

# Testing Correct Model Specification Using Extreme Learning Machines\*

JIN SEO CHO

School of Economics

Yonsei University

Email: jinseocho@yonsei.ac.kr

HALBERT WHITE

Department of Economics

University of California, San Diego

Email: hwhite@weber.ucsd.edu

First version: May, 2010. This version: September, 2010.

## Abstract

Testing the correct model specification hypothesis for artificial neural network (ANN) models of the conditional mean is not standard. The traditional Wald, Lagrange multiplier, and quasi-likelihood ratio statistics weakly converge to functions of Gaussian processes, rather than to convenient chi-squared distributions. Also, their large sample null distributions are problem dependent, limiting applicability. We overcome this challenge by applying functional regression methods of Cho, Huang, and White (2008) to extreme learning machines (ELM). The Wald ELM (WELM) test statistic proposed here is easy to compute and has a large-sample standard chi-squared distribution under the null hypothesis of correct specification. We provide associated theory for time-series data and affirm our theory with some Monte Carlo experiments.

**Key Words** Artificial neural networks; Gaussian process; Functional regression; Extreme learning machines.

---

\***Acknowledgements** Cho acknowledges research support from the Korea Sanhak foundation.

# 1 Introduction

Artificial neural networks (ANNs) are extensively used to approximate stochastic relationships of unknown form. Their appeal is based on their universal approximation properties (see, e.g., Hornik, Stinchcombe, and White [21, 22], and Stinchcombe and White [32]). Nevertheless, ANNs are often difficult to apply, as they may require estimating a large number of unknown parameters (network weights). Consequently, ANNs may suffer from overfitting, and their predictive power can be poor for this reason.

This limitation motivates the search for parsimonious ANN models. For this, one can first train a small ANN network and then train a larger network. One can then test whether or not the fit of the larger network improves to a statistically significant degree. If it does, one has evidence that the parsimonious model suffers from misspecification, in the sense that the smaller model's errors contain approximation error as well as pure prediction error. If not, one has evidence that the larger network is unnecessary, as the smaller network's errors are mostly pure prediction error.

There are many test statistics in the literature that can be applied for this purpose; their variety is due to the fact that this is a nonstandard testing problem. Specifically, there are unidentified parameters under the null of correct model specification (see Davies, [12, 13]); this causes standard test statistics to behave in nonstandard ways. For example, the Wald test examined by Bierens [5] and Hansen [19] for cross-section and time-series data, respectively, does not follow a standard chi-squared distribution under the null of correct specification. Instead, it weakly converges to a function of a Gaussian process, and its limiting null distribution is problem dependent. As another example, the quasi-likelihood ratio (QLR) test statistic designed to overcome the twofold identification problem pointed out by Cho, Ishida, and White [10] also does not follow a standard distribution under the null. As White [36] and Kuan and White [24] note, it can be a challenging task to construct test statistics in such a way that they follow standard distributions under the null and, at the same time, have non-negligible power to detect misspecification.

The goal of the current study is motivated by this observation. We seek a statistical test for the correct model specification hypothesis, whose application is straightforward and that has a standard asymptotic null distribution. This test can then be used to help construct parsimonious ANN models.

We achieve our goal by combining the theory of functional regression with that of extreme learning machines (ELM). Cho, Huang, and White [9] study functional

regression, in which functional data are regressed against known deterministic functions. These authors develop a statistic to test whether or not the population mean of the functional data is a constant function. Conveniently, the statistic follows a standard limiting chi-squared distribution under the null. Nevertheless, the required integrations make computing the test statistic quite demanding. This limits applicability, except when the observed functional data have a fairly simple structure. Here, we avoid this difficulty by applying extreme learning machines (ELM) to generate functional data for the test. Applying ELM methods proposed by Huang, Zhu, and Siew [23] and by White [43] (“QuickNet”) indeed enables our Wald ELM (WELM) test to be computed very conveniently.

The plan of this paper is as follows. We first introduce and heuristically motivate the WELM test in Sections 2 and 3. The WELM test is analyzed in Section 4 under a set of formal regularity conditions. There, we examine large sample properties of the WELM statistic under the null and alternative hypotheses. We report the results of some Monte Carlo experiments in Section 5; mathematical proofs are gathered into the Appendix.

## 2 The Data Generating Process (DGP) and ANN Models

We suppose the data to be analyzed are weakly dependent time-series data:

**Assumption A1 (DGP):** Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a complete probability space, and let  $k \in \mathbb{N}$ . Let  $\{(Y_t, \mathbf{X}_t)' : \Omega \rightarrow \mathbb{R}^{1+k} : t = 1, 2, \dots\}$  be a strictly stationary and absolutely regular process with mixing coefficients  $\beta_\tau$  such that for some  $\rho > 1$ ,  $\sum_{\tau=1}^{\infty} \tau^{2\rho/(\rho-1)} \beta_\tau < \infty$ .

Here,  $Y_t$  and  $\mathbf{X}_t$  are serially dependent target and explanatory variables, respectively. For convenience, we permit  $\mathbf{X}_t$  to include a constant element. It may also include  $Y_{t-1}, Y_{t-2}$ , etc. The mixing coefficients  $\beta_\tau$  measure the dependence of the stochastic process as

$$\beta_\tau \equiv \sup_{s \in \mathbb{N}} E \left[ \sup_{A \in \mathcal{F}_{s+\tau}^\infty} |P(A | \mathcal{F}_{-\infty}^s) - P(A)| \right],$$

where  $\mathcal{F}_s^t$  is the  $\sigma$ -field (“information set”) generated by  $(Y_s, \mathbf{X}_s'), \dots, (Y_t, \mathbf{X}_t)'$ , and Assumption 1 implies that  $\beta_\tau$  is size of  $-2\rho/(\rho-1)$ . That is, for some  $\varepsilon > 0$ ,  $\beta_\tau = O(\tau^{-2\rho/(\rho-1)-\varepsilon})$ . The dependence allowed here is less than that assumed in

the previous literature, such as Hansen [19, 20], Cho and White [8], Cho, Huang, and White [9], Cho, Ishida, and White [10]. On the other hand, this stronger condition enables us not only to consistently estimate the asymptotic covariance matrix of our estimator, despite the dependence, but also to ensure applicability of the functional central limit (FCLT) theorem, both of which are needed in deriving the asymptotic null distribution of our test statistic.

We suppose that the researcher's interest lies in estimating  $E[Y_t|\mathbf{X}_t]$ , the mean-squared-error optimal predictor of  $Y_t$  given  $\mathbf{X}_t$ . Suppose further that the researcher approximates  $E[Y_t|\mathbf{X}_t]$  using a function  $\Phi$  as

$$E[Y_t|\mathbf{X}_t] \approx \Phi(\mathbf{X}_t, \boldsymbol{\theta}_*)$$

where  $\boldsymbol{\theta}_*$  is a suitably chosen parameter vector. For example,  $\Phi(\mathbf{X}_t, \boldsymbol{\theta}_*)$  could be the output of a hidden layer feedforward network (e.g., as in Kuan and White [24]) with network weights  $\boldsymbol{\theta}_*$ . We call  $\Phi$  the researcher's "specification" for  $E[Y_t|\mathbf{X}_t]$ . The question we consider here is whether for some  $\boldsymbol{\theta}_*$  the approximation is exact ("correct specification"), so there is no possible way to improve the prediction, or if the approximation using  $\Phi$  is not perfect ("misspecification"), implying that prediction improvements are possible. If the former is true, we can avoid suboptimal predictions resulting from unnecessarily elaborating the prediction method. If the latter, we can avoid suboptimal predictions resulting from an overly simple prediction method.

