}(\hat{\zeta}_n - \zeta_n)' \frac{1}{n} \sum_{t=1+h}^{n} z_t[d_{\zeta,t}(\pi)\varepsilon_{t-h}(\psi_{0,n}, \pi) + d_{\zeta,t-h}(\pi)\varepsilon_t(\psi_{0,n}, \pi)]$$

$$+ o_{p,\pi}(1).$$

Let S_ζ be the ζ selection matrix that selects the rows of ψ corresponding to ζ. Recall that $\sqrt{n}(\hat{\zeta}_n -$

$\zeta_n) = \sqrt{n}(\hat{\psi}_n(\pi) - \psi_n)S_\zeta = O_{p,\pi}(1)$ and observe

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1+h}^{n} z_t[d_{\zeta,t}(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)] \\
&\qquad = \frac{1}{n}\sum_{t=1}^{n} z_t[d_{\zeta,t}(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)] + o_{p,\pi}(1) \\
&\qquad = \frac{1}{n}\sum_{s=1}^{n/b_n}\xi_s \sum_{t=(s-1)b_n+1}^{sb_n} [d_{\zeta,t}(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)] + o_{p,\pi}(1)
\end{aligned}$$

by the moment bounds in Assumption 9 and the construction of z_t. Recall ξ_s is iid, independent of the sample, and has zero mean and unit variance. Then stationarity, and the moment bounds imply that for each $\pi \in \Pi$,

$$\begin{aligned}
&E_{\gamma_n}\Big[\Big(\frac{1}{n}\sum_{s=1}^{n/b_n}\xi_s \sum_{t=(s-1)b_n+1}^{sb_n} [d_{\zeta,t}(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)]\Big)^2\Big] \\
&\qquad = \frac{b_n}{n}E_{\gamma_n}\Big[\Big(\frac{1}{b_n}\sum_{t=1}^{b_n}[d_{\zeta,t}(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)]\Big)^2\Big] \\
&\qquad \leq \frac{b_n}{n}E_{\gamma_n}\Big[[d_{\zeta,t}(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)]^2\Big] = o(1)
\end{aligned}$$

because $b_n = o(n)$. Now apply Chebyshev's inequality to see

$$P(\frac{1}{n}\sum_{t=1+h}^{n} z_t[d_{\zeta,t}\varepsilon_{t-h}(\psi_{0,n},\pi)] > \eta) \leq \frac{b_n}{n}E\Big[[d_{\zeta,t}(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)]^2\Big]/\eta^2 \to 0,$$

so that $\frac{1}{n}\sum_{t=1+h}^{n} z_t[d_{\zeta,t}(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)] \xrightarrow{p} 0$ point-wise on Π. Stochastic Equicontinuity follows from a mean value theorem argument and the moment bounds in Assumption 9 (See the proof of Lemma A.2.4). This implies that $\sqrt{n}(\hat{\zeta}_n - \zeta_n)' \frac{1}{n}\sum_{t=1+h}^{n} z_t\,[d_{\zeta,t}(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi) + d_{\zeta,t-h}(\psi_{0,n},\pi)\varepsilon_t(\psi_{0,n},\pi)] = o_{p,\pi}(1)$, so that

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t(\varepsilon_t(\hat{\psi}_{0,n},\pi)\varepsilon_{t-h}(\hat{\psi}_{0,n},\pi) - \varepsilon_t\varepsilon_{t-h}) \\
&\qquad = \frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t(\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi) - \varepsilon_t\varepsilon_{t-h}) + o_{p,\pi}(1).
\end{aligned}$$

Now use the construction of z_t to see

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t(\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h}) \\
&\qquad=\frac{1}{\sqrt{n/b_n}}\sum_{s=1}^{n/b_n}\xi_s\left[\sqrt{b_n}\frac{1}{b_n}\sum_{t=(s-1)b_n+1}^{sb_n}(\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h})\right] \\
&\qquad=\frac{1}{\sqrt{n/b_n}}\sum_{s=1}^{n/b_n}\xi_s\left[\sqrt{b_n}E_{\gamma_n}(\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h})\right]+o_p(1) \\
&\qquad=\frac{1}{\sqrt{n}}\sum_{t=1}^{n} z_t E_{\gamma_n}(\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h})+o_p(1)
\end{aligned}$$

because ξ_s are independent of the data and by stationarity, ergodicity, and moment bounds in Assumptions 7, 8, and 9. Hence A.26 holds.

Next, consider A.27.

$$\begin{aligned}
&\hat{\mathcal{D}}_n(h,\pi)'H_n^{-1}(\hat{\psi}_{0,n},\pi)\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t m_t^{\psi}(\hat{\psi}_{0,n},\pi) \\
&\qquad=\mathcal{D}^{\psi}(h,\pi)'H^{-1}(\pi;\gamma_0)\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t m_t^{\psi}(\psi_{0,n},\pi) \\
&\qquad+\mathcal{D}^{\psi}(h,\pi)'H^{-1}(\pi;\gamma_0)\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t\big(m_t^{\psi}(\hat{\psi}_{0,n},\pi)-m_t^{\psi}(\psi_{0,n},\pi)\big) \\
&\qquad+\Big(\hat{\mathcal{D}}_n(h,\pi)'H_n^{-1}(\hat{\psi}_{0,n},\pi)-\mathcal{D}^{\psi}(h,\pi)'H^{-1}(\pi;\gamma_0)\Big)\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t m_t^{\psi}(\psi_{0,n},\pi) \\
&\qquad+\Big(\hat{\mathcal{D}}_n(h,\pi)'H_n^{-1}(\hat{\psi}_{0,n},\pi)-\mathcal{D}^{\psi}(h,\pi)'H^{-1}(\pi;\gamma_0)\Big) \\
&\qquad\qquad\times\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t\big(m_t^{\psi}(\hat{\psi}_{0,n},\pi)-m_t^{\psi}(\psi_{0,n},\pi)\big)
\end{aligned}$$

Recall that $\sup_{\pi\in\Pi}||\hat{\mathcal{D}}_n(h,\pi)-\mathcal{D}^{\psi}(h,\pi)||\xrightarrow{p}0$ by Lemma A.2.5, and $\sup_{\pi\in\Pi}||H_n(\hat{\psi}_{0,n},\pi)-H(\pi;\gamma_0)||\xrightarrow{p}0$ by assumption; hence $\hat{\mathcal{D}}_n(h,\pi)'H_n^{-1}(\hat{\psi}_{0,n},\pi)-\mathcal{D}^{\psi}(h,\pi)'H^{-1}(\pi;\gamma_0)=o_{p,\pi}(1)$. Next, observe that $\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t m_t^{\psi}(\psi_{0,n},\pi)=O_p(1)$ point-wise on Π by the moment bounds, the NED property, Davidson (1994), Theorem 17.5 and McLeish (1975), Theorem 1.6, and Chebyshev's inequality.

Finally, observe

$$||\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t\big(m_t^{\psi}(\hat{\psi}_{0,n},\pi) - m_t^{\psi}(\psi_{0,n},\pi)\big)||$$
$$\leq \sup_{\zeta^*\in\mathcal{Z}} ||\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t\frac{\partial}{\partial\zeta}m_t^{\psi}(0,\zeta^*,\pi)|| \times ||\hat{\zeta}_n - \zeta_n||$$

by the mean value theorem. Since $||\hat{\psi}_n(\pi) - \psi_n|| = O_{p,\pi}(1/\sqrt{n})$, it remains to show that $\sup_{\pi\in\Pi}\sup_{\zeta^*\in\mathcal{Z}}||\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t\frac{\partial}{\partial\zeta}m_t^{\psi}(0,\zeta^*,\pi)|| = o_p(\sqrt{n})$, which is shown in Lemma A.2.2.

Now following the argument above, we need only prove

$$\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t\frac{1}{n}\sum_{s=1}^{n} m_s^{\psi}(\hat{\psi}_{0,n},\pi) = \frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t E_{\gamma_n}[m_s^{\psi}(\psi_{0,n},\pi)] + o_p(1),$$

which follows from an identical argument.

Next we prove step A.28. Recall that by Davidson (1994), Theorems 17.9, $\{\varepsilon_t\varepsilon_{t-h}\}$ is zero mean, stationary, L_p-bounded for some $p > 2$, and L_2-NED with size $-1/2$ on an α-mixing base with decay rate $O(h^{-p/(p-2)-\iota})$, so $E[\big(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\varepsilon_t\varepsilon_{t-h} - E[\varepsilon_t\varepsilon_{t-h}])\big)^2] = O(1)$ by Davidson (1994), Theorem 17.5 and McLeish (1975), Theorem 1.6. Hence, $\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\varepsilon_t\varepsilon_{t-h} - E[\varepsilon_t\varepsilon_{t-h}] = O_p(1)$, and $\frac{1}{n}\sum_{t=1+h}^{n}\varepsilon_t\varepsilon_{t-h} - E[\varepsilon_t\varepsilon_{t-h}] = O_p(1/\sqrt{n})$.

Next, we prove A.29. Recall

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} z_t = \frac{\sqrt{b_n}}{\sqrt{n/b_n}}\sum_{t=1}^{n/b_n}\xi_t = O_p(\sqrt{b_n}).$$

Recall that $\varepsilon_t(\theta)$ is a continuous function and $\varepsilon_t(\psi^*,\pi)$ does not depend on π for all ψ^* with $\beta^* = 0$, and $\sup_{\pi\in\Pi}||\hat{\psi}_{0,n}(\pi) - \psi_{0,n}|| \xrightarrow{p} 0$. Then by the moment bounds, stationarity and ergodicity in Assumptions 7 and 8,

$$\sup_{\pi\in\Pi}||\frac{1}{n}\sum_{s=1+h}^{n}(\varepsilon_s(\hat{\psi}_{0,n},\pi)\varepsilon_{s-h}(\hat{\psi}_{0,n},\pi) - \varepsilon_s\varepsilon_{s-h}) - E_{\gamma_n}(\varepsilon_s(\psi_{0,n},\pi)\varepsilon_{s-h}(\psi_{0,n},\pi) - \varepsilon_s\varepsilon_{s-h})||$$
$$\xrightarrow{p} 0.$$

Now recall that $\sqrt{n}E_{\gamma_n}(\varepsilon_s(\psi_{0,n},\pi)\varepsilon_{s-h}(\psi_{0,n},\pi)-\varepsilon_s\varepsilon_{s-h})=O(1)$ by Assumption 9 and that the construction of z_t has ξ_s iid, mean zero, and independent of the data, so

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1+h}^{n} z_t\big[\sqrt{n}E_{\gamma_n}(\varepsilon_s(\psi_{0,n},\pi)\varepsilon_{s-h}(\psi_{0,n},\pi)-\varepsilon_s\varepsilon_{s-h})\big] \\
&\qquad=\frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\xi_s\frac{1}{b_n}\sum_{t=(s-1)b_n+1}^{sb_n}\big[\sqrt{n}E_{\gamma_n}(\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h})\big] \\
&\qquad=\frac{1}{n/b_n}\sum_{s=1}^{n/b_n}\xi_s\big[\sqrt{n}E_{\gamma_n}(\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h})\big] \\
&\qquad=o_p(1).
\end{aligned}$$

Finally, we prove A.30:

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_t\hat{\mathcal{D}}_n(h,\pi)'H_n^{-1}(\hat{\psi}_{0,n},\pi)\frac{1}{n}\sum_{s=1+h}^{n} m_s(\hat{\psi}_{0,n},\pi) \\
&\qquad=\mathcal{D}^{\psi}(h,\pi)'H^{-1}(\pi;\gamma_0)\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n} z_tE_{\gamma_n}[m_s^{\psi}(\pi)]+o_{p,\pi}(1) \\
&\qquad=o_{p,\pi}(1)
\end{aligned}$$

First, recall $\frac{1}{\sqrt{n}}\sum_{t=1}^{n} z_t = O_p(\sqrt{b_n})$ and $\sup_{\pi\in\Pi}||\hat{\mathcal{D}}_n(h,\pi)'H_n^{-1}(\hat{\psi}_{0,n},\pi)-\mathcal{D}^{\psi}(h,\pi)'H^{-1}(\pi;\gamma_0)||\xrightarrow{p}0$. It is therefore sufficient to show $\frac{1}{n}\sum_{t=1}^{n} m_t^{\psi}(\hat{\psi}_{0,n},\pi)=E_{\gamma_n}[m_s^{\psi}(\pi)]+o_{p,\pi}(1)$.

Next, recall the expansion

$$m_t^{\psi}(\psi_{0,n},\pi)=m_t^{\psi}(\psi_{0,n},\pi)-E_{\gamma_n}[m_t^{\psi}(\psi_{0,n},\pi)]+\beta_n\frac{\partial}{\partial\tilde{\beta}}E_{\tilde{\gamma}_n}[m_t^{\psi}(\psi_{0,n},\pi)]$$

for some $\tilde{\gamma}_n$ such that $||\tilde{\gamma}_n-\gamma_n||\leq||\gamma_{0,n}-\gamma_n||$.

Next, recall $\{m_t(\psi_{0,n},\pi)-E_{\gamma_n}[m_t(\psi_{0,n},\pi)]\}$ is zero mean, stationary, L_p-bounded for some $p>2$, and L_2-NED with size $-1/2$ on an α-mixing base with decay rate $O(h^{-p/(p-2)-\iota})$, so

$E[\big(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(m_t(\psi_{0,n},\pi)-E_{\gamma_n}[m_t(\psi_{0,n},\pi)])\big)^2] = O(1)$ point-wise on Π by McLeish (1975), Theorem 1.6. Hence, $\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(m_t(\psi_{0,n},\pi)-E_{\gamma_n}[m_t(\psi_{0,n},\pi)] = O_p(1)$, and $\frac{1}{n}\sum_{t=1+h}^{n}(m_t(\psi_{0,n},\pi)-E_{\gamma_n}[m_t(\psi_{0,n},\pi)]) = O_p(1/\sqrt{n})$ point-wise on Π. It remains to show stochastic equicontinuity, which follows from a mean value theorem argument paired with the uniform moment bounds in Assumption 4.

For the last term in the expansion, observe that $\frac{1}{n}\sum_{t=1}^{n}E_{\tilde{\gamma}_n}[m_t(\psi_{0,n},\pi)] \equiv K_n(\tilde{\psi}_n,\pi;\tilde{\gamma}_n) \to K(\psi_0,\pi;\gamma_0)$ uniformly over $\pi \in \Pi$ by Assumptions 1 and because $\beta_n = O(1/\sqrt{n})$ under weak identification. Thus, $\frac{1}{n}\sum_{t=1}^{n}\beta_n\frac{\partial}{\partial\tilde{\beta}}E_{\tilde{\gamma}_n}[m_t(\psi_{0,n},\pi)] = O(1/\sqrt{n})$.

Finally, observe that

$$||\frac{1}{n}\sum_{t=1+h}^{n}m_t(\hat{\psi}_{0,n},\pi) - \frac{1}{n}\sum_{t=1+h}^{n}m_t(\psi_{0,n},\pi)|| \leq \sup_{\pi\in\Pi}\sup_{\zeta^*\in\mathcal{Z}}||\frac{\partial}{\partial\zeta}m_t(0,\zeta^*,\pi)|| \times ||\hat{\zeta}_n - \zeta_n||$$

by the mean value theorem. Now recall that $||\hat{\zeta}_n - \zeta_n|| = O_{p,\pi}(1/\sqrt{n})$ and the moment bounds in Assumption 4 to complete the proof.

This completes the proof of A.25. Hence, A.23 holds, so combined with A.22, we see that A.20 holds:

$$\{\sqrt{n}\hat{\rho}_n^{(w)}(h,\pi):\ 1\leq h\leq\mathcal{L},\ \pi\in\Pi\} \Rightarrow^p \{\mathring{\mathcal{Z}}\ (h,\pi):\ 1\leq h\leq\mathcal{L},\ \pi\in\Pi\}$$

It remains to prove the joint convergence result A.21:

$$\{\sqrt{n}\hat{\rho}_n^{(w)}(h,\pi),\pi^*_{(bs)}(b,\gamma_0) : 1\leq h\leq\mathcal{L},\ \pi\in\Pi\}$$
$$\Rightarrow^p \{\mathring{\mathcal{Z}}\ (h,\pi),\pi^*(b,\gamma_0):\ 1\leq h\leq\mathcal{L},\ \pi\in\Pi\}$$

which will follows simply from the construction of $\hat{\rho}_n^{(w)}(h,\pi)$, the fact that $\pi^*_{(bs)}(b,\gamma_0)$ is a continuous function of $\hat{\sigma}_n\hat{G}_n^{(bs)}(\pi)$, $K_n(\hat{\psi}_{0,n},\pi;\gamma_0)$, and $H_n(\psi_{0,n},\pi)$, and the continuous mapping

theorem. The result A.16 follows. Hence,

$$\{\sqrt{n}\hat{\rho}_n^{(w)}(h):\ 1\leq h\leq \mathcal{L}\} \Rightarrow^p \{\overset{\circ}{\mathcal{Z}}\ (h,\pi^*(b,\gamma_0)):\ 1\leq h\leq \mathcal{L}\}$$

for each $\mathcal{L}\in\mathbb{N}$, where $\{\overset{\circ}{\mathcal{Z}}\ (h,\pi):h\in\mathbb{N},\ \pi\in\Pi\}$ is an independent copy of $\{\mathcal{Z}^{\psi}(h,\pi):h\in\mathbb{N},\ \pi\in\Pi\}$, the zero mean Gaussian process in Lemma 3.2(a).

Now we prove A.17:

$$\sup_{c>0}\left|P(\max_{1\leq h\leq\mathcal{L}_n}|\sqrt{n}\hat{\rho}_n^{(w)}(h)|\leq c|\mathcal{W}_n)-P(\max_{1\leq h\leq\mathcal{L}_n}|\ \overset{\circ}{\mathcal{Z}}\ (h,\pi^*(b,\gamma_0))|\leq c)\right|\xrightarrow{p}0 \qquad (A.31)$$

for the process $\{\overset{\circ}{\mathcal{Z}}\ (h,\pi):\ 1\leq h\leq\mathcal{L},\ \pi\in\Pi\}$ and some sequence of positive integers $\{\mathcal{L}_n\}$, $\mathcal{L}_n\to\infty$ and $\mathcal{L}_n=o(n)$.

This follows the proof of A.7 exactly by defining

$$\mathcal{A}_{\mathcal{L},n}\equiv\sup_{c>0}\left|P(\max_{1\leq h\leq\mathcal{L}}|\sqrt{n}\hat{\rho}_n^{(w)}(h)|\leq c|\mathcal{W}_n)-P(\max_{1\leq h\leq\mathcal{L}}|\ \overset{\circ}{\mathcal{Z}}\ (h,\pi^*(b,\gamma_0))|\leq c)\right|\xrightarrow{p}0.$$

Step 3. Finally, we show the consistency of the critical values. Define the quantile functions $\hat{F}_n^{-1}(u|\cdot)=\inf\{c\geq 0:\ P(\hat{\mathcal{T}}_n^{(w)}\leq c|\cdot)\geq u\}$, $F_n^{-1}(u)=\inf\{c\geq 0:\ P(\hat{\mathcal{T}}_n\leq c)\geq u\}$.

Operate conditionally on the sample $\mathcal{W}_n$. From A.17, $\{\hat{\mathcal{T}}_{n,j}^{(w)}(\gamma_n,b)\}_{j=1}^M$ is a sequence of iid draws from $\max_{1\leq h\leq\mathcal{L}_n}|\ \overset{\circ}{\mathcal{Z}}\ (h,\pi^*(b,\gamma_0))|$ asymptotically with probability approaching one with respect to the sample $\mathcal{W}_n$. Thus under $\{\gamma_n\}\in\Gamma(\gamma_0,0,b)$ with $||b||<\infty$, $\hat{\mathcal{T}}_n$ and $\hat{\mathcal{T}}_n^{(w)}(\gamma_n,b)$ have the same limits under H_0. Hence, under H_0,

$$\sup_{c>0}\left|P(\hat{\mathcal{T}}_n^{(w)}(\gamma_n,b)\leq c|\mathcal{W}_n)-P(\hat{\mathcal{T}}_n\leq c)\right|\xrightarrow{p}0.$$

Therefore $\sup_{u\in[0,1]}\left|\hat{F}_n^{-1}(u|\mathcal{W}_n)-F_n^{-1}(u)\right|\xrightarrow{p}0$. Further, by independence and letting $M_n\to\infty$, the bootstrapped critical value $\hat{c}_{n,1-\alpha,M_n}^{(w)}(\gamma_n,b)=\hat{\mathcal{T}}_{n,[(1-\alpha)\cdot M_n]}^{(w)}(\gamma_n,b)$ is a central order statistic (see e.g. Galambos (1987)) of a conditionally iid random variable, so $\left|\hat{c}_{n,1-\alpha,M_n}^{(w)}(\gamma_n,b)-\hat{F}_n^{-1}(1-\alpha|\mathcal{W}_n)\right|\xrightarrow{p}0$. Combining these statements yields $\left|\hat{c}_{n,1-\alpha,M_n}^{(w)}(\gamma_n,b)-F_n^{-1}(1-\alpha)\right|\xrightarrow{p}0$. Since

$c_{n,1-\alpha}(\gamma_n, b) = F_n^{-1}(1-\alpha)$, the proof is complete.

Finally, we must show that under the alternative hypothesis, $P(\hat{T}_n > \hat{c}^{(k)}_{1-\alpha,n}) \to 1$ for $k = w, s$. Under the alternative, $\rho(h) \neq 0$ for some $h \in \mathbb{N}$.