To address this issue we impose the following assumption.

**Assumption A2 (Model):** Let  $\Theta \subset \mathbb{R}^d$ ,  $d \in \mathbb{N}$ , be a non-empty compact convex set and suppose that for each  $\boldsymbol{\theta} \in \Theta$ ,  $\Phi(\cdot, \boldsymbol{\theta}) : \mathbb{R}^k \mapsto \mathbb{R}$  is a measurable function such that for each  $\omega \in F \in \mathcal{F}$  with  $\mathbb{P}(F) = 1$ ,  $\Phi(\mathbf{X}_t(\omega), \cdot) : \Theta \mapsto \mathbb{R}$  is twice continuously differentiable. Let  $\Lambda \subset \mathbb{R}$  and  $\Delta \subset \mathbb{R}^k$  be non-empty compact convex sets with  $0 \in \text{int}(\Lambda)$ , and suppose that  $\Psi : \mathbb{R} \mapsto \mathbb{R}$  is a non-polynomial analytic function.

We can now formally state the null hypothesis of correct specification of  $\Phi$  and the alternative of misspecification as

$$\mathcal{H}_0 : \text{For some } \boldsymbol{\theta}_* \in \Theta, \mathbb{P}[E(Y_t|\mathbf{X}_t) = \Phi(\mathbf{X}_t, \boldsymbol{\theta}_*)] = 1 \text{ versus}$$

$$\mathcal{H}_1 : \text{For all } \boldsymbol{\theta} \in \Theta, \mathbb{P}[E(Y_t|\mathbf{X}_t) = \Phi(\mathbf{X}_t, \boldsymbol{\theta})] < 1.$$

The *augmented* specification  $f(\mathbf{X}_t; \boldsymbol{\theta}, \lambda, \boldsymbol{\delta}) := \Phi(\mathbf{X}_t, \boldsymbol{\theta}) + \lambda\Psi(\mathbf{X}_t'\boldsymbol{\delta})$  is the original specification,  $\Phi$ , augmented by the contribution from an additional hidden

unit with activation function  $\Psi$ , input-to-hidden weights  $\delta$ , and hidden-to-output weight  $\lambda$ . The augmented specification generates the (augmented) *model*  $\mathcal{M}$  as

$$\mathcal{M} := \{f(\cdot; \boldsymbol{\theta}, \lambda, \boldsymbol{\delta}) : (\boldsymbol{\theta}, \lambda, \boldsymbol{\delta}) \in \Theta \times \Lambda \times \Delta\}.$$

Thus, the *null* model  $\mathcal{M}_0 := \{\Phi(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$  obtains when  $\lambda = 0$  and is nested in  $\mathcal{M}$ . When  $\Phi$  is correctly specified,  $E[Y_t | \mathbf{X}_t] \in \mathcal{M}_0$ , that is,  $\mathcal{H}_0$  holds. Otherwise,  $\mathcal{H}_1$  holds. The advantage to specifying  $\mathcal{M}$  as we have is that, as Stinchcombe and White [32] show, if  $\Psi$  is *generically comprehensively revealing* (GCR), then whenever  $\mathcal{H}_1$  holds, the augmented specification  $f$  is guaranteed to provide a better prediction than the original specification  $\Phi$ , revealing the presence of arbitrary misspecification of  $\Phi$ . That is, we just need to check whether adding a single suitably chosen (i.e., GCR) hidden unit to the original specification can improve prediction performance. By assuming that  $\Psi$  is non-polynomial analytic, we ensure that it is GCR.

There are many admissible choices for  $\Psi$ , and each has its own merits. For example, White [36] considers the logistic cumulative distribution function (CDF) for  $\Psi$ ; Bierens [5] examines the correct model specification assumption by letting  $\Psi = \exp$ ; and Candés [7] analyzes ridgelet functions. In addition, the familiar trigonometric functions will also work.

Note also that  $\Phi$  is only mildly restricted. It can be any feedforward network with any finite number of hidden units and weights and any sufficiently smooth activation functions. A particularly simple but important case is that of a linear input-output network,  $\Phi(\mathbf{X}_t, \boldsymbol{\theta}) = \mathbf{X}_t' \boldsymbol{\theta}$ , as this form is widely used for making predictions. White [36] and Lee, Granger, and White [25], among others, explicitly test the linearity hypothesis using various analytic functions  $\Psi$ .

The literature provides many testing procedures suited to our present goal. For example, the goodness-of-fit test examined by Delgado and Stute [14] can be used to test  $\mathcal{H}_0$ . Nevertheless, to maintain a tight focus in what follows, we limit attention to tests utilizing the universal approximation feature of ANNs.

First, Wald and Lagrange multiplier (LM) test statistics are specifically examined by Bierens [5] and Hansen [19]. Their approach mainly focuses on the coefficient  $\lambda$ . The null model  $\mathcal{M}_0$  is generated if  $\lambda_*$  is zero, where  $\lambda_*$  is the probability limit of the nonlinear least squares (NLS) estimator, say  $\hat{\lambda}_n$ . Thus, a diagnostic test for correct specification can be constructed using the standard Wald or LM statistics, testing  $\lambda_* = 0$ . But when  $\lambda_* = 0$ , the associated optimal prediction parameter  $\boldsymbol{\delta}_*$  is not identified, where  $\boldsymbol{\delta}_*$  is the probability limit of the NLS estimator, say  $\hat{\boldsymbol{\delta}}_n$ . The so-called Davies problem [12, 13] of nuisance parameters

identified only under the alternative hypothesis is present, and the asymptotic null distributions of the test statistics are different from the standard chi-squared distribution. Accordingly, a great deal of effort has focused on determining these test statistics' asymptotic null distribution. In most cases, this is a function of a Gaussian process defined on  $\Delta$ , but the asymptotic null distribution is different for each model  $\mathcal{M}$ . This is also true of Neyman's [27, 29, 30]  $C(\alpha)$  test.

Second, White [36] and Lee, Granger, and White [25] take a different approach. They avoid the problem of nuisance parameters identified only under the alternative by randomly selecting input-to-hidden weights  $\delta_j$ ,  $j = 1, \dots, p$ , and estimating the hidden-to-output weights by simple least-squares methods. This random selection and estimation approach is exactly that known as extreme learning machines (ELM) in the analysis of Huang, Zhu, and Siew [23] and also proposed by White [43], based on universal approximation results in Stinchcombe and White [32]. The estimated hidden-to-output weights have a limiting normal distribution, with mean zero under the null. This property can be used to construct a Wald or LM statistic following a chi-squared distribution, yielding a test for correct specification that is no longer model dependent. Nevertheless, the random selection of a finite number of weights makes it possible that two researchers using the same data could arrive at different conclusions due to different weight selections.

Third, Cho, Ishida, and White [10] consider a quasi-likelihood ratio (QLR) statistic, pointing out a previously overlooked issue that they call the twofold identification problem. Specifically, they note that nuisance parameters not identified under the alternative arise not only when  $\lambda_*$  is zero, but also when  $\delta_* = \mathbf{0}$ . In the latter case, we have  $f(\mathbf{X}_t; \boldsymbol{\theta}_*, \lambda_*, \delta_*) = \Phi(\mathbf{X}_t, \boldsymbol{\theta}_*) + \lambda_* \Psi(0)$ ; and the second component on the right-hand side (RHS) can form part of  $\Phi(\mathbf{X}_t, \boldsymbol{\theta}_*)$ , provided  $\Phi(\mathbf{X}_t, \boldsymbol{\theta}_*)$  contains an intercept (bias) term, as it usually does. In this case, the asymptotic null distribution of the QLR test can be obtained by carefully accounting for the interrelationship of the weak limits obtained under each hypothesis:  $\lambda_* = 0$  and  $\delta_* = \mathbf{0}$ . Cho, Ishida, White [10] provide asymptotic distribution results for QLR under new conditions, stronger than those used by Hansen [19]. Their stronger conditions are needed to approximate the QLR statistic with a fourth-order expansion, similar to the approach taken in related contexts by Bartlett [3, 4] and McCullagh [26]. Unfortunately, however, the QLR statistic is also model dependent, and constructing its asymptotic distribution is computationally intensive, involving use of the weighted bootstrap proposed by Giné and Zinn [17] and Hansen [19].

Recently, Cho, Huang, and White [9] have proposed another approach to test-

ing correct model specification. This is based on the fact that under the null of no misspecification, the residuals obtained from the null model must be orthogonal to any measurable function of explanatory variables  $\mathbf{X}_t$ . This is true in particular for  $\Psi(\mathbf{X}'_t\boldsymbol{\delta})$ , so that if we let  $U_t \equiv Y_t - \Phi(\mathbf{X}_t, \boldsymbol{\theta}_*)$  be the residual derived from the null model, then it follows that for every  $\boldsymbol{\delta}$ ,  $E[U_t\Psi(\mathbf{X}'_t\boldsymbol{\delta})] = 0$  under the null. Cho, Huang, and White [9] propose regressing  $U_t\Psi(\mathbf{X}'_t\boldsymbol{\delta})$  on a vector of deterministic functions of  $\boldsymbol{\delta}$ , say  $\mathbf{g}(\cdot)$ , and testing whether or not the coefficients of  $\mathbf{g}(\cdot)$  are equal to zero using a Wald-type test statistic. As this is a regression of  $U_t\Psi(\mathbf{X}'_t\boldsymbol{\delta})$  on  $\mathbf{g}(\boldsymbol{\delta})$  for every  $\boldsymbol{\delta}$ , they call this *functional regression*. The Wald-type test is obtained by integrating out the functional coefficients of  $\boldsymbol{\delta}$  using an arbitrarily chosen probability measure. Cho, Huang, and White show that their Wald-type statistic converges to a standard chi-squared random variable under the null, whereas it is not bounded in probability under the alternative. The test is therefore not model dependent and has power to detect arbitrary misspecification. But this can be computationally demanding, as computing integrals of random functions is a time-consuming process.

These considerations motivate us to devise another approach that overcomes the limitations of previous methods. Specifically, we combine the functional regression of Cho, Huang, and White [9] with ELM, which replaces computationally demanding methods with simple and straightforward procedures that yield tests consistent against arbitrary misspecification under mild conditions.

### 3 A Heuristic Introduction to the WELM Test

We provide a heuristic motivation for our test statistic by first supposing that the parameter  $\boldsymbol{\theta}_*$  is known or given. In the next section, we provide a formal result supposing more realistically that  $\boldsymbol{\theta}_*$  is unknown.

Let  $\boldsymbol{\theta}_*$  be the mean-squared-error optimal prediction parameter,

$$\boldsymbol{\theta}_* \equiv \arg \min_{\boldsymbol{\theta} \in \Theta} E[(E[Y_t|\mathbf{X}_t] - \Phi(\mathbf{X}_t, \boldsymbol{\theta}))^2].$$

It is readily verified that  $\boldsymbol{\theta}_*$  also solves

$$\min_{\boldsymbol{\theta} \in \Theta} E[(Y_t - \Phi(\mathbf{X}_t, \boldsymbol{\theta}))^2],$$

and that, under mild conditions,  $\boldsymbol{\theta}_*$  is the probability limit of the NLS estimator,

$$\hat{\boldsymbol{\theta}}_n \equiv \arg \min_{\boldsymbol{\theta} \in \Theta} n^{-1} \sum_{t=1}^n (Y_t - \Phi(\mathbf{X}_t, \boldsymbol{\theta}))^2.$$

Letting the prediction error be  $U_t := Y_t - \Phi(\mathbf{X}_t, \boldsymbol{\theta}_*)$ , we may decompose it into the sum of the specification error and a (pure) forecasting error. That is,

$$Y_t \equiv \Phi(\mathbf{X}_t, \boldsymbol{\theta}_*) + \underbrace{E[Y_t|\mathbf{X}_t] - \Phi(\mathbf{X}_t, \boldsymbol{\theta}_*)}_{\text{Specification Error}} + \underbrace{Y_t - E[Y_t|\mathbf{X}_t]}_{\text{Forecasting Error}}$$

$U_t := \text{Prediction Error}$

The behavior of  $U_t$  is different under  $\mathcal{H}_0$  and  $\mathcal{H}_1$ . Specifically, if  $\mathcal{H}_0$  holds, then  $E[Y_t|\mathbf{X}_t] = \Phi(\mathbf{X}_t, \boldsymbol{\theta}_*)$  with probability one, so  $U_t = Y_t - E[Y_t|\mathbf{X}_t]$  and for any measurable function  $\Psi$ ,  $E[U_t\Psi(\mathbf{X}'_t\boldsymbol{\delta})] = 0$  for every  $\boldsymbol{\delta}$ . On the other hand, under  $\mathcal{H}_1$ ,  $U_t = E[Y_t|\mathbf{X}_t] - \Phi(\mathbf{X}_t, \boldsymbol{\theta}_*) + Y_t - E[Y_t|\mathbf{X}_t]$ , so that

$$E[U_t\Psi(\mathbf{X}'_t\boldsymbol{\delta})] = E[(E[Y_t|\mathbf{X}_t] - \Phi(\mathbf{X}_t, \boldsymbol{\theta}_*))\Psi(\mathbf{X}'_t\boldsymbol{\delta})], \quad (1)$$

is not necessarily equal to zero for every  $\boldsymbol{\delta}$ . Indeed, Bierens [5] shows that when  $\Psi = \exp$ , the  $\boldsymbol{\delta}$ 's for which  $E[U_t\Psi(\mathbf{X}'_t\boldsymbol{\delta})] = 0$  form a set of Lebesgue measure zero that is not dense in  $\Delta$  under the alternative. Stinchcombe and White [32] generalize this result by showing that it remains true for any GCR function  $\Psi$ , such as nonpolynomial analytic functions.

This property provides the basis for a diagnostic test statistic. That is, if  $E[U_t\Psi(\mathbf{X}'_t\boldsymbol{\delta})] \neq 0$  for some  $\boldsymbol{\delta}$ , this is evidence that  $\Phi$  is misspecified, whereas we cannot reject correct specification when  $E[U_t\Psi(\mathbf{X}'_t\boldsymbol{\delta})] = 0$  for all  $\boldsymbol{\delta}$ . Accordingly, we construct a test statistic by regressing  $\Psi(\mathbf{X}'_t\boldsymbol{\delta})$  on the constant 1 and  $U_t := Y_t - \Phi(\mathbf{X}_t, \boldsymbol{\theta}_*)$  for each  $\boldsymbol{\delta}$ . The estimated regression coefficients are

$$\begin{bmatrix} \widehat{\alpha}_n(\boldsymbol{\delta}) \\ \widehat{\beta}_n(\boldsymbol{\delta}) \end{bmatrix} := \begin{bmatrix} n & \sum_{t=1}^n U_t \\ \sum_{t=1}^n U_t & \sum_{t=1}^n U_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^n \Psi(\mathbf{X}'_t\boldsymbol{\delta}) \\ \sum_{t=1}^n U_t \Psi(\mathbf{X}'_t\boldsymbol{\delta}) \end{bmatrix} \quad (2)$$

for each  $\boldsymbol{\delta}$ . Following Cho, Huang, and White [9], we call this procedure *functional* regression, as the estimated coefficients are functions of  $\boldsymbol{\delta}$ .