Let $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$. By the triangle inequality and Theorem 2.3.2,

$$\begin{aligned}\max_{1\leq h\leq \mathcal{L}_n} |\sqrt{n}\hat{\rho}_n(h)| &\leq \max_{1\leq h\leq \mathcal{L}_n} |\sqrt{n}(\hat{\rho}_n(h) - \rho(h))| + \max_{1\leq h\leq \mathcal{L}_n} |\sqrt{n}\rho(h)| \\ &= \max_{1\leq h\leq \mathcal{L}_n} (|\mathcal{Z}_n^{\theta}(h)|) + \max_{1\leq h\leq \mathcal{L}_n} |\sqrt{n}\rho(h)| + o_p(1) \xrightarrow{p} \infty.\end{aligned}$$

Similarly, if $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $||b|| < \infty$, then

$$\begin{aligned}\max_{1\leq h\leq \mathcal{L}_n} \sup_{\pi\in\Pi} |\sqrt{n}\hat{\rho}_n(h,\pi)| &\leq \max_{1\leq h\leq \mathcal{L}_n} \sup_{\pi\in\Pi} (\sqrt{n}|\hat{\rho}_n(h;\pi) - \rho(h)|) + \max_{1\leq h\leq \mathcal{L}_n} |\sqrt{n}\rho(h)| \\ &= \max_{1\leq h\leq \mathcal{L}_n} \sup_{\pi\in\Pi} (|\mathcal{Z}_n^{\psi}(h,\pi)|) + \max_{1\leq h\leq \mathcal{L}_n} |\sqrt{n}\rho(h)| + o_p(1) \xrightarrow{p} \infty.\end{aligned}$$

Then using arguments above,

$$\begin{aligned}&P(\hat{T}_n \geq \hat{c}^{(k)}_{1-\alpha,n} | \mathcal{W}_n) \\ &\quad \geq \min\{P(\max_{1\leq h\leq \mathcal{L}_n} | \overset{\circ}{\mathcal{Z}}^{\psi}(h, \pi^*(b,\gamma_0))| \leq \hat{T}_n), P(\max_{1\leq h\leq \mathcal{L}_n} | \overset{\circ}{\mathcal{Z}}^{\theta}(h)| \leq \hat{T}_n)\} + o_p(1) \\ &\quad \to 1.\end{aligned}$$

□

A.2 Appendix: Supporting Lemmas and Proofs

A.2.1 Lemmas and Proofs relating to ULLNs for m_t

Lemma A.2.1. *Under Assumption 4, and* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,

$$\sup_{\theta\in\Theta} ||\frac{1}{n}\sum_{t=1}^{n} z_t \frac{\partial}{\partial\theta} m_t(\theta)|| \xrightarrow{p} 0 \tag{A.32}$$

$$\sup_{\theta\in\Theta}||\frac{1}{n}\sum_{t=1}^{n}\frac{\partial}{\partial\theta}m_t(\theta)-E\big[\frac{\partial}{\partial\theta}m_t(\theta)\big]||\xrightarrow{p}0 \tag{A.33}$$

$$\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}z_t m_t^{\theta}=O_p(1). \tag{A.34}$$

Proof. Consider the first statement. For iid $\xi_s\sim N(0,1)$ that is independent of the data

$$\frac{1}{n}\sum_{t=1}^{n}z_t\frac{\partial}{\partial\theta}m_t(\theta)=\frac{b_n}{n}\sum_{s=1}^{n/b_n}\xi_s\frac{1}{b_n}\sum_{t=(s-1)b_n+1}^{b_n}\frac{\partial}{\partial\theta}m_t(\theta),$$

and $\frac{\partial}{\partial\theta}m_t(\theta)$ is uniformly integrable on Θ under Assumption 4. Thus, for each $i=1,\dots,k_\theta$ and $j=1,\dots,k_m$, stationarity and Minkowski's inequality imply

$$\begin{aligned}E\Big[\Big(\frac{1}{n}\sum_{t=1}^{n}\frac{\partial}{\partial\theta_i}m_{j,t}(\theta)\Big)^2\Big]&=E\Big[\Big(\frac{b_n}{n}\sum_{s=1}^{n/b_n}\xi_s\frac{1}{b_n}\sum_{t=(s-1)b_n+1}^{b_n}\frac{\partial}{\partial\theta_i}m_{j,t}(\theta)\Big)^2\Big]\\&=\Big(\frac{b_n}{n}\Big)^2E\Big[\Big(\frac{1}{b_n}\sum_{t=(s-1)b_n+1}^{b_n}\frac{\partial}{\partial\theta_i}m_{j,t}(\theta)\Big)^2\Big]\\&\le\Big(\frac{b_n}{n}\Big)^2E\Big[\Big(\sup_{\theta\in\Theta}\frac{\partial}{\partial\theta_i}m_{j,t}(\theta)\Big)^2\Big]\to 0,\end{aligned}$$

so we see that $\frac{1}{n}\sum_{t=1}^{n}z_t\frac{\partial}{\partial\theta}m_t(\theta)\xrightarrow{p}0$ pointwise on Θ. Further, $\frac{1}{n}\sum_{t=1}^{n}z_t\frac{\partial}{\partial\theta}m_t(\theta)$ is stochastically equicontinuous. Observe that by the mean value theorem

$$\begin{aligned}\sup_{|\theta-\tilde\theta|<\delta}\Big|\frac{1}{n}\sum_{t=1}^{n}z_t\Big(\frac{\partial}{\partial\theta}m_t(\theta)-\frac{\partial}{\partial\theta}m_t(\tilde\theta)\Big)\Big|&\le\sup_{|\theta-\tilde\theta|<\delta}\Big|\frac{1}{n}\sum_{t=1}^{n}z_t\Big(\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta'}m_t(\theta^*)\times|\theta-\tilde\theta|\Big)\Big|\\&\le\sup_{\theta\in\Theta}\Big|\frac{1}{n}\sum_{t=1}^{n}z_t\Big(\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta'}m_t(\theta)\times\delta\Big)\Big|.\end{aligned}$$

It follows that, given $(\epsilon,\eta)>0$, there is a $\delta\in(0,\eta\epsilon/E[\sup_{\theta\in\Theta}|(\partial^2/\partial\theta\partial\theta')m_t(\theta)|])$ such that

$$\begin{aligned}&\lim_{n\to\infty}P\Big(\sup_{|\theta-\tilde\theta|<\delta}\Big|\frac{1}{n}\sum_{t=1}^{n}z_t\Big(\frac{\partial}{\partial\theta}m_t(\theta)-\frac{\partial}{\partial\theta}m_t(\tilde\theta)\Big)\Big|>\eta\Big)\\&\quad\le\lim_{n\to\infty}P\Big(\sup_{\theta\in\Theta}\Big|\frac{1}{n}\sum_{t=1}^{n}z_t\Big(\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta'}m_t(\theta)\Big)>\frac{\eta}{\delta}\Big)\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\delta}{\eta} \times \lim_{n\to\infty} E\Big(\sup_{\theta\in\Theta} \Big| \frac{1}{n} \sum_{t=1}^{n} z_t \Big(\frac{\partial}{\partial\theta} \frac{\partial}{\partial\theta'} m_t(\theta) \Big) \Big| \Big) \\
&= \frac{\delta}{\eta} \times \lim_{n\to\infty} E\Big(\sup_{\theta\in\Theta} \Big| \frac{1}{n/b_n} \sum_{s=1}^{n/b_n} \xi_s \frac{1}{b_n} \sum_{t=(s-1)b_n+1}^{sb_n} \Big(\frac{\partial}{\partial\theta} \frac{\partial}{\partial\theta'} m_t(\theta) \Big) \Big| \Big) \\
&\leq \frac{\delta}{\eta} \times \lim_{n\to\infty} E\Big(\sup_{\theta\in\Theta} \Big| \frac{\partial}{\partial\theta} \frac{\partial}{\partial\theta'} m_t(\theta) \Big| \Big) < \epsilon.
\end{aligned}
$$

The second inequality follows from Markov, and the final inequality follows from the uniform integrability of $\sup_{\theta\in\Theta} \left| \frac{\partial}{\partial\theta} \frac{\partial}{\partial\theta'} m_t(\theta) \right|$ in Assumption 4. Hence, $\frac{1}{n} \sum_{t=1}^{n} z_t \frac{\partial}{\partial\theta} m_t(\theta)$ is stochastically equicontinuous. Corollary 3.1 in Newey (1991) gives the desired result.

The second statement follows similarly to the first.

Consider now the final statement. Recall that m_t^θ is zero mean, stationary, L_p-bounded for some $p > 2$, and L_2-NED with size $1/2$ on an α-mixing base with decay $O(h^{-p/(p-2)-\iota})$. Then by Theorem 17.5 in Davidson (1994) and Theorem 1.6 in McLeish (1975) $E[(1/\sqrt{n} \times \sum_{t=1}^{n} m_{i,t}^\theta)^2] = O(1)$ for each $i = 1, \ldots, k_m$. Use the construction of z_t to see that

$$
\begin{aligned}
E\Big[\Big(\frac{1}{\sqrt{n}} \sum_{t=1}^{n} z_t m_{i,t}^\theta \Big)^2 \Big] &= E\Big[\Big(\frac{1}{\sqrt{n/b_n}} \sum_{s=1}^{n/b_n} \xi_s \frac{1}{\sqrt{b_n}} \sum_{t=(s-1)b_n+1}^{sb_n} m_{i,t}^\theta \Big)^2 \Big] \\
&= \frac{1}{n/b_n} \sum_{s=1}^{n/b_n} E\Big[\Big(\frac{1}{\sqrt{b_n}} \sum_{t=(s-1)b_n+1}^{sb_n} m_{i,t}^\theta \Big)^2 \Big] = O(1).
\end{aligned}
$$

Thus $\frac{1}{\sqrt{n}} \sum_{t=1}^{n} z_t m_{i,t}^\theta = O_p(1)$. □

Lemma A.2.2. *Under Assumption 4, and* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $||b|| < \infty$,

$$
\sup_{\pi\in\Pi} \sup_{\zeta\in\mathcal{Z}} || \frac{1}{n} \sum_{t=1}^{n} z_t \frac{\partial}{\partial\zeta} m_t(0, \zeta, \pi) || \xrightarrow{p} 0 \tag{A.35}
$$

$$
\sup_{\pi\in\Pi} \sup_{\zeta\in\mathcal{Z}} || \frac{1}{n} \sum_{t=1}^{n} \frac{\partial}{\partial\zeta} m_t(0, \zeta, \pi) - E_{\gamma_n}\big[\frac{\partial}{\partial\theta} m_t(0, \zeta, \pi) \big] || \xrightarrow{p} 0 \tag{A.36}
$$

$$
\frac{1}{\sqrt{n}} \sum_{t=1+h}^{n} z_t m_t^\psi(\pi) = O_{p,\pi}(1). \tag{A.37}
$$

Proof. Consider the first statement. Recall $\xi_s \sim$ iid $N(0,1)$ and is independent of the data, so

$$\frac{1}{n}\sum_{t=1}^{n} z_t \frac{\partial}{\partial \zeta} m_t(0,\zeta,\pi) = \frac{b_n}{n}\sum_{s=1}^{n/b_n} \xi_s \frac{1}{b_n} \sum_{t=(s-1)b_n+1}^{b_n} \frac{\partial}{\partial \zeta} m_t(0,\zeta,\pi),$$

and $\frac{\partial}{\partial \zeta} m_t(0,\zeta,\pi)$ is uniformly integrable on $\mathcal{Z} \times \Pi$ under Assumption 4. Thus, for each $i = 1,\ldots,k_\zeta$ and $j = 1,\ldots,k_m$, stationarity and Minkowski's inequality imply

$$\begin{aligned} E_{\gamma_n}\Big[\Big(\frac{1}{n}\sum_{t=1}^{n} \frac{\partial}{\partial \zeta_i} m_{j,t}(0,\zeta,\pi)\Big)^2\Big] &= E_{\gamma_n}\Big[\Big(\frac{b_n}{n}\sum_{s=1}^{n/b_n} \xi_s \frac{1}{b_n} \sum_{t=(s-1)b_n+1}^{b_n} \frac{\partial}{\partial \zeta_i} m_{j,t}(0,\zeta,\pi)\Big)^2\Big] \\ &= \Big(\frac{b_n}{n}\Big)^2 E_{\gamma_n}\Big[\Big(\frac{1}{b_n} \sum_{t=(s-1)b_n+1}^{b_n} \frac{\partial}{\partial \zeta_i} m_{j,t}(0,\zeta,\pi)\Big)^2\Big] \\ &\leq \Big(\frac{b_n}{n}\Big)^2 E_{\gamma_n}\Big[\Big(\sup_{\pi\in\Pi}\sup_{\zeta\in\mathcal{Z}} \frac{\partial}{\partial \zeta_i} m_{j,t}(0,\zeta,\pi)\Big)^2\Big] \to 0, \end{aligned}$$

so we see that $\frac{1}{n}\sum_{t=1}^{n} z_t \frac{\partial}{\partial \zeta} m_t(0,\zeta,\pi) \xrightarrow{p} 0$ pointwise on $\mathcal{Z} \times \Pi$. Further, $\frac{1}{n}\sum_{t=1}^{n} z_t \frac{\partial}{\partial \zeta} m_t(0,\zeta,\pi)$ is stochastically equicontinuous. Observe that by the mean value theorem

$$\begin{aligned} &\sup_{||(\zeta,\pi)-(\tilde{\zeta},\tilde{\pi})||<\delta} \Big|\frac{1}{n}\sum_{t=1}^{n} z_t\Big(\frac{\partial}{\partial \zeta} m_t(0,\zeta,\pi) - \frac{\partial}{\partial \zeta} m_t(0,\tilde{\zeta},\tilde{\pi})\Big)\Big| \\ &\qquad \leq \sup_{||(\zeta,\pi)-(\tilde{\zeta},\tilde{\pi})||<\delta} \Big|\frac{1}{n}\sum_{t=1}^{n} z_t\Big(\frac{\partial}{\partial \zeta}\frac{\partial}{\partial (\zeta,\pi)'} m_t(0,\zeta^*,\pi^*) \times ||(\zeta,\pi)-(\tilde{\zeta},\tilde{\pi})||\Big)\Big| \\ &\qquad \leq \sup_{\pi\in\Pi}\sup_{\zeta\in\mathcal{Z}} \Big|\frac{1}{n}\sum_{t=1}^{n} z_t\Big(\frac{\partial}{\partial \zeta}\frac{\partial}{\partial (\zeta,\pi)'} m_t(0,\zeta,\pi) \times \delta\Big)\Big|. \end{aligned}$$

Using the construction of z_t, Markov's inequality, and the uniform integrability of $(\partial^2/\partial\zeta\partial(\zeta,\pi)')m_t(0,\zeta,\pi)$ in Assumption 4, it follows that, given $(\epsilon,\eta) > 0$, there is a

$$\delta \in (0, \eta\epsilon/E[\sup_{\pi\in\Pi}\sup_{\zeta\in\mathcal{Z}} |(\partial^2/\partial\zeta\partial(\zeta,\pi)')m_t(0,\zeta,\pi)|])$$

such that

$$\lim_{n\to\infty} P\Big(\sup_{||(\zeta,\pi)-(\tilde{\zeta},\tilde{\pi})||<\delta} \Big|\frac{1}{n}\sum_{t=1}^{n} z_t\Big(\frac{\partial}{\partial \zeta} m_t(0,\zeta,\pi) - \frac{\partial}{\partial \zeta} m_t(0,\tilde{\zeta},\tilde{\pi})\Big)\Big| > \eta\Big)$$

$$\leq \frac{\delta}{\eta} \times \lim_{n\to\infty} E_{\gamma_n}\Big(\sup_{\pi\in\Pi}\sup_{\zeta\in\mathcal{Z}} \Big| \frac{1}{n}\sum_{t=1}^{n} \Big(\frac{\partial}{\partial\zeta}\frac{\partial}{\partial(\zeta,\pi)'} m_t(0,\zeta,\pi) \Big| \Big) < \epsilon.$$

Corollary 3.1 in Newey (1991) then gives the desired result.

Now consider the final statement. Recall that $m_t^{\psi}(\pi) - E_{\gamma_n}[m_t^{\psi}(\pi)]$ is zero mean, stationary, L_p-bounded for some $p > 2$, and L_2-NED with size $1/2$ on an α-mixing base with decay $O(h^{-p/(p-2)-\iota})$. Then by Theorem 17.5 in Davidson (1994) and Theorem 1.6 in McLeish (1975) $E\big[\big(1/\sqrt{n}\times\sum_{t=1}^{n}(m_t^{\psi}(\pi)-E_{\gamma_n}[m_t^{\psi}(\pi)])\big)^2\big] = O(1)$ for each $i = 1,\ldots,k_m$. Use the construction of z_t to see that

$$\begin{aligned}
&E\Big[\Big(\frac{1}{\sqrt{n}}\sum_{t=1}^{n} z_t(m_t^{\psi}(\pi) - E_{\gamma_n}[m_t^{\psi}(\pi)])\Big)^2\Big] \\
&\quad = E\Big[\Big(\frac{1}{\sqrt{n/b_n}}\sum_{s=1}^{n/b_n}\xi_s\frac{1}{\sqrt{b_n}}\sum_{t=(s-1)b_n+1}^{sb_n}(m_t^{\psi}(\pi) - E_{\gamma_n}[m_t^{\psi}(\pi)])\Big)^2\Big] \\
&\quad = \frac{1}{n/b_n}\sum_{s=1}^{n/b_n} E\Big[\Big(\frac{1}{\sqrt{b_n}}\sum_{t=(s-1)b_n+1}^{sb_n}(m_t^{\psi}(\pi) - E_{\gamma_n}[m_t^{\psi}(\pi)])\Big)^2\Big] = O(1).
\end{aligned}$$

Thus $\frac{1}{\sqrt{n}}\sum_{t=1}^{n} z_t(m_t^{\psi}(\pi) - E_{\gamma_n}[m_t^{\psi}(\pi)]) = O_p(1)$. □

A.2.2 Lemmas and Proofs relating to the covariance expansion

In order to conserve space in this appendix, we use the following shorthand notation:

$$\hat{R}_n(h,\theta) = \tfrac{1}{n}\textstyle\sum_{t=1+h}^{h}\varepsilon_t(\theta)\varepsilon_{t-h}(\theta) \quad \hat{R}_n(h) \equiv \hat{R}_n(h,\hat{\theta}_n)$$

$$R_n(h,\theta) = E_{\gamma_n}(\varepsilon_t(\theta)\varepsilon_{t-h}(\theta)) \qquad R_n(h) \equiv R_n(h,\theta_n)$$

$$R_{0,n}(h,\theta) = E_{\gamma_{0,n}}(\varepsilon_t(\theta)\varepsilon_{t-h}(\theta)) \qquad R_{0,n}(h) \equiv R_{0,n}(h,\theta_{0,n})$$

$$R(h,\theta) = E_{\gamma_0}(\varepsilon_t(\theta)\varepsilon_{t-h}(\theta)) \qquad R(h) \equiv R(h,\theta_0)$$

$$\hat{\rho}_n(h) = \tfrac{\hat{R}_n(h)}{\hat{R}_n(0)} \qquad \rho_n(h) = \tfrac{R_n(h)}{R_n(0)}$$

$$\hat{\rho}_n(h;\pi) = \tfrac{\hat{R}_n(h,\hat{\psi}_n(\pi),\pi)}{\hat{R}_n(0)}$$

Observe that for true parameter γ_*, $\varepsilon_t(\theta_*) = \varepsilon_t$ by definition, so

$$R_{0,n}(h) \equiv R_{0,n}(h, \theta_{0,n}) \equiv R_n(h) \equiv R_n(h, \theta_n) \equiv R_0(h) \equiv R_0(h, \theta_0) \equiv E(\varepsilon_t \varepsilon_{t-h}).$$

Lemma A.2.3. *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $||b|| < \infty$ *and Assumption 5,* $\hat{R}_n(h) - R_n(h) = O_p(1/\sqrt{n})$.

The proof follows trivially from Lemma A.2.4.

Lemma A.2.3 establishes convergence in probability to zero of the difference between the denominator in the test statistic, $\hat{R}_n(0) = \frac{1}{n}\sum_{t=1}^n \varepsilon_t^2(\hat{\theta}_n)$, and the second moment of ε_t, $E_{\gamma_n}(\varepsilon_t^2) = E(\varepsilon_t^2) \equiv \sigma^2$.