For each  $\boldsymbol{\delta}$ ,

$$\begin{bmatrix} \widehat{\alpha}_n(\boldsymbol{\delta}) \\ \widehat{\beta}_n(\boldsymbol{\delta}) \end{bmatrix} \rightarrow \begin{bmatrix} \alpha_*(\boldsymbol{\delta}) \\ \beta_*(\boldsymbol{\delta}) \end{bmatrix} := \begin{bmatrix} 1 & B_* \\ B_* & C_* \end{bmatrix}^{-1} \begin{bmatrix} D_*(\boldsymbol{\delta}) \\ F_*(\boldsymbol{\delta}) \end{bmatrix} \quad a.s. - \mathbb{P}$$

under mild regularity conditions, where

$$B_* := E[Y_t - \Phi(\mathbf{X}_t, \boldsymbol{\theta}_*)]$$

$$C_* := E[(Y_t - E[Y_t|\mathbf{X}_t])^2] + E[(E[Y_t|\mathbf{X}_t] - \Phi(\mathbf{X}_t, \boldsymbol{\theta}_*))^2]$$

$$D_*(\boldsymbol{\delta}) := E[\Psi(\mathbf{X}'_t \boldsymbol{\delta})], \quad \text{and}$$

$$F_*(\boldsymbol{\delta}) := E[(E[Y_t | \mathbf{X}_t] - \Phi(\mathbf{X}_t, \boldsymbol{\theta}_*)) \Psi(\mathbf{X}'_t \boldsymbol{\delta})].$$

When  $\Phi(\mathbf{X}_t, \boldsymbol{\theta}_*)$  is an unbiased predictor for  $Y_t$ , then  $B_* = 0$  for both  $\mathcal{H}_0$  and  $\mathcal{H}_1$ . Under  $\mathcal{H}_0$ ,  $F_*(\boldsymbol{\delta}) = 0$  for each  $\boldsymbol{\delta}$ , so that  $\widehat{\beta}_n(\boldsymbol{\delta})$  converges to 0 *a.s.*  $-\mathbb{P}$ . Under  $\mathcal{H}_1$ , however, the almost sure limit of  $\widehat{\beta}_n(\boldsymbol{\delta})$  differs from 0 for almost every  $\boldsymbol{\delta}$ , by the GCR property of  $\Psi$ .

We exploit these different limits for  $\widehat{\beta}_n(\boldsymbol{\delta})$  to construct a specification test statistic. For this, we first obtain the asymptotic distribution of  $\widehat{\beta}_n(\cdot)$ , using the Functional Central Limit Theorem (FCLT):

$$\sqrt{n} \begin{bmatrix} \widehat{\alpha}_n - \alpha_* \\ \widehat{\beta}_n - \beta_* \end{bmatrix} \Rightarrow \mathcal{G} := \begin{bmatrix} \mathcal{G}_1 \\ \mathcal{G}_2 \end{bmatrix}. \quad (3)$$

This weak convergence holds under mild conditions given below, where  $\mathcal{G}_1$  and  $\mathcal{G}_2$  are mean zero Gaussian stochastic processes with

$$\begin{aligned} E[\mathcal{G}(\boldsymbol{\delta}_1) \mathcal{G}(\boldsymbol{\delta}_2)'] &= \mathbf{r}(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) := \begin{bmatrix} r_1(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) & r_2(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) \\ r_2(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) & r_3(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) \end{bmatrix} \\ &:= \begin{bmatrix} 1 & B_* \\ B_* & C_* \end{bmatrix}^{-1} \begin{bmatrix} \kappa_1(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) & \kappa_2(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) \\ \kappa_2(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) & \kappa_3(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) \end{bmatrix} \begin{bmatrix} 1 & B_* \\ B_* & C_* \end{bmatrix}^{-1} \end{aligned}$$

for each  $\boldsymbol{\delta}_1$  and  $\boldsymbol{\delta}_2$  in  $\Delta$ , where

$$\begin{aligned} \kappa_1(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) &:= \text{acov} \left[ n^{-1/2} \sum \Psi(\mathbf{X}'_t \boldsymbol{\delta}_1), n^{-1/2} \sum \Psi(\mathbf{X}'_t \boldsymbol{\delta}_2) \right], \\ \kappa_2(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) &:= \text{acov} \left[ n^{-1/2} \sum \Psi(\mathbf{X}'_t \boldsymbol{\delta}_1), n^{-1/2} \sum \Psi(\mathbf{X}'_t \boldsymbol{\delta}_2) U_t \right], \\ \kappa_3(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) &:= \text{acov} \left[ n^{-1/2} \sum \Psi(\mathbf{X}'_t \boldsymbol{\delta}_1) U_t, n^{-1/2} \sum \Psi(\mathbf{X}'_t \boldsymbol{\delta}_2) U_t \right], \end{aligned}$$

respectively; and  $\text{acov}[\cdot, \cdot]$  denotes the asymptotic covariance between the indicated variables.

Here, we can apply the FCLT for absolutely regular processes given in Doukhan, Maasart, and Rio [15], showing the tightness of  $(\widehat{\alpha}_n, \widehat{\beta}_n)'$  as in Billingsley [6] and van der Vaart [33]. In particular, for every  $\boldsymbol{\delta}$ ,  $\sqrt{n} \widehat{\beta}_n(\boldsymbol{\delta})$  is asymptotically bounded in probability under  $\mathcal{H}_0$ , whereas it is not under  $\mathcal{H}_1$ . Using this simple fact, one can define a variety of statistics that test this particular property.

Here, we seek a procedure that is not model dependent. Note that the given covariance kernel  $\mathbf{r}$  depends on the joint distribution of  $(\mathbf{X}_t, U_t)$ , and the particular

form of  $\Psi$ , implying the model dependence of  $\mathcal{G}$ . Cho, Huang, and White [9] therefore advocate integrating over the associated random functions, which can yield a test statistic whose asymptotic distribution is not model dependent.

Following this advice, we integrate  $\sqrt{n}[\widehat{\beta}_n(\cdot) - \beta_*(\cdot)]$  over  $\Delta$ . This gives

$$\int_{\Delta} \sqrt{n}[\widehat{\beta}_n(\boldsymbol{\delta}) - \beta_*(\boldsymbol{\delta})] d\mathbb{Q}(\boldsymbol{\delta}) \Rightarrow N(0, \tau_*^2),$$

where

$$\tau_*^2 := \int_{\Delta} \int_{\Delta} r_3(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) d\mathbb{Q}(\boldsymbol{\delta}_1)d\mathbb{Q}(\boldsymbol{\delta}_2),$$

and  $\mathbb{Q}(\cdot)$  is a probability measure on  $\Delta$ , specified by the researcher to focus power in particular directions of concern. This permits construction of a Wald [34] statistic,

$$\mathcal{W}_n := \frac{n}{\widehat{\tau}_n^2} \left[ \int_{\Delta} \widehat{\beta}_n(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) \right]^2,$$

where  $\widehat{\tau}_n^2$  is a consistent estimator for  $\tau_*^2$ . Under  $\mathcal{H}_0$ ,  $\mathcal{W}_n \Rightarrow \chi_1^2$ , whereas  $\mathcal{W}_n$  is not bounded in probability under  $\mathcal{H}_1$ .

Although  $\mathcal{W}_n$  is conceptually straightforward, computing  $\mathcal{W}_n$  can require immense computational resources. To see why, note that computing  $\int_{\Delta} \widehat{\beta}_n(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta})$  requires first computing  $\int_{\Delta} \Psi(\mathbf{X}'_t \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta})$  for every  $t$ . That is,  $n$  integrals of random functions have to be computed with respect to  $\boldsymbol{\delta}$ . When  $n$  is even modestly large, this can be extremely time consuming.

We can reduce this computational burden using ELM. Specifically, we treat  $\boldsymbol{\delta}$  as a random variable distributed according to  $\mathbb{Q}$  on  $\Delta$ , so that  $\mathcal{W}_n$  can be also expressed as

$$\mathcal{W}_n = \frac{n}{\widehat{\tau}_n^2} \{E_{\mathbb{Q}}[\widehat{\beta}_n(\boldsymbol{\delta})]\}^2,$$

where  $E_{\mathbb{Q}}[\cdot]$  denotes the expectation under  $\mathbb{Q}(\cdot)$ . This also implies that the integral defining  $\mathcal{W}_n$  can be approximated by applying the law of large numbers (LLN). That is, when  $\{\boldsymbol{\delta}_j\}_{j=1}^m$  is a sequence of IID random variables following  $\mathbb{Q}(\cdot)$ , the LLN yields that

$$m^{-1} \sum_{j=1}^m \widehat{\beta}_n(\boldsymbol{\delta}_j) \rightarrow E_{\mathbb{Q}}[\beta_*(\boldsymbol{\delta})] \quad a.s. - \mathbb{P}\mathbb{Q} \quad \text{as } m, n \rightarrow \infty.$$

Thus, the Wald ELM (WELM) test statistic is computed as

$$\mathcal{W}_{n,m} := \frac{n}{\widehat{\tau}_n^2} \left( m^{-1} \sum_{j=1}^m \widehat{\beta}_n(\boldsymbol{\delta}_j) \right)^2.$$

Before moving to the next section, where we drop the assumption that  $\boldsymbol{\theta}_*$  is known, we note that the functional regression approach considered here differs somewhat from that considered by Cho, Huang, and White [9]. Specifically, they construct a test statistic by regressing  $U_t \Psi(\mathbf{X}_t' \boldsymbol{\delta})$  on the constant 1 and deterministic functions  $\mathbf{g}(\boldsymbol{\delta})$ . Here, however, we regress  $\Psi(\mathbf{X}_t' \boldsymbol{\delta})$  on the constant 1 and  $U_t$ . This permits us to avoid specifying and integrating out the functions  $\mathbf{g}$ , further reducing the computational burden.

## 4 Formal Asymptotics for the WELM Test

Instead of assuming that  $\boldsymbol{\theta}_*$  is known, we now treat it as unknown. We suppose the researcher estimates  $\boldsymbol{\theta}_*$  for the null model by NLS, so that  $\hat{\boldsymbol{\theta}}_n := \arg \min_{\boldsymbol{\theta} \in \Theta} Q_n(\boldsymbol{\theta})$ , where for each  $\boldsymbol{\theta}$ ,

$$Q_n(\boldsymbol{\theta}) := \frac{1}{2n} \sum_{t=1}^n \{Y_t - \Phi(\mathbf{X}_t, \boldsymbol{\theta})\}^2.$$

We impose the following regularity condition.

**Assumption A3 (Estimator):** *There is a sequence of measurable functions  $\{\hat{\boldsymbol{\theta}}_n : \Omega \mapsto \mathbb{R}^d\}$  converging to  $\boldsymbol{\theta}_*$  a.s. such that  $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*) = -n^{1/2} \mathbf{D}_*^{-1} \mathbf{M}_n(\boldsymbol{\theta}_*) + o_{\mathbb{P}}(1) = O_{\mathbb{P}}(1)$ , where  $\boldsymbol{\theta}_* := \arg \min_{\boldsymbol{\theta} \in \Theta} E[Q_n(\boldsymbol{\theta})]$  is unique and interior to  $\Theta$ ;  $\mathbf{D}_*$  is a nonstochastic finite nonsingular  $d \times d$  matrix; and  $\mathbf{M}_n(\boldsymbol{\theta}_*) := n^{-1} \sum_{t=1}^n \mathbf{s}_t(\boldsymbol{\theta}_*)$  such that for every  $t$ ,  $\mathbf{s}_t(\boldsymbol{\theta}_*) := U_t \nabla_{\boldsymbol{\theta}} \Phi(\mathbf{X}_t, \boldsymbol{\theta}_*)$ .*

Assumption A3 is a standard condition ensuring consistency and asymptotic normality for the NLS estimator (see, e.g., White, 1994). Here,  $\mathbf{M}_n(\boldsymbol{\theta}_*)$  is the gradient of  $Q_n(\cdot)$  at  $\boldsymbol{\theta}_*$ ;  $\mathbf{D}_*$  is the probability limit of the Hessian matrix of  $Q_n(\cdot)$  at  $\boldsymbol{\theta}_*$ ; and  $\boldsymbol{\theta}_*$  is not necessarily such that  $E[Y_t | \mathbf{X}_t] = \Phi(\mathbf{X}_t, \boldsymbol{\theta}_*)$ , although this holds under  $\mathcal{H}_0$ .

The NLS estimator is not the only estimator consistent for the conditional mean, as pointed out by White [35]. Other estimators, such as the quasi-maximum likelihood estimator (QMLE) can be used instead. This is useful, for example, when  $Y_t$  is binary, taking only the values 0 or 1. In such cases, we impose A3 with  $Q_n(\cdot)$  chosen to be a quasi-likelihood function. We then interpret  $\hat{\boldsymbol{\theta}}_n$  and  $\mathbf{s}_t(\boldsymbol{\theta}_*)$  as the QMLE estimator and score, respectively. Here, we specify the NLS estimator for concreteness and to convey the key ideas without unnecessary technicality. We note that A3 imposes certain implicit conditions:  $\boldsymbol{\theta}_*$  is not defined without

assuming the existence of  $E[Y_t^2]$  and of  $E[\Phi(\mathbf{X}_t, \boldsymbol{\theta})^2]$  for every  $\boldsymbol{\theta}$ . This implies that A3 cannot be effective without imposing these conditions. The following condition explicitly ensures these, and suffices for a uniform law of large numbers (ULLN).

**Assumption A4 (ULLN):** *There is a sequence of stationary and ergodic random variables  $\{W_t\}$  with  $E[W_t^2] < \infty$  such that*

- (i)  $|Y_t| \leq W_t$ ;
- (ii)  $\sup_{\boldsymbol{\theta} \in \Theta} |\Phi(\mathbf{X}_t, \boldsymbol{\theta})| \leq W_t$ ; and
- (iii)  $|\frac{\partial}{\partial \theta_j} \Phi(\mathbf{X}_t, \boldsymbol{\theta})| \leq W_t, j = 1, \dots, d$ .