Proof of Lemma A.2.3. Lemma A.2.4 shows that $\sqrt{n}\Big(\hat{R}_n(h, \hat{\psi}_n(\pi), \pi) - R_n(h)\Big) = O_{p\pi}(1)$, so for any $h = 0, 1, 2, \ldots$ and $\pi = \hat{\pi}_n$, $\hat{R}_n(h, \hat{\psi}_n(\hat{\pi}_n), \hat{\pi}_n) - R_n(h) = O_p(1/\sqrt{n})$. □

Lemma A.2.4. *(a) Recall* $\mathcal{D}_n(h, \pi) = \frac{1}{n}\sum_{t=1+h}^n \big[d_{\psi,t-h}(\psi_{0,n}, \pi)\varepsilon_t(\psi_{0,n}, \pi) + d_{\psi,t}(\psi_{0,n}, \pi)\varepsilon_{t-h}(\psi_{0,n}, \pi)\big]$, *where* $d_{\psi,t}(\psi_{0,n}, \pi) = \frac{\partial}{\partial \psi}\varepsilon_{t-h}(\psi, \pi)\big|_{(\psi,\pi)=(\psi_{0,n},\pi)}$. *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $||b|| < \infty$ *and Assumptions 5 and 9,*

$$\begin{aligned}
&\sqrt{n}\Big(\hat{R}_n(h, \hat{\psi}_n(\pi), \pi) - E(\varepsilon_t \varepsilon_{t-h})\Big) \\
&\quad = \sqrt{n}\Big(\frac{1}{n}\sum_{t=1+h}^n \Big[\varepsilon_t \varepsilon_{t-h} - E(\varepsilon_t \varepsilon_{t-h})\Big]\Big) \\
&\qquad + \Big(H_n^{-1}(\psi_{0,n}, \pi)\frac{1}{\sqrt{n}}\sum_{t=1}^n m_t^{\psi}(\psi_{0,n}, \pi)\Big)' \mathcal{D}_n(h, \pi) \\
&\qquad + \sqrt{n} E_{\gamma_n}\Big[\varepsilon_t(\psi_{0,n}, \pi)\varepsilon_{t-h}(\psi_{0,n}, \pi) - \varepsilon_t \varepsilon_{t-h}\Big] + o_{p\pi}(1)
\end{aligned}$$

(b) Recall $\mathcal{D}_n^{\theta}(h) = \frac{1}{n}\sum_{t=1+h}^n \big[d_{\theta,t-h}(\theta_n)\varepsilon_t(\theta_n) + d_{\theta,t}(\theta_n)\varepsilon_{t-h}(\theta_n)\big]$, *where* $d_{\theta,t}(\theta_n) = \frac{\partial}{\partial \theta}\varepsilon_{t-h}(\theta)\big|_{\theta=\theta_n}$. *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ *and Assumptions 6 and 10,*

$$\begin{aligned}
&\sqrt{n}\Big(\hat{R}_n(h) - E(\varepsilon_t \varepsilon_{t-h})\Big) \\
&\quad = \Big(\frac{1}{\sqrt{n}}\sum_{t=1+h}^n \Big[\varepsilon_t(\theta_n)\varepsilon_{t-h}(\theta_n) - E(\varepsilon_t \varepsilon_{t-h})\Big]\Big)
\end{aligned}$$

$$+\left(J_n^{-1}(\theta_n)\frac{1}{\sqrt{n}}\sum_{t=1}^{n} m_t^{\theta}(\theta_n)\right)' B^{-1}(\beta_n)\mathcal{D}_n^{\theta}(h) + o_p(1)$$

Proof of Lemma A.2.4. (a) Consider an expansion of $\hat{R}_n(h, \hat{\psi}_n(\pi), \pi)$ about $\psi_{0,n}$:

$$\begin{aligned}
&\sqrt{n}\Big(\hat{R}_n(h, \hat{\psi}_n(\pi), \pi) - R_n(h)\Big) \\
&= \sqrt{n}\Big(\hat{R}_n(h, \psi_{0,n}, \pi) - R_n(h)\Big) + \sqrt{n}(\hat{\psi}_n(\pi) - \psi_{0,n})' \frac{\partial}{\partial \psi}\hat{R}_n(h, \psi_{0,n}, \pi) \\
&\qquad + \frac{1}{2}\sqrt{n}(\hat{\psi}_n(\pi) - \psi_{0,n})' \left(\frac{\partial}{\partial \psi}\frac{\partial}{\partial \psi'}\hat{R}_n(h, \tilde{\psi}_n, \pi)\right)(\hat{\psi}_n(\pi) - \psi_{0,n}) \\
&= \sqrt{n}\Big(\hat{R}_n(h, \psi_{0,n}, \pi) - R_n(h)\Big) \\
&\qquad + \left(H_n^{-1}(\psi_{0,n}, \pi)\frac{1}{\sqrt{n}}\sum_{t=1}^{n} m_t^{\psi}(\psi_{0,n}, \pi)\right)' \mathcal{D}_n(h, \pi) + o_{p\pi}(1)
\end{aligned}$$

for some $\tilde{\psi}_n$ st $0 \leq ||\tilde{\psi}_n - \psi_{0,n}|| \leq ||\hat{\psi}_n(\pi) - \psi_{0,n}||$ where the first equality follows from a second order expansion, and the second follows from Assumption 5.ii, stationarity, ergodicity, and the moment bounds in Assumption 9, and Lemma A.2.5. In particular, we show

$$||\frac{\partial}{\partial \psi}\hat{R}_n(h, \psi_{0,n}, \pi) - \mathcal{D}_n(h, \pi)|| = o_{p\pi}(1)$$

and

$$||\frac{\partial}{\partial \psi}\frac{\partial}{\partial \psi'}\hat{R}_n(h, \tilde{\psi}_n, \pi) - \tilde{\mathcal{D}}_n(h, \pi)|| = o_{p\pi}(1)$$

in Lemma A.2.5. From the second statement and Assumption 5, we have that

$$\begin{aligned}
&\sqrt{n}(\hat{\psi}_n(\pi) - \psi_{0,n})' \left(\frac{\partial}{\partial \psi}\frac{\partial}{\partial \psi'}\hat{R}_n(h, \tilde{\psi}_n, \pi)\right)(\hat{\psi}_n(\pi) - \psi_{0,n}) \\
&\qquad = \left(H_n^{-1}(\psi_{0,n}, \pi)\frac{1}{\sqrt{n}}\sum_{t=1}^{n} m_t^{\psi}(\psi_{0,n}, \pi)\right)' \tilde{\mathcal{D}}_n(h, \pi) \times O_{p\pi}(1/\sqrt{n}) + o_{p\pi}(1) \\
&\qquad = o_{p\pi}(1).
\end{aligned}$$

In order to deal with the term $\sqrt{n}\Big(\hat{R}_n(h, \psi_{0,n}, \pi) - R_n(h)\Big)$, add and subtract

$\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}\varepsilon_t\varepsilon_{t-h}$:

$$
\begin{aligned}
&\sqrt{n}\Big(\hat{R}_n(h,\psi_{0,n},\pi)-R_n(h)\Big)\\
&\quad=\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}\Big[\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h}\Big]+\Big[\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}\varepsilon_t\varepsilon_{t-h}-R_n(h)\Big]\\
&\quad=\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}\Big[\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h}\Big]\\
&\qquad+\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}\Big[\varepsilon_t\varepsilon_{t-h}-R_n(h)\Big]-\Big(\frac{1+h}{\sqrt{n}}\Big)R_n(h)\\
&\quad=\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}\Big[\varepsilon_t\varepsilon_{t-h}-R_n(h)\Big]\\
&\qquad+\sqrt{n}\Big(\frac{1}{n}\sum_{t=1+h}^{n}\Big[\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h}\Big]\Big)-\Big(\frac{1+h}{\sqrt{n}}\Big)R_n(h)\\
&\quad=\frac{1}{\sqrt{n}}\sum_{t=1+h}^{n}\Big[\varepsilon_t\varepsilon_{t-h}-R_n(h)\Big]\\
&\qquad+\sqrt{n}\Big(\frac{1}{n}\sum_{t=1+h}^{n}\Big[\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h}\Big]\Big)+O(h/\sqrt{n}).
\end{aligned}
$$

Remark 7. *Recall that $h \leq \mathcal{L}_n = o(n)$. This is sufficient as the argument above only relies on arguments pointwise in h.*

Finally, recall $\sqrt{n}\Big[E_{\gamma_n}\big[\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h}\Big]$ is $O(1/\sqrt{n})$ by Assumption 9. Further, $\varepsilon_t(\psi_{0,n},\pi)$ does not depend on π by Assumption 3. Then by stationarity and ergodicity in Assumption 8, $\frac{1}{n}\sum_{t=1+h}^{n}\Big[\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h}\Big] \xrightarrow{p} E_{\gamma_n}\big[\varepsilon_t(\psi_{0,n},\pi)\varepsilon_{t-h}(\psi_{0,n},\pi)-\varepsilon_t\varepsilon_{t-h}\Big] = O(1/\sqrt{n})$.

(b) Under Assumption 6, $\hat{\pi}_n$ is consistent, so we can expand the sample covariance estimator $\hat{R}_n(h)$ about the true parameter θ_n. Consider an expansion of $\hat{R}_n(h,\hat{\psi}_n(\pi),\pi)$ about $\psi_{0,n}$:

$$
\begin{aligned}
&\sqrt{n}\Big(\hat{R}_n(h,\hat{\theta}_n)-R_n(h)\Big)\\
&=\sqrt{n}\Big(\hat{R}_n(h,\theta_n)-R_n(h)\Big)+\sqrt{n}(\hat{\theta}_n-\theta_n)'\frac{\partial}{\partial\theta}\hat{R}_n(h,\theta_n)
\end{aligned}
$$

$$+\frac{1}{2}\sqrt{n}(\hat{\theta}_n-\theta_n)'\left(\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta'}\hat{R}_n(h,\tilde{\theta}_n)\right)(\hat{\theta}_n-\theta_n)$$
$$=\sqrt{n}\Big(\hat{R}_n(h,\theta_n)-R_n(h)\Big)+\Big(J_n^{-1}(\gamma_0)G_n^{\theta}(\gamma_0)\Big)'B^{-1}(\beta_n)\mathcal{D}_n^{\theta}(h)+o_p(1)$$

for some $\tilde{\theta}_n$ st $0\leq||\tilde{\theta}_n-\theta_n||\leq||\hat{\theta}_n-\theta_n||$.

□

Lemma A.2.5. *Under* $\{\gamma_n\}\in\Gamma(\gamma_0,0,b)$ *with* $||b||<\infty$ *and Assumptions 5 and 9, we have for some* $\psi^*_{0,n}$ *st* $|\psi^*_{0,n}-\psi_{0,n}|\xrightarrow{p}0$

(a) $||\frac{\partial}{\partial\psi}\hat{R}_n(h,\psi_{0,n},\pi)-\mathcal{D}_n(h,\pi)||=o_{p\pi}(1)$

(b) $||\frac{\partial}{\partial\psi}\frac{\partial}{\partial\psi'}\hat{R}_n(h,\psi^*_{0,n},\pi)-\tilde{\mathcal{D}}_n(h,\pi)||=o_{p\pi}(1)$

Proof of Lemma A.2.5. We appeal to a sequence of theorems detailed in Davidson (1994), Chapter 21.[2]

(a) Using differentiability (Assumption 9(ii,iii)) of $\frac{\partial}{\partial\psi}\hat{R}_n(h,\psi,\pi)-\mathcal{D}_n(h,\psi,\pi)$, define

$$B_n=\sup_{\pi\in\Pi}||\frac{\partial}{\partial\pi}\Big(\frac{\partial}{\partial\psi}\hat{R}_n(h,\psi_{0,n},\pi)-\mathcal{D}_n(h,\psi_{0,n},\pi)\Big)||.$$

By Assumption 9(iv), $B_n=O_p(1)$. Further, by an application of the MVT,

$$||\Big(\frac{\partial}{\partial\psi}\hat{R}_n(h,\psi_{0,n},\pi)-\mathcal{D}_n(h,\psi_{0,n},\pi)\Big)-\Big(\frac{\partial}{\partial\psi}\hat{R}_n(h,\psi_{0,n},\pi')-\mathcal{D}_n(h,\psi_{0,n},\pi')\Big)||\leq B_n||\pi-\pi'||\ a.s.$$

Then by following Davidson (1994), Theorem 21.10, we see that $\Big\{\frac{\partial}{\partial\psi}\hat{R}_n(h,\psi_{0,n},\pi)-\mathcal{D}_n(h,\psi_{0,n},\pi)\Big\}$ is stochastically equicontinuous.

Next, observe that

$$||\frac{\partial}{\partial\psi}\hat{R}_n(h,\psi_{0,n},\pi)-\mathcal{D}_n(h,\psi_{0,n},\pi)||\xrightarrow{p}0$$

[2]Davidson (1994)[Theorems 21.6, 21.9, 21.10]. See also Newey (1991).

for every $\pi \in \Pi$ by stationarity and ergodicity (Assumption 9(ii)) and the moment bounds. Combining these two results, we see that Davidson (1994), Theorem 21.9 applies, so

$$\sup_{\pi \in \Pi} || \frac{\partial}{\partial \psi} \hat{R}_n(h, \psi_{0,n}, \pi) - \mathcal{D}_n(h, \pi) || \xrightarrow{p} 0$$

for the non-stochastic function $\mathcal{D}_n(h, \pi) \equiv \mathcal{D}_n(h, \psi_{0,n}, \pi)$.

(b) Define $\tilde{\mathcal{D}}_n(h, \psi, \pi) = E_{\gamma_n}\left[\frac{\partial}{\partial \psi} \frac{\partial}{\partial \psi'} \left(\varepsilon_t(\psi, \pi) \varepsilon_{t-h}(\psi, \pi) \right) \right]$ and $\mathcal{Z}_n(h, \psi, \pi) = \frac{\partial}{\partial \psi} \frac{\partial}{\partial \psi'} \hat{R}_n(h, \psi, \pi) - \tilde{\mathcal{D}}_n(h, \psi_{0,n}, \pi)$.

Using differentiability (Assumption 9(ii,iii)) of $\mathcal{Z}_n(h, \psi, \pi)$, define

$$B_n = \sup_{\pi \in \Pi} \sup_{\psi \in \Psi(\pi)} || \frac{\partial}{\partial \theta} \mathcal{Z}_n(h, \psi, \pi) ||.$$

By Assumption 9(iv), $B_n = O_p(1)$. Further, by an application of the MVT,

$$|| \mathcal{Z}_n(h, \psi, \pi) - \mathcal{Z}_n(h, \psi', \pi') || \leq B_n || \theta - \theta' || \; a.s.$$

Then by following Davidson (1994), Theorem 21.10, we see that $\{\mathcal{Z}_n(h, \psi, \pi)\}$ is stochastically equicontinuous. Next, observe that

$$|| \mathcal{Z}_n(h, \psi, \pi) || \xrightarrow{p} 0$$

by stationarity and ergodicity (Assumption 9(ii)) and the moment bounds. Combining these two results, we see that Davidson (1994), Theorem 21.9 applies, so

$$\sup_{\pi \in \Pi} \sup_{\psi \in \Psi_n(\pi)} || \mathcal{Z}_n(h, \psi, \pi) || \xrightarrow{p} 0.$$

for an open set $\Psi_{0,n}(\pi)$ containing $\psi_{0,n}$ and for the non-stochastic function $\tilde{\mathcal{D}}_n(h, \psi, \pi)$ that is continuous at $\psi_{0,n}$. Finally, we combine the previous result with $|\psi^*_{0,n} - \psi_{0,n}| \xrightarrow{p} 0$ and apply

Davidson (1994), Theorem 21.6 to yield

$$\sup_{\pi \in \Pi} ||\mathcal{Z}_n(h, \psi^*_{0,n}, \pi)|| \xrightarrow{p} 0.$$

□

Lemma A.2.6. *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ *and Assumptions 6 and 10, for some* θ^*_n *st* $|\theta^*_n - \theta_n| \xrightarrow{p} 0$*, we have that*

(a) $||\frac{\partial}{\partial \theta}\hat{R}_n(h, \theta_n) - \mathcal{D}^\theta_n(h)|| = o_p(1)$

(b) $||\frac{\partial}{\partial \theta}\frac{\partial}{\partial \theta'}\hat{R}_n(h, \theta^*_n) - \tilde{\mathcal{D}}^\theta_n(h)|| = o_p(1)$

Proof of Lemma A.2.6. The proof proceeds similarly to that of Lemma A.2.5(b). □

APPENDIX B

APPENDIX FOR TESTING MANY ZERO RESTRICTIONS UNDER MIXED IDENTIFICATION STRENGTH

B.1 Appendix: Notation

The framework utilized here is based upon that developed in Andrews and Cheng (2012a) and Cheng (2015); hence we find it convenient to borrow their notation. This section is meant to be a reference for the notation found throughout the remainder of the paper. Readers familiar with Cheng (2015) may wish to skip this section and return if needed.

The parameter vector $\theta \in \Theta^*$ can be partitioned into three subvectors $\theta = (\beta', \zeta', \pi')'$ where the parameters β and ζ are always strongly identified, and the identification strength of π is determined by β. ζ does not affect the identification of π or β. For the observations $\{W_t = (Y_t', X_t', Z_t')' : t \leq n\}$, $\{Z_t\}$ are the variables associated with parameter ζ which are not associated with β or π. The variables X_t are associated with β and π but not with ζ. For any $\theta \in \Theta^*$, we denote by F_γ the distribution of $\{W_t : t \leq n\}$ and E_γ its expectation, where $\gamma = (\theta, \phi) \in \Gamma$ and $\phi \in \Phi^*$ is a possibly infinite dimensional nuisance parameter such that the distribution is fully characterized by γ. In the framework of Andrews and Cheng (2012a), all elements of π are only allowed to exhibit a single identification strength that is determined by the value of β.

This can be demonstrated with a simple example in which we estimate scalar parameters (β, π) from the nonlinear function $Y_t = \beta g(X_t, \pi) + \varepsilon_t$ for some smooth non-linear function g. It is well known that when $\beta \neq 0$, π can be (strongly) identified, and when $\beta = 0$, π cannot be identified. In order to develop a unifying testing framework, Andrews and Cheng (2012a) utilize a thought experiment which can be characterized with the notion of drifting sequences of true parameters. Let $\beta = \beta_n$ be a sequence of true parameters that are drifting to 0, the point that induces identification failure. Then the strength of identification of π is categorized by the speed at which $\beta_n \to 0$. When $\sqrt{n}\beta_n \to \infty$, π is characterized as being semi-strongly identified, and when $\sqrt{n}\beta_n \to b \in (0, \infty)$, we say π is weakly identified. In the latter case our estimator $\hat{\pi}_n$ is not consistent for the true π_0, and converges instead to a random variable. These drifting sequences are described in greater detail below.

While this setup allows for uniform inference within the parameter space, missing from their framework is the ability to account for identification strengths that differ across elements of π. Cheng (2015) augments this theory specifically for the additive non-linear model to allow for mixed identification strength by pairing subvectors of β with subvectors of π, and allowing the subvectors of β to drift to zero at differing rates. To allow for uniformity over $\gamma \in \Gamma$, all true parameters are indexed by the sample size n. That is, the true $\gamma_n = (\theta_n', \phi_n')'$ where

$$\theta_n = (\beta_n', \zeta_n', \pi_n')'$$

with $\beta_n = (\beta_{1,n}', \ldots, \beta_{p,n}')'$ and $\pi_n = (\pi_{1,n}', \ldots, \pi_{p,n}')'$. These parameters drift to the limiting values $\theta_n \to \theta_0 = (\beta_0', \zeta_0', \pi_0')' \in \Theta^*$ and $\gamma_n \to \gamma_0 \in \Gamma$.

B.1.1 Drifting Sequences

In this framework, the identification strength of π_i, $i = 1, \ldots, p$, is determined by the rate at which $||\beta_{i,n}||$ converges to 0 as $n \to \infty$, with π_i being strongly identified only if $\beta_{i,n} \to \beta_i \neq 0$. In the case that $\beta_{i,0} = 0$, the speed at which $\beta_{i,n} \to \beta_{i,0} = 0$ affects the asymptotic analysis. In particular, when $||\beta_{i,n}|| \to 0$ fast enough, given by case (i) below, we say the parameter $\pi_{i,0}$ is *weakly* identified. In this case, the estimator $\hat{\pi}_{i,n}$ is not consistent. Hence, following Cheng (2015), we divide the space of drifting sequences into three identification categories of π_i:

(i) Weak Identification: $\beta_{i,n} \to 0$ with $n^{1/2}\beta_{i,n} \to b_i \in \mathbb{R}^{d_{\beta_i}}$

(ii) Semi-Strong Identification: $\beta_{i,n} \to 0$ with $n^{1/2}||\beta_{i,n}|| \to \infty$

(iii) Strong Identification: $\beta_{i,n} \to \beta_i \neq 0$.

Observe that the case $\beta_{i,n} = 0 \ \forall n$ is allowed under case (i); hence this case includes non-identification. The category (ii) of semi-strong identification is necessary for uniform results in Cheng's (2015) work. She groups subvectors of π by the identification category above and the rate of convergence to zero for subvectors in the semi-strong identification category. This grouping allows a convenient inductive argument to be used to prove estimation results.

B.1.2 Grouping Notation

To facilitate sequential analysis, we follow the notation in Cheng (2015). Let $||\beta_i||$ denote the norm of vector β_i. We group subvectors of β and their associated pairings in π with the following procedure.

(i) All $||\beta_{j,n}||$ that have non-zero limit are put in the first group. If all $||\beta_{j,n}||$ have zero limits, the first group is empty.

(ii) All $||\beta_{j,n}||$ that are $O(n^{-1/2})$ are put in the last group.

(iii) For those that converge to 0 but at a rate slower than $n^{-1/2}$, members in group k converge to 0 slower than members in group k' for any $k' > k$ and members in the same group converge to 0 at the same rate.

The first group is associated with strong identification, the last group is associated with weak identification, and the middle groups are associated with semi-strong identification, ordered by the rate of convergence. Note that the group index k is a property associated with the drifting sequence $\{\beta_{j,n} : n \geq 1\}$. Therefore the group index k does not change with the sample size n. See Cheng (2015) for details.