We now define the prediction error as  $\widehat{U}_t := Y_t - \Phi(\mathbf{X}_t, \widehat{\boldsymbol{\theta}}_n)$ ,  $t = 1, 2, \dots$ . Next, we define the functional regression estimator. Replacing  $U_t$  in our previous definition with  $\widehat{U}_t$  gives

$$\begin{bmatrix} \widehat{\alpha}_n(\boldsymbol{\delta}) \\ \widehat{\beta}_n(\boldsymbol{\delta}) \end{bmatrix} := \begin{bmatrix} n & \sum_{t=1}^n \widehat{U}_t \\ \sum_{t=1}^n \widehat{U}_t & \sum_{t=1}^n \widehat{U}_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^n \Psi(\mathbf{X}'_t \boldsymbol{\delta}) \\ \sum_{t=1}^n \widehat{U}_t \Psi(\mathbf{X}'_t \boldsymbol{\delta}) \end{bmatrix}. \quad (4)$$

We call this the *two-stage functional ordinary least squares (2SFOLS)* estimator following Cho, Huang, and White [9].

To establish the consistency of the 2SFOLS estimator, we impose the following additional conditions.

**Assumption A5 (Bounds I):** *There is a sequence of stationary and ergodic random variables  $\{M_t\}$  such that*

- (i) for some  $\eta \geq 2(\rho - 1)$ ,  $E[M_t^{4+2\eta}] < \infty$ ;
- (ii)  $|U_t| \leq M_t$ ;
- (iii)  $\sup_{\boldsymbol{\delta} \in \Delta} |\Psi(\mathbf{X}'_t \boldsymbol{\delta})| \leq M_t$ ; and
- (iv)  $\sup_{\boldsymbol{\delta} \in \Delta} |\frac{\partial}{\partial \delta_j} \Psi(\mathbf{X}'_t \boldsymbol{\delta})| \leq M_t, j = 1, \dots, k$ .

Our first lemma is key to establishing the consistency of the functional regression coefficients.

**Lemma 1.** *Given A1, A2, A3, A4, and A5,*

- (i)  $\{n^{-1} \sum \widehat{U}_t, n^{-1} \sum \widehat{U}_t^2\} \rightarrow \{B_*, C_*\}$  a.s.  $-\mathbb{P}$ ;
- (ii) For  $D_n(\boldsymbol{\delta}) := n^{-1} \sum \Psi(\mathbf{X}'_t \boldsymbol{\delta})$  and  $G_{1,n}(\boldsymbol{\delta}) := D_n(\boldsymbol{\delta}) - D_*(\boldsymbol{\delta})$ ,  $\sup_{\boldsymbol{\delta} \in \Delta} |G_{1,n}(\boldsymbol{\delta})| \rightarrow 0$  a.s.  $-\mathbb{P}$ ; and
- (iii) For  $F_n(\boldsymbol{\delta}, \boldsymbol{\theta}) := n^{-1} \sum \{Y_t - \Phi(\mathbf{X}_t, \boldsymbol{\theta})\} \Psi(\mathbf{X}'_t \boldsymbol{\delta})$  and  $G_{2,n}(\boldsymbol{\delta}, \boldsymbol{\theta}) := F_n(\boldsymbol{\delta}, \boldsymbol{\theta}) - F_*(\boldsymbol{\delta})$ ,  $\sup_{\boldsymbol{\delta} \in \Delta} |G_{2,n}(\boldsymbol{\delta}, \widehat{\boldsymbol{\theta}}_n)| \rightarrow 0$  a.s.  $-\mathbb{P}$ .

From this, consistency of  $[\widehat{\alpha}_n, \widehat{\beta}_n]$  for  $[\alpha_*, \beta_*]$  easily follows. To save space, before stating this result, we provide another regularity condition allowing us to establish the desired weak limit. We let  $\lambda_{\min}(\cdot)$  denote the minimum eigenvalue of its argument matrix.

**Assumption A6 (Covariance):** (i)  $\lambda_{\min}\{\mathbf{V}(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2)\} \geq 0$ , where  $\mathbf{V}(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \boldsymbol{\delta}_3, \boldsymbol{\delta}_4) := \text{acov}[\sqrt{n}\mathbf{Z}_n(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2), \sqrt{n}\mathbf{Z}_n(\boldsymbol{\delta}_3, \boldsymbol{\delta}_4)]$ ,  $\mathbf{Z}_n(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) := [\mathbf{M}_n(\boldsymbol{\theta}_*)', G_{1,n}(\boldsymbol{\delta}_1), G_{2,n}(\boldsymbol{\delta}_2, \boldsymbol{\theta}_*)]'$ ; and

(ii) Writing

$$\mathbf{V}(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \boldsymbol{\delta}_3, \boldsymbol{\delta}_4) := \begin{bmatrix} \mathbf{I}_* & \boldsymbol{\varrho}_1(\boldsymbol{\delta}_3) & \boldsymbol{\varrho}_2(\boldsymbol{\delta}_4) \\ \boldsymbol{\varrho}_1(\boldsymbol{\delta}_1)' & \kappa_1(\boldsymbol{\delta}_1, \boldsymbol{\delta}_3) & \kappa_2(\boldsymbol{\delta}_1, \boldsymbol{\delta}_4) \\ \boldsymbol{\varrho}_2(\boldsymbol{\delta}_2)' & \kappa_2(\boldsymbol{\delta}_2, \boldsymbol{\delta}_3) & \kappa_3(\boldsymbol{\delta}_2, \boldsymbol{\delta}_4) \end{bmatrix},$$

$\mathbf{I}_*$  is a nonstochastic finite symmetric positive definite  $d \times d$  matrix; and  $\boldsymbol{\varrho}_i : \Delta \mapsto \mathbb{R}$  and  $\kappa_j : \Delta \times \Delta \mapsto \mathbb{R}$  are continuous,  $i = 1, 2$ ;  $j = 1, 2, 3$ .

**Remark:** We do not require the usual condition that  $\lambda_{\min}\{\mathbf{V}(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2)\}$  is strictly positive, as zero is often encountered in applications. For example, if  $\boldsymbol{\delta}_1 = \boldsymbol{\delta}_2 = \mathbf{0}$ , and  $\Phi(\mathbf{X}_t, \boldsymbol{\theta})$  is an affine function of  $\mathbf{X}_t$ , so that  $\Phi(\mathbf{X}_t, \boldsymbol{\theta}) = \mathbf{X}_t' \boldsymbol{\theta}_1 + \theta_0$  with  $\boldsymbol{\theta} := (\theta_0, \boldsymbol{\theta}_1)'$ , then  $F_n(\mathbf{0}, \widehat{\boldsymbol{\theta}}_n) = n^{-1} \sum \{Y_t - \Phi(\mathbf{X}_t, \widehat{\boldsymbol{\theta}}_n)\} \Psi(0) = \Psi(0) n^{-1} \sum \{Y_t - \Phi(\mathbf{X}_t, \widehat{\boldsymbol{\theta}}_n)\} = 0$ . Hence,  $\lambda_{\min}\{\mathbf{V}(\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0})\} = 0$ . This is a typical twofold identification example of the sort considered by Cho, Ishida, and White [10]. They handle this problem using the QLR statistic, approximated by a quartic expansion. (See Bartlett [3, 4] and McCallagh [26].) Assumption A6 accommodates this degenerate case.

The condition A6 enables us to derive the asymptotic behavior of the scores constituting the WELM statistic. The following lemma formally states this.

**Lemma 2.** Given A1 - A6,  $\sqrt{n}\mathbf{Z}_n \Rightarrow \mathcal{Z}$ , where for each  $\boldsymbol{\delta}$ ,  $\mathcal{Z}(\boldsymbol{\delta}) := [\mathcal{S}'_0, \mathcal{G}_1(\boldsymbol{\delta}), \mathcal{G}_2(\boldsymbol{\delta})]'$ , and  $\mathcal{Z} : \Omega \times \Delta \mapsto \mathbb{R}^{d+2}$  is a zero-mean Gaussian process such that for  $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2$ ,  $E[\mathcal{Z}(\boldsymbol{\delta}_1)\mathcal{Z}(\boldsymbol{\delta}_2)] = \mathbf{V}(\boldsymbol{\delta}_1, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \boldsymbol{\delta}_2)$ .

The result of Lemma 2 corresponds to eq. (3), but Lemma 2 accommodates the parameter estimation error through  $n^{1/2}\mathbf{M}_n(\boldsymbol{\theta}_*)$ . This results in an increase in the dimension of  $\mathbf{V}(\boldsymbol{\delta}, \boldsymbol{\delta})$  relative to that in eq. (3). If  $\boldsymbol{\theta}_*$  is in fact known, there is no need to examine the asymptotic distribution of  $n^{1/2}\mathbf{M}_n(\boldsymbol{\theta}_*)$ .

The next condition ensures that the determinant of the matrix defining  $[\alpha_*, \beta_*]$  is strictly greater than zero, ensuring a unique limit for the 2SFOLS estimator.

**Assumption A7 (Identification):**  $C_* > B_*^2$ .

This condition is equivalent to requiring that  $C_* - B_*^2 = \text{var}(U_t) > 0$ .

The following result establishes the consistency and asymptotic distribution for the 2SFOLS estimator.

**Theorem 1.** *Given A1 - A7,*

- (i)  $\sup_{\delta \in \Delta} |\widehat{\alpha}_n(\delta) - \alpha_*(\delta)| \rightarrow 0$  a.s.  $-\mathbb{P}$ ;
- (ii)  $\sup_{\delta \in \Delta} |\widehat{\beta}_n(\delta) - \beta_*(\delta)| \rightarrow 0$  a.s.  $-\mathbb{P}$ ;
- (iii)  $\sup_{(\theta, \delta) \in \Theta \times \Delta} |n^{-1} \sum \Psi(\mathbf{X}_t' \delta) \frac{\partial}{\partial \theta_j} \Phi(\mathbf{X}_t, \theta) - E[\Psi(\mathbf{X}_t' \delta) \frac{\partial}{\partial \theta_j} \Phi(\mathbf{X}_t, \theta)]| \rightarrow 0$  a.s.  $-\mathbb{P}$ ,  $j = 1, \dots, d$ ; and
- (iv)  $\sqrt{n}[\widehat{\alpha}_n - \alpha_*, \widehat{\beta}_n - \beta_*]' \Rightarrow \mathbf{U}$ , where for each  $\delta$ ,  $\mathbf{U}(\delta) := \mathbf{A}_*^{-1} \mathbf{R}_*(\delta) \mathbf{Z}(\delta)$ ,

$$\mathbf{A}_* := \begin{bmatrix} 1 & B_* \\ B_* & C_* \end{bmatrix}, \quad \mathbf{R}_*(\delta) := \begin{bmatrix} \mathbf{0}' & 1 & 0 \\ -\mathbf{K}_*(\delta)' \mathbf{D}_*^{-1} & 0 & 1 \end{bmatrix},$$

and  $\mathbf{K}_*(\delta) := E[\Psi(\mathbf{X}_t' \delta) \nabla_{\theta} \Phi(\mathbf{X}_t, \theta_*)]$ .

Note that  $\mathbf{U}$  is a linear function of  $\mathbf{Z}$ . Thus, its covariance structure is also determined by that of  $\mathbf{V}$ :

$$E[\mathbf{U}(\delta_1) \mathbf{U}(\delta_2)'] = \mathbf{A}_*^{-1} \mathbf{B}(\delta_1, \delta_2) \mathbf{A}_*^{-1},$$

where  $\mathbf{B}(\delta_1, \delta_2) := \mathbf{R}_*(\delta_1) \mathbf{V}(\delta_1, \delta_1, \delta_2, \delta_2) \mathbf{R}_*(\delta_2)'$ . Here,  $\mathbf{R}_*(\delta)$  accommodates the parameter estimation error. If  $\mathbf{K}_*(\delta) = \mathbf{0}$  for each  $\delta$ , so that  $\mathbf{R}_*(\delta) = [\mathbf{0}_{2 \times d}, \mathbf{I}_{2 \times 2}]$ , then the parameter estimation error does not influence the asymptotic covariance matrix. In this case, it follows that  $E[\mathbf{U}(\delta_1) \mathbf{U}(\delta_2)] = \mathbf{r}(\delta_1, \delta_2)$ .

For notational simplicity in what follows, we partition  $E[\mathbf{U}(\delta_1) \mathbf{U}(\delta_2)]$  as

$$\begin{bmatrix} \varsigma_1(\delta_1, \delta_2) & \varsigma_2(\delta_1, \delta_2) \\ \varsigma_2(\delta_1, \delta_2) & \varsigma_3(\delta_1, \delta_2) \end{bmatrix}.$$

We use Theorem 1 to construct a test that is not model dependent. We achieve this by following Cho, Huang, and White [9], integrating over  $\delta$ . For this, we use the following conditions:

**Assumption A8 (Adjunct Probability Measure):** (i)  $(\Delta, \mathcal{D}, \mathbb{Q})$  is a complete probability space; and

- (ii)  $(\Omega \times \Delta, \mathcal{F} \otimes \mathcal{D}, \mathbb{P} \cdot \mathbb{Q})$  is a complete probability space.

**Assumption A9 (Integrability):** Suppose  $\lambda_{\max}(\mathbf{L}_*) < \infty$  and  $\lambda_{\min}(\mathbf{L}_*) > 0$ , where

$$\mathbf{L}_* := \begin{bmatrix} G_* & J_* \\ J_* & Q_* \end{bmatrix},$$

$G_* := \int_{\Delta} \int_{\Delta} \varsigma_1(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) d\mathbb{Q}(\boldsymbol{\delta}_1)d\mathbb{Q}(\boldsymbol{\delta}_2)$ ,  $J_* := \int_{\Delta} \int_{\Delta} \varsigma_2(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) d\mathbb{Q}(\boldsymbol{\delta}_1)d\mathbb{Q}(\boldsymbol{\delta}_2)$ , and  $Q_* := \int_{\Delta} \int_{\Delta} \varsigma_3(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) d\mathbb{Q}(\boldsymbol{\delta}_1)d\mathbb{Q}(\boldsymbol{\delta}_2)$ .

This condition ensures that the integrals of the 2SFOLS estimator are well defined. We obtain the asymptotic distribution of the integrated 2SFOLS estimators as follows.

**Corollary 1.** Given A1 - A9,

$$\sqrt{n} \begin{bmatrix} \int_{\Delta} \{\widehat{\alpha}_n(\boldsymbol{\delta}) - \alpha_*(\boldsymbol{\delta})\} d\mathbb{Q}(\boldsymbol{\delta}) \\ \int_{\Delta} \{\widehat{\beta}_n(\boldsymbol{\delta}) - \beta_*(\boldsymbol{\delta})\} d\mathbb{Q}(\boldsymbol{\delta}) \end{bmatrix} \overset{A}{\approx} N \left[ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G_* & J_* \\ J_* & Q_* \end{bmatrix} \right].$$

Note that the asymptotic variance  $Q_*$  is different from  $\tau_*$  in the last section, as  $Q_*$  accommodates the parameter estimation error. Unless  $\mathbf{K}_*(\cdot) = \mathbf{0}$ , its influence survives at the limit, and  $Q_* \neq \tau_*$ .