Next, suppose there are K groups and $\beta_{k_1}, \ldots, \beta_{k_{p_k}}$ are the elements in group k. Let $l_k = \{k_1, \ldots, k_{p_k}\}$ denote the indices for group k. Use the subscript l_k to denote a sub-vector associated with group k:

$$\beta_{l_k} = (\beta'_{k_1}, \ldots, \beta'_{k_{p_k}})' \in \mathbb{R}^{d_k}$$
$$\text{and} \quad \pi_{l_k} = (\pi'_{k_1}, \ldots, \pi'_{k_{p_k}})' \in \mathbb{R}^{d_{\pi_{l_k}}}.$$

$\beta_{l_k,n}$ denotes the true value of β_{l_k} when the sample size is n and $\beta_{l_k,0}$ denotes its limit. In particular, the grouping rule implies that $||\beta_{l_{k'},n}|| = o(||\beta_{l_k,n}||)$ for $k' > k$ between groups and $||\beta_{j',n}||$ converges at the same rate as $||\beta_{j,n}||$ for any $j, j' \in l_k$ and $k = 1, \ldots, K-1$. In the presence of weak identification, $\beta_{l_k,n} = O(n^{-1/2})$ for $k = K$. If all regressors are in the semi-strong or strong identification category, then we denote $l_K = \emptyset$.

Finally, we describe one more partition of the vectors β and π based on the grouping notation above that will be used to sequentially analyze the limiting behavior of the estimators.

Consider $\pi_{(i),l_k}$, and denote $\pi_{(i),k^-}$ as the elements of π in the previous groups $l_1, \ldots, l_{k-1}$ and $\pi_{(i),k^+}$ as the elements of π in the subsequent groups $l_{k+1}, \ldots, l_K$.

$$\pi_{k^-} = (\pi'_{l_1}, \ldots, \pi'_{l_{k-1}})' \quad \text{and} \quad \pi_{k^+} = (\pi'_{l_{k+1}}, \ldots, \pi'_{l_K})'.$$

Observe that $\pi = (\pi'_{k^-}, \pi'_{l_k}, \pi'_{k^+})'$, and that the identification strength of these subvectors are in decreasing order by definition. The same notation will apply to β, where we can note that the subvectors in $\beta = (\beta'_{k^-}, \beta'_{l_k}, \beta'_{k^+})'$ have smaller magnitude by definition.

It is important to note that π_{l_1} is strongly identified. All strongly identified elements of π are included in this group in order to analyze them together with the strongly identified parameters β and ζ. The semi-strongly identified and weakly-identified elements of π are analyzed using the sequential procedure outlined in Cheng (2015). If no elements of π are strongly identified, $l_1 = \emptyset$ and π_{l_1} disappears.

B.1.3 Concentrated Criterion Functions

The least squares estimator $\hat{\theta}_n$ minimizes $Q_n(\theta)$ over $\theta \in \Theta$, where $\Theta = \mathcal{B} \times \mathcal{Z} \times \Pi$. $\mathcal{B} = \times_{j=1}^{p} \mathcal{B}_j$ where $\mathcal{B}_j$ for $j = 1, \ldots, p$ are compact sets, as are $\mathcal{Z}$ and Π. We assume all true values and parsimonious model counterparts in Θ^* are in the interior of the optimization space Θ.

Proof of the consistency of the strongly and semi-strongly identified components of the estimator follows from sequential analysis in order of decreasing identification strength. In particular, we sequentially concentrate out parameters and analyze the concentrated criterion function

$$Q_n^c(\pi_{l_k}, \pi_{k^+}) = Q_n(\hat{\psi}_{k^-}(\pi_{l_k}, \pi_{k^+}), \pi_{l_k}, \pi_{k^+})$$

where $\psi_{k^-} = (\beta', \zeta', \pi'_{k^-})'$ collects the parameters that have been concentrated out, and the true values of these parameters are denoted with the additional subscripts $\psi_{k^-,n} = (\beta'_n, \zeta'_n, \pi'_{k^-,n})'$ and $\psi_{k^-,0} = (\beta'_0, \zeta'_0, \pi'_{k^-,0})'$ where the latter gives the limit of the drifting sequence: $\psi_{k^-,n} \to \psi_{k^-,0}$.

Due to the mixed identification strength along differing subvectors of π, it becomes necessary to evaluate expansions around the points of sequential identification failure, $\beta_{l_k}^0 = 0$ and $\beta_{k^+}^0 = 0$, rather than the true values $\beta_{l_k,n} = 0$ and $\beta_{k^+,n} = 0$ as is commonly done Andrews and Cheng's (2012a). We use the superscript 0 notation to define

$$\psi_{k^-,n}^0 = (\beta_{k^-,n}', \beta_{l_k}^{0\prime}, \beta_{k^+}^{0\prime}, \zeta_n', \pi_{k^-,n}')'$$

to be the parameter vector consisting of the concentrated out parameters evaluated at the point of sequential identification failure $\beta_{l_k}^0 = 0$ and $\beta_{k^+}^0 = 0$. Observe that the difference $\psi_{k^-,n} - \psi_{k^-,n}^0 = (0', \beta_{l_k,n}, \beta_{k^+,n}, 0', 0')'$. This is done so that under our basic assumptions the centering term $Q_n(\psi_{k^-,n}^0, \pi_{l_k}, \pi_{k^+})$ does not depend on $(\pi_{l_k}', \pi_{k^+}')'$.

B.2 Appendix: Limit Theory for Models with Mixed Identification Strength

We assume that the following assumptions hold throughout this section.

Assumption A.1. *The observations* $\{W_t = (Y_t', X_t', Z_t')' : t \leq n\}$ *are strictly stationary as are* $\{\varepsilon_t\}$. $\{W_t\}$ *is strongly mixing with mixing coefficient* $\alpha(j)$ *such that* $\sum_{j=1}^{\infty} \alpha(j)^{\delta/(2+\delta)} < \infty$ *for some* $\delta > 0$.

Assumption A.2. *The true value* θ^* *belongs to the set* $\Theta^* = \mathcal{B}_1^* \times \cdots \times \mathcal{B}_p^* \times \mathcal{Z}^* \times \Pi^*$ *where* $\mathcal{B}_j^*$ *is compact and includes* 0 *for each* j. Π^* *and* $\mathcal{Z}^*$ *are compact. For any* $\theta \in \Theta^*$, *the distribution of* $\{W_t\}$ *is given by* F_γ, *where* $\gamma = (\theta', \phi')' \in \Gamma$, *and* $\phi \in \Phi^*$ *is an possibly infinite dimensional nuisance parameter that fully characterizes the distribution.* Φ^* *is a compact metric space with a metric that induces weak convergence on bivariate distributions* (W_t, W_{t+m}) *for every* $t, m \geq 1$.

Assumption A.3. *The estimator* $\hat{\theta}_n$ *minimizes the criterion function* $Q_n(\theta) \equiv Q_n(\theta; W_t) = \frac{1}{n}\sum_{t=1}^n m_t(\theta; W_t)$ *over* $\theta \in \Theta = \mathcal{B}_1 \times \cdots \times \mathcal{B}_p \times \mathcal{Z} \times \Pi$ *where* $\mathcal{B}_j, \mathcal{Z}, \Pi$ *are compact for every* j *and* Θ^* *is contained in the interior of* Θ.

Assumption A.4. *For every* $\mathcal{B}_j$ *there is a* $\Pi_j = \otimes_{i=1}^{q_j} \Pi_i$ *such that* $m_t(\theta; w)$ *does not depend upon* $\pi_j \in \Pi_j$ *iff* $\beta_j = 0$. β_i *for* $i \neq j$ *does not affect the identification of* π_j. ζ *does not affect the identification of* β *or* π, *and the identification of* ζ *is not affected by* β *or* π.

Assumption A.5. *Denote E_{γ_0} as expectation taken under true parameter γ_0.*

1. *if $l_K = \emptyset$, then $E_{\gamma_0}(m_t(\theta; W_t))$ is minimized uniquely by $\theta = \theta^* \in \Theta^*$.*

2. *if $l_K \neq \emptyset$, then $E_{\gamma_0}(m_t(\psi_{K^-}, \pi_K; W_t))$ is minimized uniquely by $\psi_{K^-} = \psi^*_{K^-} \in \Psi^*_{K^-}$ for every $\pi_K \in \Pi_K$.*

Assumption A.6. *The function $m_t(\theta; \cdot)$ is measurable with respect to $\sigma(W_t)$, the sigma field generated by $\{W_t\}$, for every $\theta \in \Theta$. Further, $m_t(\theta)$ is three times continuously differentiable, and for some $\delta > 0$*

1. $\sup_{\theta \in \Theta} E_{\gamma_0} |m_t(\theta)|^{2+\delta} < \infty$

2. $\sup_{\theta \in \Theta} \lim_{n \to \infty} E_{\gamma_n} \left| \left[B(\beta_{K^-})^{-1} \nabla_{\psi_{K^-}} m_t(\theta) \right]_j \right|^{2+\delta} < \infty$

3. $\sup_{\theta \in \Theta} \lim_{n \to \infty} E_{\gamma_n} \left| \left[B(\beta_{K^-})^{-1} \left(\nabla^2_{\psi_{K^-}} m_t(\theta) \right) B(\beta_{K^-})^{-1} \right]_{i,j} \right|^{2+\delta} < \infty$

4. $\sup_{\theta \in \Theta} \lim_{n \to \infty} E_{\gamma_n} \left| \left[\frac{\partial}{\partial \psi'_{k^-}} vec\left(B(\beta_{K^-})^{-1} \nabla^2_{\psi_{K^-}} m_t(\theta) B(\beta_{K^-})^{-1} \right) \right]_{i,j} \right|^{2+\delta} < \infty$

where $[A]_{i,j}$ denotes the i, jth element of the matrix A.

Additional Assumptions:

Assumption A.7. *i) For every $k = 1, \ldots, K$,*

$$\mathcal{K}_k(\psi_{k^-}, \pi_{l_k}, \pi_{k^+}; \gamma_0) = \frac{\partial}{\partial \beta'_0} E_{\gamma_0} \nabla_{\psi_{k^-}} m_t(\theta)$$

exists for every $(\theta, \gamma_0) \in \Theta_\eta \times \Gamma_0$, where $\theta = (\psi_{k^-}, \pi_{l_k}, \pi_{k^+})$.

ii) For every $k = 1, \ldots, K$, $\mathcal{K}_k(\theta; \gamma)$ is continuous at $(\psi^0_{k^-}, \pi_{l_k}, \pi_{k^+}; \gamma^0)$ uniformly over $\pi_{l_k}, \pi_{k^+} \in \Pi_{l_k} \times \Pi_{k^+}$ for every $\gamma^0 \in \Gamma$ such that $\psi^0_{k^-}$ is a subvector of γ^0.

Assumption A.8. *For each k, $\lambda_{min}(H_k(\pi_{l_k}, \pi_{k^+}; \gamma_0)) \geq \varepsilon$ for some $\varepsilon > 0$.*

Assumption A.9. *i) If l_K is empty, then $\lambda_{min}(\Omega_\theta(\gamma_0)) \geq \varepsilon$ for some $\varepsilon > 0$ and every i.*

ii) If l_K is not empty, then each sample path of the process $\chi(\pi_{l_K})$ is continuous a.s. and minimized uniquely with probability 1. Denote the minimizer by $\pi^*_{l_K}$.

i) For every $k = 1, \ldots, K$, $\mathcal{K}_k(\psi_{k^-}, \pi_{l_k}, \pi_{k^+}; \gamma_0)$ exists for every $(\theta, \gamma_0) \in \Theta_\eta \times \Gamma_0$, where $\theta = (\psi_{k^-}, \pi_{l_k}, \pi_{k^+})$.

ii) For each $k = 1, \ldots, K$, $\mathcal{K}_k(\theta; \gamma)$ is continuous at $(\psi^0_{k^-}, \pi_{l_k}, \pi_{k^+}; \gamma^0)$ uniformly over $\pi_{l_k}, \pi_{k^+} \in \Pi_{l_k} \times \Pi_{k^+}$ for every $\gamma^0 \in \Gamma$ such that $\psi^0_{k^-}$ is a subvector of γ^0.

iii) $\lambda_{min}(H_k(\pi_{l_k}, \pi^+_k); \gamma_0) \geq \varepsilon$ for some $\varepsilon > 0$.

iv) Let $\mathcal{G}(\pi_{l_K}; \gamma_0)$ be a zero mean Gaussian process with covariance kernel $\Omega(\pi_{l_K}, \tilde{\pi}_{l_K}; \gamma_0)$. Then $\lambda_{min}(\Omega(\pi_{l_K}, \tilde{\pi}_{l_K}; \gamma_0)) \geq \varepsilon$ for some $\varepsilon > 0$.

v) Define the process

$$\chi(\pi_{l_K}) = -\frac{1}{2}\Big(\mathcal{K}_K(\pi_{l_K}; \gamma_0) b_{l_K} + \mathcal{G}(\pi_{l_K}; \gamma_0)\Big)' \Big[H_K(\pi_{l_K}; \gamma_0)\Big]^{-1} \Big(\mathcal{K}_K(\pi_{l_K}; \gamma_0) b_{l_K} + \mathcal{G}(\pi_{l_K}; \gamma_0)\Big).$$

Each sample path of the process $\chi(\pi_{l_K})$ is minimized uniquely with probability 1.

Lemma B.2.1 (Consistency for Strong Identification Groups)**.** *Suppose Assumptions A.1-A.6 hold. Then under $\gamma_n \to \gamma_0$,*

$$\sup_{\pi^+_1 \in \Pi^+_1} ||\hat{\zeta}(\pi^+_1) - \zeta_n|| \xrightarrow{p} 0$$

$$\sup_{\pi^+_1 \in \Pi^+_1} ||\hat{\beta}(\pi^+_1) - \beta_n|| \xrightarrow{p} 0$$

$$\sup_{\pi^+_1 \in \Pi^+_1} ||\hat{\pi}_{l_1}(\pi^+_1) - \pi_{l_1,n}|| \xrightarrow{p} 0$$

Proof of Lemma B.2.1. First, observe that a ULLN holds for $Q_n(\theta)$, since $\sup_\theta |Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0$ by B.3.1. Next, denote the true sequence $\psi_n \to \psi_0$ and $Q(\theta) = Q(\psi, \pi_{l_1} | \pi^+_1)$ for fixed π^+_1. By assumption, $Q(\theta)$ is uniquely minimized by $(\psi_0', \pi_{l_1,0}')'$ for any fixed π^+_1. Observe that since

$\beta_{l_k,0} = 0$ for every $k > 1$, $Q(\psi_0, \pi_{l_1,0}|\pi_1^+)$ does not depend upon π_1^+. Finally, we appeal to Lemma 3.1 in Andrews and Cheng (2012a) for the extension to uniform consistency. □

Lemma B.2.2 (Consistency for Semi-Strong Identification Groups)**.** *Suppose Assumptions A.1-A.9 hold. Then under $\gamma_n \to \gamma_0$, for $k = 2, \ldots, K-1$,*

(a) the concentrated sample criterion function satisfies

$$
\begin{aligned}
&||\beta_{l_k,n}||^{-2}\Big(Q_n^c(\pi_{l_k}, \pi_{k^+}) - Q_n(\psi^0_{k^-,n})\Big) \\
&\xrightarrow{p} -\frac{1}{2}(\omega'_{k,0}, 0'_{d_{k^+}})\mathcal{K}_k(\pi_{l_k}, \pi_{k^+}; \gamma_0)'[H_k(\pi_{l_k}, \pi_{k^+}; \gamma_0)]^{-1}\mathcal{K}_k(\pi_{l_k}, \pi_{k^+}; \gamma_0)(\omega'_{k,0}, 0'_{d_{k^+}})',
\end{aligned}
\tag{B.1}
$$

where $\omega_{k,0} = \lim_{n\to\infty} \beta_{l_k,n}/||\beta_{l_k,n}||$ is the angle parameter

(b) the estimator of $\pi_{l_k,n}$ satisfies

$$
\sup_{\pi_{k^+} \in \Pi_{k^+}} ||\hat{\pi}_{l_k}(\pi_{k^+}) - \pi_{l_k,n}|| \xrightarrow{p} 0
$$

(c) the estimator of $\psi_{k^-} = (\beta'_{(i)}, \zeta'_{(i)}, \pi'_{l_1}, \ldots, \pi'_{l_{k-1}})'$ satisfies

$$
||\beta_{l_k,n}||^{-1}\begin{pmatrix}
\hat{\beta}_{k^-}(\pi_{k^+}) - \beta_{k^-,n} \\
\hat{\beta}_{l_k}(\pi_{k^+}) - \beta_{l_k,n} \\
\hat{\beta}_{k^+}(\pi_{k^+}) \\
\hat{\zeta}_{(i)} - \zeta_n \\
B^*(\beta_{k^-,n})(\hat{\pi}_{k^-}(\pi_{k^+}) - \pi_{k^-,n})
\end{pmatrix} \xrightarrow{p} 0,
$$

uniformly over $\pi_{k^+} \in \Pi_{k^+}$ where $B^*(\beta_{k^-,n}) = diag\{(1_{d_{\pi_{l_1}}}||\beta_{l_1}||, \ldots, 1_{d_{\pi_{l_{k-1}}}}||\beta_{l_{k-1}}||)'\}$.

Proof of Lemma B.2.2. The proof follows by an inductive argument.

1. Observe (b) and (c) hold for $k = 1$ by Lemma B.2.1.

2. Let Lemma B.2.2 hold for $k-1$. We will show it holds for k.

a) i) We will first use a second order expansion of $Q_n(\hat{\psi}_{k^-}(\pi_{l_k}, \pi_{k^+}), \pi_{l_k}, \pi_{k^+})$ around $Q_n(\psi^0_{k^-,n})$. This will imply the LHS is minimized by $\hat{\pi}_{l_k}(\pi_{k^+})$, since $Q_n(\psi^0_{k^-,n})$ does not depend upon π_{l_k} or π_{k^+}.

ii) We will then appeal to ULLNs developed in the appendix to show convergence of the components of the expansion.

b) i) Proof of part (b) follows from a simple observation and the argmax continuity theorem. Observe that the left hand side of (a) is minimized by $\hat{\pi}_{l_k}(\pi_{k^+})$

ii) The right hand side of (a) can be shown to be minimized at $\pi_{l_k} = \pi_{l_k,0}$ by a matrix Cauchy-Schwarz inequality.

iii) Finally, invoke the argmax continuity theorem to arrive at the result.

c) Part (c) follows from two mean value expansions of the first order condition and score function paired with the refined rate derived in part (a).

Step 2 utilizes the following expansions of the criterion function and its derivative about the point of sequential identification failure.

A second order mean value expansion of the criterion function about the point of sequential identification failure yields for some $\tilde{\tilde{\psi}}_{k^-,n}$ between $\hat{\psi}_{k^-}(\pi_{l_k}, \pi_{k^+})$ and $\psi^0_{k^-,n}$,

$$
\begin{aligned}
&Q_n(\hat{\psi}_{k^-}(\pi_{l_k}, \pi_{k^+}), \pi_{l_k}, \pi_{k^+}) - Q_n(\psi^0_{k^-,n}) \qquad \text{(B.2)}\\
&\quad = \nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})'(\hat{\psi}_{k^-}(\pi_{l_k}, \pi_{k^+}) - \psi^0_{k^-,n})\\
&\qquad + \frac{1}{2}(\hat{\psi}_{k^-}(\pi_{l_k}, \pi_{k^+}) - \psi^0_{k^-,n})'\nabla^2_{\psi_{k^-}} Q_n(\tilde{\tilde{\psi}}_{k^-,n}, \pi_{l_k}, \pi_{k^+})(\hat{\psi}_{k^-}(\pi_{l_k}, \pi_{k^+}) - \psi^0_{k^-,n})
\end{aligned}
$$

Consider the first order condition from the optimization problem, and use the MVT to see that for some $\tilde{\psi}_{k^-,n}$ between $\hat{\psi}_{k^-}(\pi_{l_k}, \pi_{k^+})$ and $\psi^0_{k^-,n}$,

$$
\begin{aligned}
0 &= \nabla_{\psi_{k^-}} Q_n(\hat{\psi}_{k^-}(\pi_{l_k}, \pi_{k^+}), \pi_{l_k}, \pi_{k^+})\\
&= \nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+}) + \nabla^2_{\psi_{k^-}} Q_n(\tilde{\psi}_{k^-,n}, \pi_{l_k}, \pi_{k^+})(\hat{\psi}_{k^-}(\pi_{l_k}, \pi_{k^+}) - \psi^0_{k^-,n})
\end{aligned}
$$

which implies

$$\hat{\psi}_{k^-}(\pi_{l_k}, \pi_{k^+}) - \psi^0_{k^-,n} = -[\nabla^2_{\psi_{k^-}} Q_n(\tilde{\psi}_{k^-,n}, \pi_{l_k}, \pi_{k^+})]^{-1} \nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+}) \tag{B.3}$$

Finally, expansion about the point of sequential identification failure in expansion (ii) induces a bias, so we use the following mean value expansion to account for this bias.

$$\begin{aligned}
&\nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+}) \\
&\quad = E_{\gamma_n}\left[\nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})\right] \\
&\qquad + \nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+}) - E_{\gamma_n}\left[\nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})\right] \\
&\quad = E_{\gamma^0_{k^-,n}}\left[\nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})\right] \\
&\qquad + \frac{\partial}{\partial(\beta'_{l_k,n}, \beta'_{k^+})} E_{\tilde{\gamma}_n}\left[\nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})\right](\beta'_{l_k,n}, \beta'_{k^+})' \\
&\qquad + \nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+}) - E_{\gamma_n}\left[\nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})\right] \\
&\quad = \mathcal{K}_{k,n}(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+}; \tilde{\gamma}_n) \cdot (\beta'_{l_k,n}, \beta'_{k^+})' + \mathcal{G}_{k,n}(\pi_{l_k}, \pi_{k^+})
\end{aligned} \tag{B.4}$$

where $\mathcal{K}_{k,n}(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+}; \tilde{\gamma}_n) = \frac{\partial}{\partial(\beta'_{l_k,n}, \beta'_{k^+})} E_{\tilde{\gamma}_n}\left[\nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})\right]$ for some $\tilde{\gamma}_n$ between γ_n and $\gamma^0_{k^-,n}$, and $\mathcal{G}_{k,n}(\pi_{l_k}, \pi_{k^+}) = \nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+}) - E_{\gamma_n}\left[\nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})\right]$. Note that $E_{\gamma^0_{k^-,n}}\left[\nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})\right] = 0$ by definition (see Lemmas 9.1, 9.2 in Andrews and Cheng (2012b)).

Observe that by differentiability of Q_n, the definition of the estimator $\hat{\psi}_{k^-}(\pi_{l_k}, \pi_{k^+})$, the MVT, and combining the above three expansions B.2, B.3, and B.4 we have

$$\begin{aligned}
&Q_n(\hat{\psi}_{k^-}(\pi_{l_k}, \pi_{k^+}), \pi_{l_k}, \pi_{k^+}) - Q_n(\psi^0_{k^-,n}) \\
&= -\nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})'[\nabla^2_{\psi_{k^-}} Q_n(\tilde{\psi}_{k^-,n}, \pi_{l_k}, \pi_{k^+})]^{-1} \nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+}) \\
&\quad + \frac{1}{2} \nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})'[\nabla^2_{\psi_{k^-}} Q_n(\tilde{\psi}_{k^-,n}, \pi_{l_k}, \pi_{k^+})]^{-1} \nabla^2_{\psi_{k^-}} Q_n(\tilde{\tilde{\psi}}_{k^-,n}, \pi_{l_k}, \pi_{k^+}) \\
&\quad \times [\nabla^2_{\psi_{k^-}} Q_n(\tilde{\psi}_{k^-,n}, \pi_{l_k}, \pi_{k^+})]^{-1} \nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})
\end{aligned} \tag{B.5}$$

for some $\tilde{\psi}_{k^-,n}$ and $\tilde{\tilde{\psi}}_{k^-,n}$, both between $\hat{\psi}_{k^-}(\pi_{l_k}, \pi_{k^+})$ and $\psi^0_{k^-,n}$, and where

$$\begin{aligned}
\nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+}) &= \mathcal{K}_{k,n}(\pi_{l_k}, \pi_{k^+}; \tilde{\gamma}_n)(\beta'_{l_k,n}, \beta'_{k^+})' + \mathcal{G}_{k,n}(\pi_{l_k}, \pi_{k^+}) \\
\mathcal{K}_{k,n}(\pi_{l_k}, \pi_{k^+}; \tilde{\gamma}_n) &\equiv \mathcal{K}_{k,n}(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+}; \tilde{\gamma}_n) \\
&= \frac{\partial}{\partial(\beta'_{l_k,n}, \beta'_{k^+})} E_{\tilde{\gamma}_n}\left[\nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})\right] \\
&= \frac{1}{n}\sum_{t=1}^{n} \frac{\partial}{\partial(\beta'_{l_k,n}, \beta'_{k^+})} E_{\tilde{\gamma}_n}\left[\nabla_{\psi_{k^-}} m_t(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})\right] \\
\mathcal{G}_{k,n}(\pi_{l_k}, \pi_{k^+}) &= \nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+}) - E_{\gamma_n}\left[\nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})\right] \\
&= \frac{1}{n}\sum_{t=1}^{n} \Big\{\nabla_{\psi_{k^-}} m_t(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+}) \\
&\qquad - E_{\gamma_n}\left[\nabla_{\psi_{k^-}} m_t(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})\right]\Big\}
\end{aligned}$$

and

$$\nabla^2_{\psi_{k^-}} Q_n(\tilde{\psi}_{k^-,n}, \pi_{l_k}, \pi_{k^+}) = \frac{1}{n}\sum_{t=1}^{n} \nabla^2_{\psi_{k^-}} m_t(\tilde{\psi}_{k^-,n}, \pi_{l_k}, \pi_{k^+})$$

Hence, we need to establish ULLNs for $\mathcal{K}_{k,n}(\pi_{l_k}, \pi_{k^+}; \tilde{\gamma}_n)$, $\mathcal{G}_{k,n}(\pi_{l_k}, \pi_{k^+})$, and $\nabla^2_{\psi_{k^-}} Q_n(\tilde{\psi}_{k^-,n}, \pi_{l_k}, \pi_{k^+})$, which are established in the supporting lemmas section. In particular recall that $\psi_{k^-} = (\beta', \zeta', \pi'_{k^-})'$, $\pi_{k^-} = (\pi'_{l_1}, \ldots, \pi'_{l_{k-1}})'$, and that π_{l_k} for $k > 1$ does not affect Q in the limit, resulting in a hessian $\nabla^2_{\psi_{k^-}} Q_n$ that approaches singularity as $n \to \infty$. It is necessary then to normalize the columns of the hessian corresponding π_{l_k} for $k > 1$. Define $B(\beta_{k^-}) = diag\{(1_{d_\beta + d_\zeta}, 1_{d_{\pi_{l_1}}}||\beta_{l_1}||, \ldots, 1_{d_{\pi_{l_{k-1}}}}||\beta_{l_{k-1}}||)'\}$

Lemmas B.3.2, B.3.3, and B.3.4 show that for each $k = 1, \ldots, K-1$

$$\sup_{\pi_{l_k}, \pi_{k^+}} ||B(\beta_{k^-,n})^{-1} \nabla^2_{\psi_{k^-}} Q_n(\tilde{\psi}_{k^-,n}, \pi_{l_k}, \pi_{k^+}) B(\beta_{k^-,n})^{-1} - H_k(\pi_{l_k}, \pi_{k^+}; \gamma_0)|| \xrightarrow{p} 0$$

$$\sup_{\pi_{l_k}, \pi_{k^+}} ||B(\beta_{k^-,n})^{-1} \mathcal{K}_{k,n}(\pi_{l_k}, \pi_{k^+}; \tilde{\gamma}_n) - \mathcal{K}_k(\pi_{l_k}, \pi_{k^+}; \gamma_0)|| \to 0$$

$$\sup_{\pi_{l_k}, \pi_{k^+}} ||||\beta_{l_k,n}||^{-1} B(\beta_{k^-,n})^{-1} \mathcal{G}_{k,n}(\pi_{l_k}, \pi_{k^+})|| \xrightarrow{p} 0$$

Recall that $\omega_{k,0} = \lim_{n\to\infty} \beta_{l_k,n}/||\beta_{l_k,n}||$ for $k < K$, and $\beta_{l_j,n} = o(||\beta_{l_k,n}||)$ for every $j > k$. Normalize B.5 by $||\beta_{l_k,n}||^{-2}$, multiplying by $I_{d_{\psi_{k^-}}} = B(\beta_{k^-,n})^{-1}B(\beta_{k^-,n})$, and utilize the results from Lemmas B.3.2, B.3.3, and B.3.4 in equation B.5 to establish the result B.1:

$$
\begin{aligned}
&||\beta_{l_k,n}||^{-2}\Big(Q_n^c(\pi_{l_k},\pi_{k^+}) - Q_n(\psi^0_{k^-,n})\Big) \\
&\xrightarrow{p} -\frac{1}{2}(\omega'_{k,0}, 0'_{d_{k^+}})\mathcal{K}_k(\pi_{l_k},\pi_{k^+};\gamma_0)'[H_k(\pi_{l_k},\pi_{k^+};\gamma_0)]^{-1}\mathcal{K}_k(\pi_{l_k},\pi_{k^+};\gamma_0)(\omega'_{k,0}, 0'_{d_{k^+}})',
\end{aligned}
$$

Notice that the notation differs from that used in Cheng (2015). In particular, for the specific additive nonlinear model studied in Cheng (2015), there is some $\tilde{\Omega}(\pi_{l_k,1},\pi_{l_k,2}|\pi_{k^+})$ such that $H_k(\pi_{l_k},\pi_{k^+};\gamma_0) = \tilde{\Omega}(\pi_{l_k},\pi_{l_k}|\pi_{k^+})$ and $\tilde{\mathcal{K}}_k(\pi_{l_k},\pi_{k^+};\gamma_0) = \tilde{\Omega}(\pi_{l_k},\pi_{l_k,0}|\pi_{k^+})$, where $\tilde{\mathcal{K}}_k(\pi_{l_k},\pi_{k^+};\gamma_0) = \frac{\partial}{\partial\beta'_0}E_{\gamma_0}\big[\nabla_{\psi_{k^-}} m_t(\psi^0_{k^-,n},\pi_{l_k},\pi_{k^+})\big]$, so that our $\mathcal{K}_k(\pi_{l_k},\pi_{k^+};\gamma_0) = \tilde{\mathcal{K}}_k(\pi_{l_k},\pi_{k^+};\gamma_0)S_k$ where S_k is a selection matrix that selects the columns corresponding to $(\beta'_{l_k,0},\beta'_{k^+,0})'$. Our results generalize those in Cheng (2015) to a broader class of models.

For part (b), observe that the left hand side of (a) B.1 is minimized by $\hat{\pi}_{l_k}(\pi_{k^+})$. That the right hand side of (a) B.1 is minimized at $\pi_{l_k} = \pi_{l_k,0}$ can be shown by a matrix Cauchy-Schwarz inequality (Tripathi, 1999). To establish the result, one must then invoke the argmax continuous mapping theorem (van der Vaart and Wellner, 1996).

Finally, for part (c), consider the expansions B.3 and B.4 which are related to the first order condition. Given the result in part (b), the expansion about $\beta^0_{l_k} = 0$, rather than the true value $\beta_{l_k,n}$, is not necessary, so replace $\beta^0_{l_k} = 0$ by the true value $\beta_{l_k,n}$ in the expansion B.3 to yield

$$
\hat{\psi}_{k^-}(\pi_{l_k},\pi_{k^+}) - \psi^0_{k,n} = -[\nabla^2_{\psi_{k^-}} Q_n(\tilde{\psi}_{k^-,n},\pi_{l_k},\pi_{k^+})]^{-1}\nabla_{\psi_{k^-}} Q_n(\psi^0_{k,n},\pi_{l_k},\pi_{k^+})
$$

where $\psi^0_{k,n} = (\beta'_{k^-,n},\beta'_{l_k,n},\beta^{0\prime}_{k^+},\zeta'_n,\pi'_{k^-,n})'$. This additionally alters B.4 to

$$
\begin{aligned}
\nabla_{\psi_{k^-}} Q_n(\psi^0_{k,n},\pi_{l_k},\pi_{k^+}) &= E_{\gamma_n}\big[\nabla_{\psi_{k^-}} Q_n(\psi^0_{k,n},\pi_{l_k},\pi_{k^+})\big] \\
&\quad + \nabla_{\psi_{k^-}} Q_n(\psi^0_{k,n},\pi_{l_k},\pi_{k^+}) - E_{\gamma_n}\big[\nabla_{\psi_{k^-}} Q_n(\psi^0_{k,n},\pi_{l_k},\pi_{k^+})\big] \\
&= E_{\gamma^0_{k,n}}\big[\nabla_{\psi_{k^-}} Q_n(\psi^0_{k,n},\pi_{l_k},\pi_{k^+})\big]
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{\partial}{\partial \beta_{k^+,n}} E_{\tilde{\gamma}_n}\left[\nabla_{\psi_{k^-}} Q_n(\psi^0_{k,n}, \pi_{l_k}, \pi_{k^+})\right]\beta_{k^+,n} \\
&+ \nabla_{\psi_{k^-}} Q_n(\psi^0_{k,n}, \pi_{l_k}, \pi_{k^+}) - E_{\gamma_n}\left[\nabla_{\psi_{k^-}} Q_n(\psi^0_{k,n}, \pi_{l_k}, \pi_{k^+})\right] \\
&= \mathcal{K}_{k,n}(\psi^0_{k,n}, \pi_{l_k}, \pi_{k^+}; \tilde{\gamma}_n)\beta_{k^+,n} + \mathcal{G}_{k^+,n}(\pi_{l_k}, \pi_{k^+})
\end{aligned}
$$

Recall that $B(\beta_{k^-,n})^{-1}\mathcal{K}_{k,n}(\psi^0_{k,n}, \pi_{l_k}, \pi_{k^+}; \tilde{\gamma}_n)$ has a non-zero, finite limit, but that $\beta_{l_j,n} = o(||\beta_{l_k,n}||)$ for every $j > k$. Substitute the previous equation and normalize by $||\beta_{l_k}||^{-1}B(\beta_{k^-,n})$ to see that

$$
\begin{aligned}
&||\beta_{l_k}||^{-1}B(\beta_{k^-,n})\left(\hat{\psi}_{k^-}(\pi_{l_k}, \pi_{k^+}) - \psi^0_{k,n}\right) \\
&= -[B(\beta_{k^-,n})^{-1}\nabla^2_{\psi_{k^-}} Q_n(\tilde{\psi}_{k^-,n}, \pi_{l_k}, \pi_{k^+})B(\beta_{k^-,n})^{-1}]^{-1} \\
&\quad \times ||\beta_{l_k}||^{-1}B(\beta_{k^-,n})^{-1}\left(\mathcal{K}_{k,n}(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+}; \tilde{\gamma}_n)\beta_{k^+,n} + \mathcal{G}_{k^+,n}(\pi_{l_k}, \pi_{k^+})\right).
\end{aligned} \tag{B.6}
$$

Paired with the result from part (b), recall that the first quantity on the right hand side has a non-zero limit uniformly in probability, but the second quantity converges uniformly in probability to zero. This establishes the result in part (c).

□

Theorem B.2.3. *Let Assumptions A.1 - A.9 hold. Under $\gamma_n \to \gamma_0$,*

a) If $l_K \neq \emptyset$, where l_K indexes the weakly identified subvector of π, then

i)

$$
n\big(Q^c_n(\pi_{l_k}) - Q_n(\psi^0_{K,n}, \pi_{l_k})\big) \Rightarrow \chi(\pi_{l_K}) \tag{B.7}
$$

ii)

$$
\begin{pmatrix} n^{1/2}B(\beta_{K^-,n})\big(\hat{\psi}_{K^-} - \psi_{K^-,n}\big) \\ \hat{\pi}_{l_K} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \tau(\pi^*_{l_K}) - S_{l_K}b_{l_K} \\ \pi^*_{l_K} \end{pmatrix} \tag{B.8}
$$

where S_{l_K} is the selection matrix that selects the columns corresponding to β_{l_k}.

b) if $l_K = \emptyset$, then no parameters are weakly identified, so $\beta_{K^-,n} = \beta_n$ and

$$n^{1/2} B(\beta_n)(\hat{\theta} - \theta_n) \xrightarrow{d} N(0, \Sigma(\pi_0, \omega_0))$$

Proof of B.2.3. Steps:

1. Normalize the altered for $k = K$ B.4 by $n^{1/2}$ and show ULLN + weak convergence.
2. Use Lemma B.3.3 for $k = K$.
3. Use the FOC expansion together with the two previous steps to get the weak convergence result in (i).
4. Use the criterion expansion from Lemma B.2.2, normalize by n and apply ULLN and weak convergence result
5. Recognize that $\hat{\pi}_K$ minimizes the left hand side and π_K^* minimizes the right hand side by definition. Apply the argmax CMT to obtain $\hat{\pi}_{l_K} \xrightarrow{d} \pi_K^*$.
6. Recognize that $\hat{\psi}_{K^-}(\hat{\pi}_K) = \hat{\psi}_{K^-}$, add and subtract $\psi^0_{K^-,n}$ in $n^{1/2} B(\beta_{K^-,n})\big(\hat{\psi}_{K^-} - \psi_{K^-,n}\big)$, and apply the CMT to arrive at the joint convergence result in (ii).
7. Proof of part (b) is standard.

We utilize the same expansions conducted in the proof of Lemma B.2.2, and we reference these expansions without explicitly rederiving them in this proof for conciseness. Consider the first order condition and expansion in B.3 and B.4 for the case $k = K$ and recall that $K^+ = \emptyset$, as group K is the last group.

$$\hat{\psi}_{K^-}(\pi_{l_K}) - \psi^0_{K^-,n} = -[\nabla^2_{\psi_{K^-}} Q_n(\tilde{\psi}_{K^-,n}, \pi_{l_K})]^{-1} \nabla_{\psi_{K^-}} Q_n(\psi^0_{K^-,n}, \pi_{l_K}) \tag{B.9}$$

with

$$\nabla_{\psi_{K^-}} Q_n(\psi^0_{K^-,n}, \pi_{l_K}) = \mathcal{K}_{K,n}(\psi^0_{K^-,n}, \pi_{l_K}; \tilde{\gamma}_n) \cdot \beta_{l_K,n} + \mathcal{G}_{K,n}(\pi_{l_K}). \tag{B.10}$$

Substitute these into B.2 for $k = K$ to arrive at equation B.5 for $k = K$, rewritten here:

$$
\begin{aligned}
&Q_n(\hat{\psi}_{K^-}(\pi_{l_K}), \pi_{l_K}) - Q_n(\psi^0_{K^-,n}) \\
&\quad = -\nabla_{\psi_{K^-}} Q_n(\psi^0_{K^-,n}, \pi_{l_K})' [\nabla^2_{\psi_{K^-}} Q_n(\tilde{\psi}_{K^-,n}, \pi_{l_K})]^{-1} \nabla_{\psi_{K^-}} Q_n(\psi^0_{K^-,n}, \pi_{l_K}) \\
&\qquad + \frac{1}{2} \nabla_{\psi_{K^-}} Q_n(\psi^0_{K^-,n}, \pi_{l_K})' [\nabla^2_{\psi_{K^-}} Q_n(\tilde{\psi}_{K^-,n}, \pi_{l_K})]^{-1} \nabla^2_{\psi_{K^-}} Q_n(\tilde{\tilde{\psi}}_{K^-,n}, \pi_{l_K}) \\
&\qquad \times [\nabla^2_{\psi_{K^-}} Q_n(\tilde{\psi}_{K^-,n}, \pi_{l_K})]^{-1} \nabla_{\psi_{K^-}} Q_n(\psi^0_{K^-,n}, \pi_{l_K}) \qquad \text{(B.11)}
\end{aligned}
$$

for some $\tilde{\psi}_{K^-,n}$ and $\tilde{\tilde{\psi}}_{K^-,n}$, both between $\hat{\psi}_{K^-}(\pi_{l_K})$ and $\psi^0_{K^-,n}$.

Recall that $\beta_{l_K,n} n^{1/2} \to b_{l_K}$, and Lemmas B.3.2, B.3.3, and B.3.5 imply

$$
\begin{aligned}
&\sup_{\pi_{l_K}} ||B(\beta_{K^-,n})^{-1} \nabla^2_{\psi_{K^-}} Q_n(\tilde{\psi}_{K^-,n}, \pi_{l_K}) B(\beta_{K^-,n})^{-1} - H_K(\pi_{l_K}; \gamma_0)|| \xrightarrow{p} 0 \\
&\sup_{\pi_{l_K}} ||B(\beta_{K^-,n})^{-1} \mathcal{K}_{K,n}(\pi_{l_K}; \tilde{\gamma}_n) - \mathcal{K}_K(\pi_{l_K}; \gamma_0)|| \to 0 \\
&\sqrt{n} B(\beta_{K^-})^{-1} \mathcal{G}_{K,n}(\pi_{l_K}) \Rightarrow \mathcal{G}(\pi_{l_K}; \gamma_0).
\end{aligned}
$$

Normalize equation B.11, and apply the above results to arrive at B.7:

$$
n\big(Q^c_n(\pi_{l_K}) - Q_n(\psi^0_{K,n}, \pi_{l_K})\big) \Rightarrow \chi(\pi_{l_K})
$$

where

$$
\begin{aligned}
\chi(\pi_{l_K}) = -\frac{1}{2} \Big(\mathcal{K}_K(\pi_{l_K}; \gamma_0) b_{l_K} + \mathcal{G}(\pi_{l_K}; \gamma_0)\Big)' \Big[H_K(\pi_{l_K}; \gamma_0)\Big]^{-1} \\
\times \Big(\mathcal{K}_K(\pi_{l_K}; \gamma_0) b_{l_K} + \mathcal{G}(\pi_{l_K}; \gamma_0)\Big).
\end{aligned}
$$

This establishes (a.i). To establish part (a.ii), observe that $Q_n(\psi^0_{K,n}, \pi_{l_k})$ does not depend upon π_{l_k} by assumption, so the left hand side of B.7 is minimized by $\hat{\pi}_{l_k}$ by definition. Further, by assumption, the right hand side is minimized by $\pi^*_{l_k}$. Apply the argmax CMT (van der Vaart and Wellner, 1996) to see that $\hat{\pi}_{l_k} \xrightarrow{d} \pi^*_{l_k}$.

The joint result B.8 follows by normalizing the first order condition in B.9-B.10 by $n^{1/2}B(\beta_{K^-})$ and application of the Lemmas B.3.2, B.3.3, and B.3.5 together with the CMT and the result that

$$\begin{aligned} n^{1/2}B(\beta_{K^-,n})\big(\hat{\psi}_{K^-} - \psi^0_{K^-,n}\big) &= n^{1/2}B(\beta_{K^-,n})\big(\hat{\psi}_{K^-} - \psi_{K^-,n}\big) \\ &\quad + n^{1/2}B(\beta_{K^-,n})\big(\psi_{K^-,n} - \psi^0_{K^-,n}\big) \\ &= n^{1/2}B(\beta_{K^-,n})\big(\hat{\psi}_{K^-} - \psi_{K^-,n}\big) + S_k\beta_{l_K}. \end{aligned}$$

This gives

$$\begin{pmatrix} n^{1/2}B(\beta_{K^-,n})\big(\hat{\psi}_{K^-} - \psi_{K^-,n}\big) \\ \hat{\pi}_{l_K} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \tau(\pi^*_{l_K}) - S_{l_K}b_{l_K} \\ \pi^*_{l_K} \end{pmatrix} \tag{B.12}$$

where $\tau(\pi_{l_K}) = \Big[H_K(\pi_{l_K};\gamma_0)\Big]^{-1}\Big(\mathcal{K}_K(\pi_{l_K};\gamma_0)b_{l_K} + \mathcal{G}(\pi_{l_K};\gamma_0)\Big)$, as desired.

Proof of part (b) is similar to that of (a) with simplifications including use of Lemma B.3.6 in place of B.3.5, and so it is omitted. □

Theorem B.2.4. *Let the assumptions of Theorem B.2.3 hold. Under $\gamma_n \to \gamma_0$,*

a) If $l_K \neq \emptyset$, where l_K indexes the weakly identified subvector of π, then

$$n^{1/2}B(\beta_n)\begin{pmatrix} \big(\hat{\psi}_{K^-} - \psi_{K^-,n}\big) \\ \hat{\pi}_{l_K} - \pi_{l_K,n} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \tau(\pi^*_{l_K}) - S_{l_K}b_{l_K} \\ ||\tau_{\beta_K}(\pi^*_{l_K})||\Big(\pi^*_{l_K} - \pi_{l_K,n}\Big) \end{pmatrix} \tag{B.13}$$

where S_{l_K} is the selection matrix that selects the columns corresponding to $\beta_{(i),l_k}$.

b) if $l_k = \emptyset$, then no parameters are weakly identified, so $\beta_{K^-,n} = \beta_n$ and

$$n^{1/2}B(\beta_n)(\hat{\theta} - \theta_n) \xrightarrow{d} H_{K-1}(\gamma_0)^{-1}\mathcal{G}_\theta(\gamma_0) \tag{B.14}$$

where $\mathcal{G}_\theta(\gamma_0) \sim N(0, \Omega_\theta(\gamma_0))$, and $\chi_\theta(\gamma_0) = -\frac{1}{2}\mathcal{G}_\theta(\gamma_0)'H_{K-1}(\gamma_0)^{-1}\mathcal{G}_\theta(\gamma_0)$.

Proof of Theorem B.2.4. The result follows directly from Theorem B.2.3 by the CMT, since

$n^{1/2}\hat{\beta}_{l_K,n}(\pi_{l_K}) \Rightarrow \tau_{\beta_K}(\pi_{(i),l_K})$ by part a of Theorem B.2.3. □

B.3 Appendix: Supporting Lemmas and Proofs for the Estimator Limit Theory

Lemma B.3.1. *We have that* $\sup_\theta |Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0$ *under the assumptions*

i) Θ *is compact*

ii) $m_t(\theta)$ *is continuously differentiable*

iii) $\sup_{\theta^* \in \Theta^*} ||E_{\gamma_0}\left(\frac{\partial}{\partial\theta} m_t(\theta^*)\right)|| < \infty$

iv) for some $\delta > 0$, $\sup_{\theta^* \in \Theta^*} E_{\gamma_0}|m_t(\theta^*)|^{2+\delta} < \infty$

v) For every $\theta \in \Theta$, $m_t(\theta)$ *is strongly mixing with mixing coefficient* α_m *such that* $\sum_{m=1}^{\infty} \alpha_m^{\delta/(d+\delta)} < \infty$.

Proof of B.3.1. The result will follow from Davidson's (1994) Theorem 21.9 or Newey's (1991) Theorem 2.1 if we show that $|Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0$ for every $\theta \in \Theta$ and that Q_n is stochastically equicontinuous.

Observe that $m_t(\theta)$ is strongly mixing and uniformly $L_{2+\delta}$ bounded for some $\delta > 0$, so $|Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0$ by Corollary 19.6 in Davidson (1994).

Let $\theta, \theta' \in int(\Theta)$ and use differentiability of $Q_n(\theta) = \frac{1}{n}\sum_{t=1}^n m_t(\theta)$ and the MVT to see that

$$\begin{aligned} |Q_n(\theta') - Q_n(\theta)| &= \sum_{i=1}^k \frac{1}{n}\sum_{t=1}^n \frac{\partial}{\partial\theta_i} m_t(\theta^*) \cdot (\theta_i - \theta_i') \; a.s. \\ &\leq B_n \cdot ||\theta - \theta'|| \; a.s. \end{aligned}$$

where $B_n \equiv \sup_{\theta^* \in \Theta^*} ||\left(\frac{1}{n}\sum_{t=1}^n \frac{\partial}{\partial\theta} m_t(\theta^*)\right)||$. By assumption, $B_n = O_p(1)$, so Q_n is stochastically equicontinuous by Davidson's (1994) Theorem 21.10. □

Nest, we establish ULLNs for $\mathcal{K}_{k,n}(\pi_{l_k}, \pi_{k^+}; \tilde{\gamma}_n)$, $\mathcal{G}_{k,n}(\pi_{l_k}, \pi_{k^+})$, and $\nabla^2_{\psi_{k^-}} Q_n(\tilde{\psi}_{k^-,n}, \pi_{l_k}, \pi_{k^+})$, and a weak convergence result for $\mathcal{G}_{k,n}(\pi_{l_k}, \pi_{k^+})$.

Lemma B.3.2. *Define* $\Theta_\eta = \{\theta \in \Theta : ||\beta|| < \eta\}$ *and* $\Gamma_0 = \{(a\beta', \zeta', \pi', \phi')' : (\beta', \zeta', \pi', \phi')' \in \Gamma,\ ||\beta|| < \eta,\ \text{and}\ a \in [0,1]\}$ *for some* $\eta > 0$*, and recall that* $\psi^0_{k^-}$ *is the parameter vector consisting of the concentrated out parameters evaluated at the point of sequential identification failure* $\beta^0_{(i),l_k} = 0$ *and* $\beta^0_{(i),k^+} = 0$*. Let the following assumptions hold:*

i) For every $k = 1, \ldots, K$, $\mathcal{K}_k(\psi_{k^-}, \pi_{l_k}, \pi_{k^+}; \gamma_0)$ *exists for every* $(\theta, \gamma_0) \in \Theta_\eta \times \Gamma_0$*, where* $\theta = (\psi_{k^-}, \pi_{l_k}, \pi_{k^+})$.

ii) For each $k = 1, \ldots, K$, $\mathcal{K}_k(\theta; \gamma)$ *is continuous at* $(\psi^0_{k^-}, \pi_{l_k}, \pi_{k^+}; \gamma^0)$ *uniformly over* $\pi_{l_k}, \pi_{k^+} \in \Pi_{l_k} \times \Pi_{k^+}$ *for every* $\gamma^0 \in \Gamma$ *such that* $\psi^0_{k^-}$ *is a subvector of* γ^0.

iii) $\tilde{\gamma}_n \to \gamma_0$

Then $\sup_{\pi_{l_k}, \pi_{k^+}} ||B(\beta_{k^-,n})^{-1}\mathcal{K}_{k,n}(\pi_{l_k}, \pi_{k^+}; \tilde{\gamma}_n) - \mathcal{K}_k(\pi_{l_k}, \pi_{k^+}; \gamma_0)|| \to 0$ *where* $\mathcal{K}_{k,n}(\pi_{l_k}, \pi_{k^+}; \tilde{\gamma}_n) \equiv \mathcal{K}_{k,n}(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+}; \tilde{\gamma}_n)$ *and* $\mathcal{K}_k(\psi_{k^-,0}, \pi_{l_k}, \pi_{k^+}; \gamma_0) \equiv \mathcal{K}_k(\pi_{l_k}, \pi_{k^+}; \gamma_0)$.

Recall that $\psi^0_{k^-,n} \to \psi_{k^-,0}$ *by definition. Note also that the assumptions are similar to Assumption S4 in Andrews and Cheng (2013), which is related to Assumption C5 of Andrews and Cheng (2012a).*

Lemma B.3.3. *For each* $k = 1, \ldots, K$*, define* $H_k(\pi_{l_k}, \pi_{k^+}; \gamma_0) = \lim_{n\to\infty} E_{\gamma_n}[B(\beta_{k^-,n})^{-1}\nabla^2_{\psi_{k^-}} m_t(\psi_{k^-,0}, \pi_{l_k}, \pi_{k^+})B(\beta_{k^-,n})^{-1}]$. *Then under the following assumptions, we have that* $\sup_\theta ||B(\beta_{k^-,n})^{-1}\nabla^2_{\psi_{k^-}} Q_n(\tilde{\psi}_{k^-,n}, \pi_{l_k}, \pi_{k^+})B(\beta_{k^-,n})^{-1} - H_k(\pi_{l_k}, \pi_{k^+}; \gamma_0)|| \xrightarrow{p} 0$ *for each* $k = 1, \ldots, K$.

i) Θ *is compact*

ii) $\tilde{\psi}_{k^-,n} \to \psi_{k^-,0}$ *uniformly on* $\Pi_{l_k} \times \Pi_{k^+}$ *for each* k

iii) $m_t(\theta)$ *is three times continuously differentiable.*

iv) Define θ_- *such that* $\theta = (\theta'_-, \pi'_K)'$*. Then* $\sup_{\theta^* \in \Theta^*} ||E_{\gamma_0}\left(\frac{\partial}{\partial \theta_-} vec\left(\nabla^2_{\theta_-} m_t(\theta^*)\right)\right)|| < \infty$

v) for every i, j *and some* $\delta > 0$, $\sup_{\theta^* \in \Theta^*} E_{\gamma_0}|\nabla^2_{\theta_-} m_{i,j,t}(\theta^*)|^{2+\delta} < \infty$

vi) For every $\theta \in \Theta$, $\nabla^2_{\theta_-} m_t(\theta)$ is strongly mixing with mixing coefficient α_m such that $\sum_{m=1}^{\infty} \alpha_m^{\delta/(d+\delta)} < \infty$.

By operating component wise on the matrix $\nabla^2_{\psi_{k^-}} m_t(\theta)$, the proof follows exactly as in B.3.1 with the added step that involves appealing to Theorem 21.6 in Davidson (1994).

Lemma B.3.4. *Under the conditions of Lemma B.3.5 and for $k = 1, \ldots, K-1$, we have*

$$\sup_{\pi_{l_k}, \pi_{k^+}} ||||\beta_{l_k,n}||^{-1} B(\beta_{k^-,n})^{-1} \mathcal{G}_{k,n}(\pi_{l_k}, \pi_{k^+})|| \xrightarrow{p} 0$$

Proof. Lemma B.3.5 implies that $\sqrt{n} B(\beta_{k^-})^{-1} \mathcal{G}_{k,n}(\pi_{l_k}, \pi_{k^+})$ is $O_p(1)$ uniformly over $\Pi_{l_k} \times \Pi_{k^+}$. The result follows, since $\beta_{l_k} = o(n^{1/2})$ for every $k = 1, \ldots, K-1$, so that $\beta_{l_k}/\sqrt{n} = o(1)$. □

Lemma B.3.5. *Recall that for $k = 1, \ldots, K-1$*

$$\begin{aligned}\mathcal{G}_{k,n}(\pi_{l_k}, \pi_{k^+}) &= \nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+}) - E_{\gamma_n}\left[\nabla_{\psi_{k^-}} Q_n(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})\right] \\ &= \frac{1}{n}\sum_{t=1}^{n}\left\{\nabla_{\psi_{k^-}} m_t(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+}) - E_{\gamma_n}\left[\nabla_{\psi_{k^-}} m_t(\psi^0_{k^-,n}, \pi_{l_k}, \pi_{k^+})\right]\right\}\end{aligned}$$

and for $k = K$, the grouping $k^+ = \emptyset$, so

$$\mathcal{G}_{K,n}(\pi_{l_K}) = \frac{1}{n}\sum_{t=1}^{n}\left\{\nabla_{\psi_{K^-}} m_t(\psi^0_{K^-,n}, \pi_{l_K}) - E_{\gamma_n}\left[\nabla_{\psi_{K^-}} m_t(\psi^0_{K^-,n}, \pi_{l_K})\right]\right\}.$$

Let $\mathcal{G}(\pi_{l_K}; \gamma_0)$ be a zero mean Gaussian process with covariance kernel $\Omega(\pi_{l_K}, \tilde{\pi}_{l_K}; \gamma_0)$. Under $\gamma_n \to \gamma_0$ and the assumptions

i) $\{W_t\}$ is strongly mixing with mixing coefficient $\alpha(j)$ such that $\sum_{j=1}^{\infty} \alpha(j)^{\delta/(2+\delta)} < \infty$ for some $\delta > 0$.

ii) m_t is measurable with respect to $\sigma(W_t)$.

iii) Θ is compact

iv) $m_t(\theta)$ is twice continuously differentiable

v) $\sup_{\theta\in\Theta} \lim_{n\to\infty} E_{\gamma_n}||B(\beta_{K^-,n})^{-1}\nabla_{\psi_{K^-}} m_t(\theta)||^{2+\delta} < \infty$

vi) $\sup_{\theta\in\Theta} \lim_{n\to\infty} E_{\gamma_n}||B(\beta_{K^-,n})^{-1}\nabla^2_{\psi_{K^-}} m_t(\theta)B(\beta_{K^-,n})^{-1}||^{2+\delta} < \infty$

we have $\sqrt{n}B(\beta_{K^-,n})^{-1}\mathcal{G}_{K,n}(\pi_{l_K}) \Rightarrow \mathcal{G}(\pi_{l_K};\gamma_0)$.

Proof. In order to establish the result, we must show finite dimensional convergence and stochastic equicontinuity (Andrews, 1994; Pollard, 1990). Stochastic equicontinuity follows from an application of the MVT and the moment bounds in (v) as elaborated by Davidson's (1994) Theorem 21.10. Finite dimensional convergence follows from appealing to an α-mixing CLT (Ibragimov, 1962) to establish convergence of a linear combination

$$(\sqrt{n}B(\beta_{K^-,n})^{-1}\mathcal{G}_{K,n}(\pi_{l_K,1}),\ldots,\sqrt{n}B(\beta_{K^-,n})^{-1}\mathcal{G}_{K,n}(\pi_{l_K,J})),$$

and then applying the Cramér-Wold theorem.

Note that under the same conditions, we have that

$$\sqrt{n}B(\beta_{k^-,n})^{-1}\mathcal{G}_{k,n}(\pi_{l_k},\pi_{k^+}) \Rightarrow \mathcal{G}_k(\pi_{l_K},\pi_{k^+};\gamma_0)$$

for $k = 1,\ldots,K-1$, as well; where $\mathcal{G}_k(\pi_{l_K},\pi_{k^+};\gamma_0)$ is a Gaussian process with covariance kernel $\Omega_k(\pi_{l_k},\pi_{k^+},\tilde{\pi}_{l_k},\tilde{\pi}_{k^+};\gamma_0)$. □

Lemma B.3.6. *Recall that when l_K is empty, $\psi_{k^-} = \theta$, so*

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} B(\beta_{K^-,n})^{-1}\nabla_{\psi_{k^-}} m_t(\theta) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} B(\beta_n)^{-1}\nabla_\theta m_t(\theta).$$

Define $\mathcal{G}_\theta(\gamma_0)$ to be a Gaussian random variable with covariance matrix $\Omega_{(i),\theta}(\gamma_0)$.

Under $\gamma_n \to \gamma_0$ and the assumptions

i) $\{W_t\}$ is strongly mixing with mixing coefficient $\alpha(j)$ such that $\sum_{j=1}^{\infty}\alpha(j)^{\delta/(2+\delta)} < \infty$ for some $\delta > 0$.

ii) m_t is measurable with respect to $\sigma(W_t)$.

iii) Θ *is compact*

iv) $m_t(\theta)$ *is continuously differentiable*

v) $\lim_{n\to\infty} E_{\gamma_n}||B(\beta_{K^-,n})^{-1}\nabla_{\psi_{K^-}} m_t(\theta_n)||^{2+\delta} < \infty$

we have $\frac{1}{\sqrt{n}}\sum_{t=1}^{n} B(\beta_n)^{-1}\nabla_\theta m_t(\theta) \xrightarrow{d} \mathcal{G}_\theta(\gamma_0)$.

Observe that the assumptions are weaker than those imposed in B.3.5 as stochastic equicontinuity need not be established. The proof follows from application of an α*-mixing CLT (Ibragimov, 1962).*

B.4 Appendix: Proofs for the Parsimonious Estimator Limit Theory

First, we discuss the limit theory for the individual parsimonious estimators. The results in this first subsection follow directly from results derived in Appendix B.2. After detailing this limiting distribution, we prove in the following subsection the results for the joint limit theory described in the paper.

B.4.1 Appendix: Pointwise Parsimonious Estimator Limit Theory

Assumption 23. *i) If* l_K *is empty, then* $\lambda_{min}(\Omega_{(i),\theta}(\gamma_0)) \geq \varepsilon$ *for some* $\varepsilon > 0$ *and every* i.

ii) If l_K *is not empty, then for every* i*, each sample path of the process* $\chi_{(i)}(\pi_{(i),l_K})$ *is continuous a.s. and minimized uniquely with probability 1. Denote the minimizer by* $\pi^*_{(i),l_K}$.

Theorem B.4.1. *Let Assumptions 1-7 and 18 hold. Under* $\gamma_n \to \gamma_0$*,*

a) If $l_K \neq \emptyset$*, where* l_K *indexes the weakly identified subvector of* $\pi_{(i)}$*, then*

$$n\big(Q^c_{(i),n}(\pi_{(i),l_k}) - Q_{(i),n}(\psi^0_{(i),K,n}, \pi_{(i),l_k})\big) \Rightarrow \chi_{(i)}(\pi_{(i),l_K}; \gamma_0) \tag{B.15}$$

$$\begin{pmatrix} n^{1/2}B(\beta_{(i),K^-,n})\big(\hat{\psi}_{(i),K^-} - \psi_{(i),K^-,n}\big) \\ \hat{\pi}_{(i),l_K} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \tau_{(i)}(\pi^*_{(i),l_K}) - S_{l_K} b_{(i),l_K} \\ \pi^*_{(i),l_K} \end{pmatrix} \tag{B.16}$$

where S_{l_K} *is the selection matrix that selects the columns corresponding to* $\beta_{(i),l_k}$.

b) if $l_K = \emptyset$, then no parameters are weakly identified, so $\beta_{(i),K^-,n} = \beta_{(i),n}$ and

$$n\big(Q_{(i),n}(\hat{\theta}_{(i)}) - Q_{(i),n}(\theta_{(i),n})\big) \xrightarrow{d} \chi_{(i),\theta}(\gamma_0) \tag{B.17}$$

$$n^{1/2}B(\beta_{(i),n})(\hat{\theta}_{(i)} - \theta_{(i),n}) \xrightarrow{d} H_{(i),K-1}(\gamma_0)^{-1}\mathcal{G}_{(i),\theta}(\gamma_0) \tag{B.18}$$

where $\mathcal{G}_{(i),\theta}(\gamma_0) \sim N(0, \Omega_{(i),\theta}(\gamma_0))$, and $\chi_{(i),\theta}(\gamma_0) = -\frac{1}{2}\mathcal{G}_{(i),\theta}(\gamma_0)' H_{(i),K-1}(\gamma_0)^{-1}\mathcal{G}_{(i),\theta}(\gamma_0)$.

The proof of Theorem B.4.1 follows directly from Theorem B.2.3.

Theorem B.4.1 details the pointwise in i asymptotic distribution of the parsimonious estimators. However, the max test combines estimators across parsimonious models, so it is necessary that we analyze the joint limiting distribution of the parsimonious estimators. Theorem 3.4.3 provides this joint asymptotic distribution.

A test directly based on the normalization described by Theorem B.4.1 will not always be consistent when including weakly identified parameters. This is demonstrated in Lemma B.6.1 in the Appendix. The following theorem provides a more convenient normalization; however, one should still note that use of this theorem does not provide consistency against all departures from the null hypothesis. This is an ongoing issue with testing weakly identified parameters, and current research focuses on correcting the size distortions that result from ignoring the effect of weak identification.

Corollary B.4.2. *Let Assumptions 1-7 and 18 hold. Under $\gamma_n \to \gamma_0$,*

a) If $l_K \neq \emptyset$, where l_K indexes the weakly identified subvector of $\pi_{(i)}$, then

$$n^{1/2}B(\beta_{(i),n}) \begin{pmatrix} (\hat{\psi}_{(i),K^-} - \psi_{(i),K^-,n}) \\ \hat{\pi}_{(i),l_K} - S_{\pi_{(i),l_K}}\pi_{l_K,n} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \tau_{(i)}(\pi^*_{(i),l_K}) - S_{l_K}b_{(i),l_K} \\ ||\tau_{(i),\beta_K}(\pi^*_{(i),l_K})|| \Big(\pi^*_{(i),l_K} - S_{\pi_{(i),l_K}}\pi_{l_K,n}\Big) \end{pmatrix} \tag{B.19}$$

where S_{l_K} is the selection matrix that selects the columns corresponding to $\beta_{(i),l_k}$, and $S_{\pi_{(i),l_K}}$ selects the elements of the vector $\pi_{l_K,n}$ corresponding to $\pi_{(i),l_K,n}$.

b) if $l_k = \emptyset$, then no parameters are weakly identified, so $\beta_{(i),K^-,n} = \beta_{(i),n}$ and

$$n^{1/2} B(\beta_{(i),n})(\hat{\theta}_{(i)} - \theta_{(i),n}) \xrightarrow{d} H_{(i),K-1}(\gamma_0)^{-1} \mathcal{G}_{(i),\theta}(\gamma_0) \tag{B.20}$$

where $\mathcal{G}_{(i),\theta}(\gamma_0) \sim N(0, \Omega_{(i),\theta}(\gamma_0))$, and $\chi_{(i),\theta}(\gamma_0) = -\frac{1}{2}\mathcal{G}_{(i),\theta}(\gamma_0)' H_{(i),K-1}(\gamma_0)^{-1} \mathcal{G}_{(i),\theta}(\gamma_0)$.

The proof of Corollary B.4.2 follows directly from Theorem B.2.4.

At first, it seems that the centering term for the weakly identified parameters $\pi_{(i),l_K,n}$ instead of $S_{\pi_{(i),l_K}} \pi_{l_K,n}$. However, this term is arbitrary, since $\hat{\pi}_{(i),l_K}$ is not a consistent estimator. Later, we will see that it is convenient to center the weakly identified parameters around the null hypothesized values.

B.4.2 Appendix: Proofs for the Joint Parsimonious Estimator Limit Theory

Proof of Theorem 3.4.3. The proof of Theorem 3.4.3 follows from an argument nearly identical to that in Theorem B.4.1 and B.4.2 where Lemma B.4.3 is used in place of Lemma B.3.5 and Assumption 9 is used in place of Assumption 8. □

Lemma B.4.3. *Let $l_{(i),K}$ denote the index set l_K for parsimonious model i, and recall $m_{(i),t}(\theta_{(i)}) = m_t([\theta]_{(i)})$ and*

i) when $l_{(i),K}$ is not empty,

$$\mathcal{G}_{(i),K,n}(\pi_{(i),l_K}) = \frac{1}{n}\sum_{t=1}^{n} \Big\{ \nabla_{\psi_{(i),K^-}} m_{(i),t}(\psi^0_{(i),K^-,n}, \pi_{(i),l_K}) - E_{\gamma_n}\left[\nabla_{\psi_{(i),K^-}} m_{(i),t}(\psi^0_{(i),K^-,n}, \pi_{(i),l_K})\right] \Big\}.$$

ii) when $l_{(i),K}$ is empty, $\psi_{(i),k^-} = \theta_{(i)}$, so

$$\nabla_{\psi_{(i),k^-}} m_{(i),t}(\theta_{(i)}) = \nabla_{\theta_{(i)}} m_{(i),t}(\theta_{(i)}).$$

Define $\mathcal{G}_{(i),\theta,n} = \frac{1}{n}\sum_{t=1}^{n} \nabla_{\theta_{(i)}} m_{(i),t}(\theta_{(i),n})$.

Define

$$G_{(i),n}(\pi_{(i),l_K}) = \begin{cases} \mathcal{G}_{(i),K,n}(\pi_{(i),l_K}) & \text{if } l_{(i),K} \neq \emptyset \\ \mathcal{G}_{(i),\theta,n} & \text{if } l_{(i),K} = \emptyset. \end{cases}$$

Let $\mathcal{G}_{(i)}(\pi_{(i),l_K}; \gamma_0)$ *be a zero mean Gaussian process with covariance kernel* $\Omega_{(i)}(\pi_{(i),l_K}, \tilde{\pi}_{(i),l_K}; \gamma_0)$ *and* $\mathcal{G}_{(i),\theta}(\gamma_0)$ *be a Gaussian random variable with covariance matrix* $\Omega_{(i),\theta}(\gamma_0)$.

Let $\gamma_n \to \gamma_0$ *and the assumptions hold:*

i) $\{W_t\}$ *is strongly mixing with mixing coefficient* $\alpha(j)$ *such that* $\sum_{j=1}^{\infty} \alpha(j)^{\delta/(2+\delta)} < \infty$ *for some* $\delta > 0$.

ii) m_t *is measurable with respect to* $\sigma\{W_t\}$, *the sigma field generated by* $\{W_t\}$.

iii) Θ *is compact*

iv) $m_t(\theta)$ *is twice continuously differentiable*

v) $\sup_{\theta \in \Theta} \lim_{n \to \infty} E_{\gamma_n} ||B(\beta_{K^-,n})^{-1} \nabla_{\psi_{K^-}} m_t(\theta)||^{2+\delta} < \infty$

vi) $\sup_{\theta \in \Theta} \lim_{n \to \infty} E_{\gamma_n} ||B(\beta_{K^-,n})^{-1} \nabla^2_{\psi_{K^-}} m_t(\theta) B(\beta_{K^-,n})^{-1}||^{2+\delta} < \infty$

Then

$$\left\{ \sqrt{n} B(\beta_{(i),K^-,n})^{-1} G_{(i),n}(\pi_{(i),l_K}) : \ 1 \leq i \leq \mathring{k} \right\} \Rightarrow \left\{ \tilde{G}_{(i)}(\pi_{(i),l_K}; \gamma_0) : \ 1 \leq i \leq \mathring{k} \right\},$$

a zero mean Gaussian process with covariance kernel $\Omega_{(i,j)}(\pi_{(i),l_{(i),K}}, \pi_{(j),l_{(j),K}}; \gamma_0)$ *and where*

$$\tilde{G}_{(i)}(\pi_{(i),l_K}; \gamma_0) = \begin{cases} \mathcal{G}_{(i)}(\pi_{(i),l_K}; \gamma_0) & \text{if } l_{(i),K} \neq \emptyset \\ \mathcal{G}_{(i),\theta}(\gamma_0) & \text{if } l_{(i),K} = \emptyset. \end{cases}$$

Proof. Establishing the result requires showing finite dimensional convergence and stochastic equicontinuity (Andrews, 1994; Pollard, 1990). Stochastic equicontinuity follows from the fact

that the set $\{1, \ldots, \mathring{k}\}$ is compact and discrete, that each of the components $\mathcal{G}_{(i),K,n}(\pi_{(i),l_K})$ are stochastically equicontinuous as shown in Lemma B.3.5, and probability sub-additivity.

To establish finite dimensional convergence, let $A = [a_i]_{i=1,\ldots,\tilde{r}}$ with each $a_i \in \mathbb{R}^{d_{\theta_{(i)}}}$ and with $A'A = 1$, and consider the linear combination

$$\begin{aligned}
&\sum_{i=1}^{r} \sqrt{n} B(\beta_{(i),K^-,n})^{-1} \sum_{m=1}^{s} a_i' \mathcal{G}_{(i),K,n}(\pi_{(i),l_K,m}) + \sum_{j=rs+1}^{\tilde{r}} a_j' \sqrt{n} B(\beta_{(j),K^-,n})^{-1} \mathcal{G}_{(j),\theta,n} \\
&\quad = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \sum_{i=1}^{r} \sum_{m=1}^{s} a_i' B(\beta_{(i),K^-,n})^{-1} \\
&\qquad \times \Big\{ \nabla_{\psi_{(i),K^-}} m_{(i),t}(\psi^0_{(i),K^-,n}, \pi_{(i),l_K,m}) \\
&\qquad\quad - E_{\gamma_n}\left[\nabla_{\psi_{(i),K^-}} m_{(i),t}(\psi^0_{(i),K^-,n}, \pi_{(i),l_K,m}) \right] \Big\} \\
&\qquad + \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \sum_{j=rs+1}^{\tilde{r}} a_j' B(\beta_{(j),K^-,n})^{-1} \nabla_{\theta_{(i)}} m_{(i),t}(\theta_{(i),n})
\end{aligned}$$

where without loss of generality, the indices have been ordered so that all i with weakly identified parameters come first. Use the assumptions above an invoke an α-mixing CLT (Ibragimov, 1962) to establish that this converges to a zero mean Normal random variable with variance that depends upon A and the vector $[\pi_{(i),l_K,m}]_{\substack{i=1,\ldots,r \\ m=1,\ldots,s}}$. A Cramér-Wold device then establishes the finite dimensional convergence result. □

B.5 Appendix: Proofs for the Max Test

Proof of Theorem 3.5.1. Together, $W_{n,i} \xrightarrow{p} W_i$ and Theorem 3.4.3 imply that under H_0 and $\gamma_n \to \gamma_0$,

$$\mathcal{N}_{(i),\lambda,n} W_{n,i} \hat{\lambda}_{(i)} - W_i S'_{(i),\lambda} \mathfrak{Z}_{(i)} \xrightarrow{p} 0$$

for each $i = 1, \ldots, \mathring{k}$ where $\mathring{k} \leq \lim_{n\to\infty} \mathring{k}_n$ and where $\mathcal{N}_{(i),\lambda,n} = S'_{(i),\lambda}(diag(n^{1/2}\ B_{(i)}(\beta_{(i),K^-,n}))', 1'_{d_{\pi_{(i),l_K}}})'$ and $S_{(i),\lambda}$ is the selection matrix that selects the element corresponding to $\lambda_{(i)}$. Now apply Lemma 4.2 in Hill and Dennis (2018) to arrive at the result. □

Proof of Lemma 3.5.2. Here we show that the bootstrapped estimator converges weakly in probability to a random variable with the same distribution as given in the limit of Theorem 3.4.3. Intuitively, this follows because, conditional on the sample, $\hat{\mathcal{G}}^{bs}_{(i)}(\pi_{(i),l_K})$ converges to a Gaussian process with the same distribution as $\mathcal{G}_{(i)}(\pi_{(i),l_K};\gamma_0)$, resulting in $\hat{\tau}^{bs}_{(i)}(\pi_{(i),l_K};\gamma_0,b)$ and $\hat{\chi}^{bs}_{(i)}(\pi_{(i),l_K};\gamma_0,b)$ converging to the respective Gaussian and Chi-square processes. Invoking the argmax continuity theorem then gives that $\pi^{*,bs}_{(i),l_K}(\gamma_0,b)$ converges to $\pi^{*}_{(i),l_K}(\gamma_0,b)$. Joint convergence occurs by the same arguments used in the proof of 3.4.3.

We only prove the claim under the case $l_K \neq \emptyset$ for which weakly identified parameters are present. The proof for the claim when there are no weakly identified parameters is similar but simpler, as several of the steps needed when $l_K \neq \emptyset$ are not necessary. This is due to the inconsistency of $\hat{\pi}_n$ for π_0 under the case $l_K \neq \emptyset$ and the required bootstrap step for calculating the bootstrapped π^*, and the joint convergence of $\hat{\pi}_n$ with the other variables.

Recall that

$$\hat{\mathcal{G}}^{bs}_{(i)}(\pi_{(i),l_K}) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} z_t \Big\{ m_{(i),t}(\hat{\psi}^{0}_{(i),K^-,n}(\pi_{(i),l_K}),\pi_{(i),l_K}) - \frac{1}{n}\sum_{t=1}^{n} m_{(i),t}(\hat{\psi}^{0}_{(i),K^-,n}(\pi_{(i),l_K}),\pi_{(i),l_K})\Big\}.$$

for $N(0,1)$ z_t. First, we prove

$$\{\hat{\mathcal{G}}^{(bs)}_{(i)}(\pi) : \pi \in \Pi\} \Rightarrow^{p} \{\mathcal{G}_{(i)}(\pi;\gamma_0) : \pi \in \Pi\} \tag{B.21}$$

where $\mathcal{G}_{(i)}(\pi;\gamma_0)$ is the mean zero Gaussian process, with covariance kernel $\Omega_{(i)}(\pi,\tilde{\pi};\gamma_0)$, the weak limit of $\mathcal{G}_{(i),n}(\cdot)$ when some parameters are weakly identified. Together with uniform convergence in probability of $H_{(i),n}(\hat{\psi}_{0,n},\pi)$ to $H_{(i)}(\pi;\gamma_0)$ and $K_{(i),n}(\tilde{\psi}_n,\pi;\tilde{\gamma}_n)$ to $K_{(i)}(\psi_0,\pi;\gamma_0)$, this step will imply $\{\tau^{(bs)}_{(i)}(\pi;\gamma_0,b) : \pi \in \Pi\} \Rightarrow^{p} \{\tau_{(i)}(\pi;b,\gamma_0) : \pi \in \Pi\}$. Then the argmax continuity theorem (cf van der Vaart and Wellner (1996), Lemma 3.2.1 and Andrews and Cheng (2012b),

Theorem 9.10.) will yield

$$\pi^*_{(i),(bs)}(\gamma_0, b) \xrightarrow{d} \pi^*_{(i)}(\gamma_0, b). \tag{B.22}$$

Joint convergence over i follows from a Cramér-Wold device.

Operate conditionally on the sample $\mathcal{W}_n \equiv \{X_t, Y_t, Z_t\}_{t=1}^n$. First, we prove B.21. We must prove convergence in finite dimensional distributions and establish stochastic equicontinuity (see Giné and Zinn (1990), Andrews (1994), or Pollard (1990)).

We prove convergence in finite dimensional distributions with an argument in Hansen (1996). By construction of z_t, $\hat{G}_n^{(bs)}(\pi)$ is normally distributed with mean zero and covariance kernel

$$\begin{aligned}
&E\Big(\hat{G}^{(bs)}_{(i)}(\pi)\hat{G}^{(bs)}_{(i)}(\tilde{\pi})'|\mathcal{W}_n\Big) \\
&\quad = \frac{1}{n}\sum_{t=1}^{n}\Bigg[\Big(m^{\psi}_{(i),t}(\psi_{0,n},\pi) - \frac{1}{n}\sum_{t=1}^{n} m^{\psi}_{(i),t}(\psi_{0,n},\pi)\Big) \\
&\qquad\qquad \times \Big(m^{\psi}_{(i),t}(\psi_{0,n},\tilde{\pi}) - \frac{1}{n}\sum_{t=1}^{n} m^{\psi}_{(i),t}(\psi_{0,n},\tilde{\pi})\Big)'\Bigg] \\
&\quad = \hat{\Omega}_{(i)}(\pi,\tilde{\pi})
\end{aligned}$$

where $\hat{\Omega}_{(i)}(\pi,\tilde{\pi})$ is defined implicitly. Let $\mathcal{W}$ be the set of samples such that

$$\sup_{\pi,\tilde{\pi}\in\Pi\times\Pi} ||E\Big(\hat{\mathcal{G}}^{(bs)}_{(i)}(\pi)\hat{\mathcal{G}}^{(bs)}_{(i)}(\tilde{\pi})'|\mathcal{W}_n\Big) - \Omega_{(i)}(\pi,\tilde{\pi};\gamma_0)|| \xrightarrow{p} 0.$$

Then $\sup_{\pi,\tilde{\pi}\in\Pi\times\Pi} ||\hat{\Omega}_{(i)}(\pi,\tilde{\pi}) - \Omega(\pi,\tilde{\pi};\gamma_0)|| \xrightarrow{p} 0$ follows from stationary mixing and the moment bounds in Assumptions 1 and 5 establishing that $P(\mathcal{W}_n \in \mathcal{W}) = 1$. Thus $\hat{\mathcal{G}}^{(bs)}_{(i)}(\pi)$ converges in finite dimensional distributions to a zero mean Gaussian process with covariance kernel $\Omega(\pi,\tilde{\pi};\gamma_0)$. Since Gaussian processes are characterized by their first two moments, the finite dimensional distributions of $\hat{\mathcal{G}}^{(bs)}_{(i)}(\pi)$ and $\mathcal{G}_{(i)}(\pi)$ converge to the same limit.

Next, we show stochastic equicontinuity. Since the set $\{1,\ldots,\mathring{k}\}$ is compact and discrete, and accounting for this set with i involves only invoking probability sub-additivity, we ignore subscript

i for clarity (see Lemma D.3). Stochastic equicontinuity follows from the same argument used in Lemma C.5. Let $r \in \mathbb{R}^{dim(\psi_{K_-})}$ be such that $r'r = 1$. The mean value theorem yields

$$r'\Big(m_t^{\psi}(\psi_{0,n}, \pi) - m_t^{\psi}(\psi_{0,n}, \tilde{\pi})\Big) \leq \sup_{\mathring{\pi} \in \Pi} ||r' \frac{\partial}{\partial \pi} m_t^{\psi}(\psi_{0,n}, \mathring{\pi})|| \times ||\tilde{\pi} - \pi||.$$

Next, use the construction of z_t and the fact that z_t is independent of the data and Chebychev's inequality, and observe the following:

$$\begin{aligned}
\mathcal{P}_n(\eta) &= P\Big(\sup_{\pi, \tilde{\pi} \in \Pi : ||\tilde{\pi} - \pi|| \leq \delta} \Big| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} z_t r' \Big(m_t^{\psi}(\psi_{0,n}, \pi) - m_t^{\psi}(\psi_{0,n}, \tilde{\pi})\Big)\Big| > \eta | \mathcal{W}_n \Big) \\
&\leq \frac{1}{\eta^2} E\Big[\sup_{\pi, \tilde{\pi} \in \Pi : ||\tilde{\pi} - \pi|| \leq \delta} \Big(\frac{1}{\sqrt{n}} \sum_{t=1}^{n} z_t r' \Big(m_t^{\psi}(\psi_{0,n}, \pi) - m_t^{\psi}(\psi_{0,n}, \tilde{\pi})\Big)\Big)^2 | \mathcal{W}_n \Big] \\
&= \frac{1}{\eta^2} \sup_{\pi, \tilde{\pi} \in \Pi : ||\tilde{\pi} - \pi|| \leq \delta} \frac{1}{n} \sum_{t=1}^{n} \Big(r' \Big(m_t^{\psi}(\psi_{0,n}, \pi) - m_t^{\psi}(\psi_{0,n}, \tilde{\pi})\Big)\Big)^2 \\
&\leq \frac{1}{\eta^2} \frac{1}{n} \sum_{t=1}^{n} \sup_{\pi, \tilde{\pi} \in \Pi : ||\tilde{\pi} - \pi|| \leq \delta} \Big(r' \Big(m_t^{\psi}(\psi_{0,n}, \pi) - m_t^{\psi}(\psi_{0,n}, \tilde{\pi})\Big)\Big)^2 \\
&\leq \frac{\delta^2}{\eta^2} \frac{1}{n} \sum_{t=1}^{n} \sup_{\mathring{\pi} \in \Pi} \Big(||r' \frac{\partial}{\partial \pi} m_t^{\psi}(\psi_{0,n}, \mathring{\pi})|| \Big)^2 \\
&= \frac{\delta^2}{\eta^2} C_n
\end{aligned}$$

Now observe that

$$\begin{aligned}
E\left[\frac{1}{n} \sum_{t=1}^{n} \sup_{\mathring{\pi} \in \Pi} \Big(||r' \frac{\partial}{\partial \pi} m_t^{\psi}(\psi_{0,n}, \mathring{\pi})|| \Big)^2 \right] &= E\left[\sup_{\mathring{\pi} \in \Pi} \Big(||r' \frac{\partial}{\partial \pi} m_t^{\psi}(\psi_{0,n}, \mathring{\pi})|| \Big)^2 \right] \\
&= O(1)
\end{aligned}$$

by Assumption 5. Hence stationarity and ergodicity imply that $C_n \xrightarrow{p} C$ for a finite non-negative constant C. Take $\delta > 0$ such that $0 < \delta \leq (\varepsilon \eta^2 / C)^{1/2}$ to see that for every $(\varepsilon, \eta) > 0$, there is a $\delta > 0$ such that $\lim_{n \to \infty} \mathcal{P}_n(\eta) < \varepsilon$ with probability approaching one with respect to the sample $\mathcal{W}_n$.

Next, we prove B.22. Recall that $\sup_{\pi_{(i)} \in \Pi_{(i)}} ||H_{(i),n}(\hat{\psi}_{(i),0,n}, \pi_{(i)}) - H_{(i)}(\pi_{(i)}; \gamma_0)|| \xrightarrow{p} 0$

and $\sup_{\pi_{(i)} \in \Pi_{(i)}} ||K_{(i),n}(\tilde{\psi}_{(i),n}, \pi_{(i)}; \tilde{\gamma}_n) - K_{(i)}(\psi_{(i),0}, \pi_{(i)}; \gamma_0)|| \xrightarrow{p} 0$ for every pair of sequences $\tilde{\psi}_n \to \psi_0$ and $\tilde{\gamma}_n \to \gamma_0$. This paired with B.21 implies $\{\tau^{(bs)}_{(i),n}(\pi_{(i)}; \gamma_0, b) : \pi_{(i)} \in \Pi_{(i)}\} \Rightarrow^p \{\tau_{(i)}(\pi_{(i)}; b, \gamma_0) : \pi_{(i)} \in \Pi_{(i)}\}$. The argmax continuity theorem (cf van der Vaart and Wellner (1996), Lemma 3.2.1 and Andrews and Cheng (2012b), Theorem 9.10.) then yields $\pi^*_{(i),(bs)}(\gamma_0, b) \xrightarrow{d} \pi^*_{(i)}(\gamma_0, b)$. Joint convergence of $(\tau^{(bs)}_{(i),n}(\pi^*_{(i),(bs)}(\gamma_0, b); \gamma_0, b)', \pi^*_{(i),(bs)}(\gamma_0, b))'$ follows since both objects are functions of the same underlying objects that we have shown to converge. Finally, joint convergence over i follows from arguments mentioned above.

□

Proof of Theorem 3.5.4. The proof of Theorem 3.5.4 proceeds in the same fashion as the proof of Theorem 3.5.1. We first note that Lemma 3.5.2 implies converge in probability of the difference between the relevant distribution functions, and then we invoke Lemma 4.2 in Hill and Dennis (2018) to arrive at the result. □

B.6 Appendix: Additional Proofs

A test based on the normalization described in Theorem B.4.1 will be inconsistent when including elements from π_{l_K}. The reason is that the standardization described for ψ_{K^-} is $\sqrt{n}B(\beta_{K^-})$ while that for π is the constant 1. The following lemma shows that this standardization will result in an inconsistent test. For this reason, we detail the correct standardization in Theorem B.4.2.

Lemma B.6.1. *A test on a subvector of θ that includes elements from π_{l_K} is not consistent when the standardization is based on Theorem B.4.1.*

Proof. Theorem B.4.1 implies that the appropriate standardization for $\hat{\psi}_{K^-}$ is $\sqrt{n}B(\beta_{K^-})$ and that for $\hat{\pi_{l_K}}$ is 1. Hence the max test statistic over a vector $\lambda = (\beta^\lambda, \zeta^\lambda, \pi^\lambda)$

$$\begin{aligned}\hat{\mathcal{T}}_n &= \max_{1 \le i \le \mathring{k}_n} |\mathcal{N}_{i,n} W_{i,n} \hat{\lambda}_i| \\ &\le \max_{1 \le i \le \mathring{k}_{\beta,n}} |\sqrt{n}\hat{\beta}^\lambda_{(i)}| + \max_{1 \le i \le \mathring{k}_{\zeta,n}} |\sqrt{n}\hat{\zeta}^\lambda_{(i)}| + \max_{1 \le i \le \mathring{k}_{\pi,n}} |\sqrt{n}B(\beta_{K^-})\hat{\pi}^\lambda_{(i),K^-}| + \max_{1 \le i \le \mathring{k}_{\pi,n}} |\hat{\pi}^\lambda_{(i),l_K}|.\end{aligned}$$

Under the null hypothesis $H_0 : \lambda_{j,0} = 0 \; \forall j$, $\hat{\mathcal{T}}_n = O_p(1)$. However, under any alternative with $\psi_{j,K^-,0} = 0 \; \forall j$ and $\pi_{j,0} \neq 0$ for some j Theorem B.4.1 implies that $\hat{\pi}_{(i),l_K,j} \xrightarrow{d} \pi^*_{(i),l_K,j} = O_p(1)$

for every i. Hence under such an alternative, $\hat{\mathcal{T}}_n = O_p(1) \not\to \infty$, so the test is not consistent against these alternatives. □

Proof of Theorem 3.4.5. Recall that if l_K is empty, then $\psi_{K^-} = \theta$ and $\psi_{(i),K^-} = \theta_{(i)}$ for every i. Then Assumption 10 implies that $E_{\gamma_0}(m_t(\psi_{K^-}, \pi_K; W_t))$ is minimized uniquely by $\psi_{K^-} = \psi_{K^-,0} \in \Psi^*_{K^-}$ for every $\pi_K \in \Pi_K$, and Assumption 4 implies that $E_{\gamma_0}(m_t(\psi_{(i),K^-}, \pi_{(i),K}; W_t))$ is minimized uniquely by $\psi_{(i),K^-} = \psi_{(i),K^-,0} \in \Psi^*_{(i),K^-}$ for every $\pi_{(i),K} \in \Pi_{(i),K}$ for every i. Further, the construction of the criterion function implies that

$$\nabla_{\psi_{(i),K^-}} m_{(i),t}(\theta_{(i)}) = \nabla_{\psi_{(i),K^-}} m_t([\theta]_{(i)})$$

for every i, where $[\theta]_{(i)}$ is the restricted full parameter with $\lambda_j = 0$ for every $j \neq i$. By assumption 5, the expectations exist and are finite. Hence if $\lambda_0 = 0_{d_\lambda}$, then $\lambda_{(i),0} = 0$ for every i, and conversely, if $\lambda_{(i),0} = 0$, then $\lambda_0 = 0_{d_\lambda}$. □

REFERENCES

Anderson, H. M. (1997). Transaction costs and nonlinear adjustment towards equilibrium in the us treasury bill market. *Oxford Bulletin of Economics and Statistics 59*, 465–484.

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

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

Andrews, D. W. K. and X. Cheng (2012b, September). Supplemental material for "estimation and inference with weak, semi-strong, and strong identification". *Econometrica 80*(5), 2153–2211.

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

Andrews, D. W. K. and X. Cheng (2014). Gmm estimation and uniform subvector inference with possible identification failure. *Econometic Theory 30*, 287–333.

Andrews, D. W. K., X. Liu, and W. Ploberger (1998, Sep). Tests for white noise against alternatives with both seasonal and nonseasonal serial correlation. *Biometrika 85*(3), 727–740.

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

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

Andrews, D. W. K. and W. Ploberger (1996). Testing for serial correlation against an arma(1,1) process. *Journal of the American Statistical Association 91*, 1331–1342.

Andrews, D. W. K. and J. H. Stock (2007). Testing with many weak instruments. *Journal of Econometrics 24-46*, 138.

Andrews, I. and A. Mikusheva (2015). Maximum likelihood inference in weakly identified dynamic stochastic general equilibrium models. *Quantitative Economics 6*, 123–152.

Andrews, I. and A. Mikusheva (2016, July). Conditional inference with a functional nuisance parameter. *Econometrica 84*(4), 1571–1612.

Antoine, B. and E. Renault (2015). Testing identification strength. *Simon Fraser University Working Paper*.

Bekker, P. (1994). Alternative approximations to the distribution of instrumental variables estimators. *Econometrica 62*, 657–681.

Bekker, P. and F. Kleibergen (2003). Finite-sample instrumental variables inference using an asymptotically pivotal statistic. *Econometric Theory 19*, 744–753.

Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, and K. Kato (2018). High-dimensional econometrics and regularized gmm. *arXiv:1806.01888v2*.

Belloni, A., V. Chernozhukov, and C. Hansen (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives 28*, 29–50.

Belloni, A., V. Chernozhukov, and C. Hansen (2014b). Inference on treatment effects after selection amongst high-dimensional controls. *Review of Economic Studies 81*, 608–650.

Belloni, A., V. Chernozhukov, C. Hansen, and D. Kozbur (2016). Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics 34*, 590–605.

Berman (1964). Limit theorems for the maximum term in stationary sequences. *Annals of Mathematical Statistics 35*, 502–516.

Bound, J., A. Jaeger, and R. Baker (1996). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association 90*, 443–450.

Box, G. E. P. and D. A. Pierce (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association 65*, 1509–1526.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay (1997). *The Econometrics of Financial Markets*. Princeton University Press.

Caner, M. and A. B. Kock (2018). Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative lasso. *Journal of Econometrics 203*, 143–168.

Cassel, G. (1918). Abnormal deviations in international exchanges. *Economic Journal 28*, 413–15.

Chamberlain, G. and G. Imbens (2004). Random effects estimators with many instrumental variables. *Econometrica 72*, 295–306.

Chao, J. and N. Swanson (2005). Consistent estimation with a large number of weak instruments. *Econometrica 73*, 1673–1692.

Chao, J. and N. Swanson (2007). Alternative approximations of the bias and mse of the iv estimator under weak identification with an application to bias correction. *Journal of Econometrics 137*, 515–555.

Chen, H., Y. Fan, and R. Liu (2016). Inference for the correlation coefficient between potential outcomes in the gaussian switching regime model. *Journal of Econometrics 195*, 255–270.

Chen, Y.-T. (2008). A unified approach to standardized-residuals-based correlation tests for garch-type models. *Journal of Applied Econometrics 23*, 111–133.

Cheng, X. (2015). Robust inference in nonlinear models with mixed identification strength. *Journal of Econometrics 189*, 207–228.

Chernozhukov, V. (2005). Extremal quantile regression. *The Annals of Statistics 33*, 806–839.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Annals of Statistics 41*, 2786–2819.

Chernozhukov, V., D. Chetverikov, and K. Kato (2017). Central limit theorems and bootstrap in high dimensions. *Annals of Probability 45*, 2309–2352.

Chernozhukov, V., I. Fernández-val, and T. Kaji (2017). Extremal quantile regression: An overview. *Arxiv Preprint Working Paper*.

Davidson (1994). *Stochastic Limit Theory*. Oxford, U.K: Oxford University Press.

Davidson, R. (2010). Size distortion of bootstrap tests: an example from unit root testing. *Review of Economic Analysis 2*, 169–193.

Davidson, R. and J. G. MacKinnon (1999). The size distrotion of bootstrap tests. *Econometric Theory 15*, 361–376.

Davidson, R. and J. G. MacKinnon (2010). Wild bootstrap tests for iv regression. *Journal of Business & Economic Statistics 28*, 128–144.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika 64*(2), 247–254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternatives. *Biometrika 74*(1), 33–43.

de Haan, L. (1976). Sample extremes: An elementary introduction. *Statistica Neerlandica 30*, 161–172.

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

Dennis, J. (2019). Testing white noise when some parameters may be weakly identified. *UNC Working Paper*.

Dezeure, R., P. Bühlmann, and C.-H. Zhang (2017). High-dimensional simultaneous inference with the bootstrap. *Test 26*, 685–719.

Donohue, J. J. I. and S. D. Levitt (2001). The impact of legalized abortion on crime. *Quarterly Journal of Economics 116(2)*, 379–420.

Donohue, J. J. I. and S. D. Levitt (2008). Measurement error, legalized abortion, and the decline

in crime: A response to foote and goetz. *Quarterly Journal of Economics 123(1)*, 425–440.

Dwyer, G. P., P. Locke, and W. Yu (1996). Index arbitrage and nonlinear dynamics between the s&p 500 futures and cash. *Review of Financial Studies 9*, 301–332.

Eitrheim, O. and T. Teräsvirta (1996). Testing the adequacy of smooth transition autoregressive models. *Journal of Econometrics 74*, 59–76.

Escanciano, J. C. and I. N. Lobato (2009). An automatic portmanteau test for serial correlation. *Journal of Econometrics 151*, 140–149.

Feir, D., T. Lemieux, and V. Marmer (2016). Weak identification in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics 34*(2), 185–196.

Fisher, R. A. and L. H. C. Tippet (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceeding of the Cambridge Philosophical Society 24*, 180–290.

Foote, C. L. and C. F. Goetz (2008). The impact of legalized abortion on crime: Comment. *Quarterly Journal of Economics 123(1)*, 407–423.

Francq, C., L. Horvath, and J.-M. Zakoïan (2010). Sup-tests for linearity in a general nonlinear ar(1) model. *Econometric Theory 26*, 965–993.

Francq, C., R. Roy, and J. M. Zakoian (2005). Diagnostic checking in arma models with uncorrelated errors. *Journal of the American Statistical Association 100*, 532–544.

Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*. Malabar: Krieger.

Ghysels, E. and A. Guay (2004, Dec). Testing for structural change in the presence of auxiliary models. *Econometric Theory 20*(6), 1168–1202.

Ghysels, E., J. B. Hill, and K. Motegi (2016a). Simple granger causality tests for mixed frequency data. *Working Paper, UNC Chapel Hill*.

Ghysels, E., J. B. Hill, and K. Motegi (2016b). Testing for granger causality with mixed frequency data. *Journal of Econometrics 192*, 207–230.

Ghysels, E., J. B. Hill, and K. Motegi (Forthcoming). Testing a large set of zero restrictions in regression models, with an application to mixed frequency granger causality. *Journal of Econometrics*.

Giné, E. and J. Zinn (1990). Bootstrapping gneral empirical measures. *Annals of Probability 18*, 851–869.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics 44*, 423–453.

Granger, C. W. J. and T. Teräsvirta (1993). Modelling nonlinear economic relationships. *Oxford: Oxford University Press*.

Guerron-Quintana, P., A. Inoue, and L. Kilian (2013). Frequentist inference in weakly identified dynamic stochastic general equilibrium models. *Quantitative Economics 4*, 197–229.

Gumbel (1958). *Statistics of Extremes*. Columbia University Press.

Han, S. and A. McCloskey (2016). Estimation and inference with a (nearly) singular jacobian. *Working Paper*.

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

Hansen, B. E. (2011). Threshold autoregression in economics. *Statistics and Its Inference 4*, 123–127.

Hansen, C., J. Hausman, and W. Newey (2012). Estimation with many instrumental variables. *Journal of Business & Economic Statistics 26:4*, 398–422.

Hansen, P. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics 23*, 365–380.

Hausman, J. A., W. K. Newey, T. Woutersen, J. C. Chao, and N. R. Swanson (2012). Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics 3*, 211–255.

Hill, J. and J. Dennis (2018). Testing many zero restrictions where a subset may lie on the boundary. *UNC Working Paper*.

Hill, J. and K. Motegi (2018). A max-correlation white noise test for weakly dependent time series. *UNC Working Paper*.

Hill, J. and K. Motegi (2019). Testing the white noise hypothesis of stock returns. *Economic Modeling 76*, 231–242.

Hill, J. B. (2008). Consistent and non-degenerate model specification tests against smooth transition and neural network alternatives. *Annales d'Économie et de Statistique 90*, 145–179.

Hoffmann-Jorgensen, J. (1984). Convergence of stochastic processes on polish spaces. *mimeo.*.

Hoffmann-Jorgensen, J. (1991). Convergence of stochastic processes on polish spaces. *Various Publications Series, Aarhus Universitet, Aarhus, Denmark.*.

Hong, Y. (1996). Consistent testing for serial correlation of unknown form. *Econometrica 64*, 837–864.

Ibragimov, I. A. (1962). Some limit theorems for stationary processes. *Theory of Probability and*

its Applications VII(4), 349–382.

Kilic, R. (2016). Tests for linearity in star models: Sup wald and lm-type tests. *Journal of Time Series Analysis 37*, 660–674.

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

Leadbetter, M. R., G. Lindgren, and H. Rootzèn (1983). *Extremes and Related Properties of Random Sequences and Processes*. New York: Springer.

Leeb, H. and B. M. Pötscher (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory 24*, 338–376.

Lewbel, A. (Forthcoming). The identification zoo - meanings of identification in econometrics. *Journal of Economic Literature*.

Liu, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *Annals of Statistics 16*, 1696–1708.

Ljung, G. M. and G. E. P. Box (1978). On a measure of a lack of fit in time series models. *Biometrika 65*(2), 297–303.

Lundbergh, S., T. Teräsvirta, and D. van Dijk (2000). Time-varying smooth transition autoregressive models. *Working Paper Series in Economics and Finance No 376, Stockholm School of Economics*.

Martens, M., P. Kofman, and A. C. F. Vorst (1998). A threshold error correction for intraday futures and index returns. *Journal of Applied Econometrics 13*, 245–263.

McLeish, D. L. (1975). A maximal inequality and dependent strong laws. *Annals of Probability 3*, 829–839.

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

Moreira, M. J., J. R. Porter, and G. A. Suarez (2009). Bootstrap validity for the score test when instruments may be weak. *Journal of Econometrics 149*, 52–64.

Nankervis, J. C. and N. E. Savin (2010). Testing for serial correlation: Generalized andrews-ploberger tests. *Journal of Business and Economic Statistics 28*, 246–255.

Nankervis, J. C. and N. E. Savin (2012). Testing for uncorrelated errors in arma models: Non-standard andrews-ploberger tests. *Econometrics Journal 15*, 516–534.

Newey (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica 59*, 1161–1167.

Obstfeld, M. and M. Taylor (1997). Nonlinear aspcects of goods-market arbitrage and adjustment: Heckscher's commodity points revisited. *Journal of the Japanese and International Economies 11*, 441–479.

Pollard, D. (1990). Empirical processes: Theory and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, Volume II. Institute of Mathematical Statistics and the American Statistical Association.

Poterba, J. M. and L. H. Summers (1988). Mean reversion in stock prices: Evidence and implications. *Journal of Financial Economics 22*, 27–59.

Pötscher, B. M. (2009). Confidence sets based on sparse estimators are necessarily large. *Sankhyā: The Indian Journal of Statistics, Series A 71*, 1–18.

Ramsey, F. P. (1929). On a problem of formal logic. *Proceedings of the London Mathematical Society 30*, 264–286.

Resnick, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes*. New York: Springer.

Rogoff, K. (1996). The purchasing power parity puzzle. *Journal of Economic Literature 34*, 647–668.

Romano, J. and L. Thombs (1996). Inference for autocorrelations under weak assumptions. *Journal of the American Statistical Association 91*, 590–600.

Rothman, P., D. van Dijk, and P. H. Franses (2001). A multivariate star analysis of the relationship between money and output. *Macroeconomic Dynamics*.

Schwert, G. W. (1989). Testing for unit roots: a monte carlo investigation. *Journal of Business and Economic Statistics 7*, 147–159.

Shao, X. (2010). The dependent wild bootstrap. *Journal of the American Statistical Association 105*(489), 218–235.

Shao, X. (2011a). A bootstrap-assisted spectral test of white noise under unknown dependence. *Journal of Econometrics 162*, 213–224.

Shao, X. (2011b, April). Testing for white noise under unknown dependence and its applications to diagnostic checking for time series models. *Econometric Theory 27*(2), 312–343.

Skalin, J. and T. Teräsvirta (2001). Modeling asymmetries and moving equilibria in unemployment rates. *Macroeconomic Dynamics*.

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

Stinchcombe, M. B. and H. White (1998). Consistent specification testing with nuisance parameters present only under the alternative. *Econometric Theory 14*, 295–325.

Swanson, N. R. (1999). Finite sample properties of a simple lm test for neglected non-linearity in error-correcting regression equations. *Statistica Neerlandica 53*, 76–95.

Taylor, M. P., D. A. Peel, and L. Sarno (2001). Nonlinear mean-reversion in real exchange rates: towards a solution to the purchasing power parity puzzles. *International Economic Review 42*, 1015–1042.

Taylor, N., D. van Dijk, P. H. Franses, and A. Lucas (2000). Sets, arbitrage activity, and stock price dynamics. *Journal of Banking and Finance 24*, 1289–1306.

Taylor, S. (2005). *Asset Price Dynamics, Volatility, and Prediction*. Princeton, NJ: Princeton University Press.

Terasvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association 89*, 208–218.

Teräsvirta, T. (1998). Modelling economic relationships with smooth transition regressions. In *Handbook of Applied Economic Statistics*, pp. 507–552. New York: Marcel Dekker.

Teräsvirta, T. (2004). Smooth transition regression modeling,. *In Applied Time Series Econometrics. Lutkepohl H, Kratzig M (eds.) Cambridge University Press: Cambridge*, 222–242.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statisticl Society Series B 58*, 267–288.

Tripathi, G. (1999). A matrix extension of the cauchy-schwarz inequality. *Economic Letters 63*, 1–3.

Tsay, R. S. (1998). Testing and modeling multivariate threshold models. *Journal of the American Statistical Association 93*, 1188–1202.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*, 1166–1202.

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

van Dijk, D. and P. H. Franses (1999). Modeling multiple regimes in the business cycle. *Macroeconomic Dynamics 3*, 311–340.

van Dijk, D. and P. H. Franses (2000). Nonlinear error-correction models for interest rates in the netherlands. In W. A. Barnett, D. F. Hendry, S. Hylleberg, T. Teräsvirta, D. Tjostheim, and

A. H. Würtz (Eds.), *Nonlinear Econometric Modeling in Time Series Analysis*, pp. 203–227. Cambridge: Cambridge University Press.

van Dijk, D., B. Strikholm, and T. Teräsvirta (2001). The effects of institutional and technological change and business cycle fluctuations on seasonal patterns in quarterly industrial production series. *Working Paper Series in Economics and Finance No. 429, Stockholm School of Economics*.

van Dijk, D., T. Teräsvirta, and P. H. Franses (2002). Smooth transition autoregressive models - a survey of recent developments. *Econometric Reviews 21*, 1–47.

White, H. (1981). Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association 76*, 419–433.

Wooldridge, J. M. and Y. Zhu (Forthcoming). Inference in approximately sparse correlated random effects probit models. *Journal of Business and Economic Statistics*.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics 14*, 1261–1350.

Xiao, H. and W. B. Wu (2014). Portmanteau test and simultaneous inference for serial covariances. *Statistica Sinica 24*, 577–599.

Zhang, D. and W. B. Wu (2017). Gaussian approximation for high dimensional time series. *Annals of Statistics 45*, 1895–1919.

Zhang, X. (2016). White noise testing and model diagnostic checking for functional time series. *Journal of Econometrics 194*, 76–95.

Zhang, X. and G. Cheng (2018). Gaussian approximation for high dimensional vector under physical dependence. *Bernoulli 24*, 2640–2675.

Zhu, K. and W. Li (2015). A bootstrapped spectral test for adequacy in weak arma models. *Journal of Econometrics 187*, 113–130.