Corollary 1 has important consequences. Under  $\mathcal{H}_0$ ,  $\beta_* \equiv 0$ , so that  $\int_{\Delta} \beta_*(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) = 0$ . Further,  $B_* = 0$  and  $C_* = E[(Y_t - E[Y_t|\mathbf{X}_t])^2]$ , so that

$$\varsigma_1(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) = \kappa_1(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$$

$$\varsigma_2(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) = C_*^{-1} \{ \kappa_2(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) - \boldsymbol{\varrho}_1(\boldsymbol{\delta}_1)' \mathbf{D}_*^{-1} \mathbf{K}_*(\boldsymbol{\delta}_2) \} \quad \text{and}$$

$$\begin{aligned} \varsigma_3(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) = & C_*^{-2} \{ \kappa_3(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) - \boldsymbol{\varrho}_2(\boldsymbol{\delta}_1)' \mathbf{D}_*^{-1} \mathbf{K}_*(\boldsymbol{\delta}_2) \\ & - \mathbf{K}_*(\boldsymbol{\delta}_1)' \mathbf{D}_*^{-1} \boldsymbol{\varrho}_2(\boldsymbol{\delta}_2) + \mathbf{K}_*(\boldsymbol{\delta}_1)' \mathbf{D}_*^{-1} \mathbf{I}_* \mathbf{D}_*^{-1} \mathbf{K}_*(\boldsymbol{\delta}_2) \}. \end{aligned}$$

Under  $\mathcal{H}_1$ ,  $\beta_*(\boldsymbol{\delta})$  differs from 0 almost everywhere, so that  $\int_{\Delta} \beta_*(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) \neq 0$  except in exceptional circumstances. Accordingly, we can test

$$\mathcal{H}'_0 : \int_{\Delta} \beta_*(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) = 0 \quad \text{versus} \quad \mathcal{H}'_1 : \int_{\Delta} \beta_*(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) \neq 0$$

using the Wald statistic

$$\widehat{\mathcal{W}}_n := \frac{n}{\widehat{Q}_n} \left[ \int_{\Delta} \widehat{\beta}_n(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) \right]^2,$$

where  $\widehat{Q}_n$  is a consistent estimator for  $Q_*$ . We provide this below. The asymptotic behavior of  $\widehat{W}_n$  under  $\mathcal{H}_0$  and  $\mathcal{H}_1$  and can be given as follows.

**Theorem 2.** *Given A1 - A9, if  $\widehat{Q}_n = Q_* + o_{\mathbb{P}}(1)$ , then*

- (i) *Under  $\mathcal{H}_0$ ,  $\widehat{W}_n \Rightarrow \chi_1^2$ ; and*
- (ii) *Under  $\mathcal{H}'_1$ ,  $\mathbb{P}[\widehat{W}_n \geq c_n] \rightarrow 1$  for any sequence  $\{c_n\}$  such that  $c_n = o(n)$ .*

Theorem 2 follows easily from Corollary 1, although we provide its proof in the Appendix.

To obtain a consistent estimator  $\widehat{Q}_n$  for  $Q_*$ , we let

$$q_t(\boldsymbol{\theta}) := U_t(\boldsymbol{\theta})[\bar{\Psi}_t - \nabla_{\boldsymbol{\theta}}\Phi(\mathbf{X}_t, \boldsymbol{\theta})\mathbf{J}_*],$$

where for each  $\boldsymbol{\theta} \in \Theta$  and  $t = 1, 2, \dots$ ,

$$U_t(\boldsymbol{\theta}) := Y_t - \Phi(\mathbf{X}_t, \boldsymbol{\theta}), \quad \bar{\Psi}_t := \int_{\Delta} \Psi(\mathbf{X}'_t \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}),$$

$$\mathbf{J}_* := \mathbf{D}_*^{-1}\mathbf{F}_*, \quad \text{and} \quad \mathbf{F}_* := E[\bar{\Psi}_t \nabla_{\boldsymbol{\theta}}\Phi(\mathbf{X}_t, \boldsymbol{\theta}_*)].$$

We use the following conditions to estimate  $Q_*$  consistently.

**Assumption A10 (Bounds II):** *There is a sequence of stationary and ergodic random variables  $\{\widetilde{M}_t\}$  such that*

- (i) *for some  $\widetilde{\eta} \geq 2(\rho - 1)$ ,  $E[\widetilde{M}_t^{4+2\widetilde{\eta}}] < \infty$ ;*
- (ii)  *$|q_t(\cdot)| \leq \widetilde{M}_t$ ;*
- (iii)  *$|\frac{\partial}{\partial \theta_j} q_t(\cdot)| \leq \widetilde{M}_t$  for  $j = 1, 2, \dots, d$ ;*
- (iv)  *$E[\sup_{\boldsymbol{\theta} \in \Theta} |\bar{\Psi}_t \frac{\partial}{\partial \theta_j} \Phi(\mathbf{X}_t, \boldsymbol{\theta})|] < \infty$  for  $j = 1, 2, \dots, d$ ; and*
- (v) *for any  $\boldsymbol{\theta}$  and  $\boldsymbol{\theta}^\dagger \in \Theta$ ,  $|\frac{\partial}{\partial \theta_j} q_t(\boldsymbol{\theta}) - \frac{\partial}{\partial \theta_j} q_t(\boldsymbol{\theta}^\dagger)| \leq A_t \|\boldsymbol{\theta} - \boldsymbol{\theta}^\dagger\|$  such that  $E[A_t] < \infty$  for  $j = 1, \dots, d$ .*

**Assumption A11 (Near Epoch Dependence):**  *$\{q_t(\boldsymbol{\theta})\}$  is near epoch dependent (NED) on  $\{Y_t, \mathbf{X}_t\}$  of size  $-2(2\rho - 1)/(2\rho - 2)$  uniformly on  $\Theta$ , where  $\rho$  is given in A1.*

See Gallant and White (1988) for discussion of NED processes and the definition of NED size. The following lemma provides a class of consistent covariance estimators.

**Lemma 3.** Given A1 - A11, let  $\widehat{Q}_n := (\widehat{\sigma}_n^2)^{-2} \widehat{H}_n$ , where

$$\widehat{H}_n := \frac{1}{n} \left\{ \omega_{n0} \sum_{t=1}^n \widehat{\Gamma}_{n,t}^2 + 2 \sum_{\ell=1}^{k_n} \omega_{n\ell} \sum_{t=\ell+1}^n \widehat{\Gamma}_{n,t} \widehat{\Gamma}_{n,t-\ell} \right\};$$

for each  $t$ ,  $\widehat{\Gamma}_{n,t} := \widehat{U}_t [\bar{\Psi}_t - \nabla'_{\theta} \Phi(\mathbf{X}_t, \widehat{\theta}_n) \widehat{\mathbf{J}}_n]$ ;  $\widehat{\mathbf{J}}_n := \widehat{\mathbf{D}}_n^{-1} \widehat{\mathbf{F}}_n$ ;

$$\widehat{\mathbf{F}}_n := n^{-1} \sum_{t=1}^n \bar{\Psi}_t \nabla_{\theta} \Phi(\mathbf{X}_t, \widehat{\theta}_n);$$

$\widehat{\sigma}_n^2 := n^{-1} \sum_{t=1}^n \widehat{U}_t^2$ ;  $\widehat{\mathbf{D}}_n$  is a consistent estimator for  $\mathbf{D}_*$ , where  $\{k_n\}$  is a sequence of integers such that  $k_n \rightarrow \infty$  as  $n \rightarrow \infty$  with  $k_n = o(n^{1/4})$ ; and  $\omega_{n\ell}$  is such that  $|\omega_{n\ell}| < \bar{\Delta} < \infty$  and for each  $\ell$ ,  $\omega_{n\ell} \rightarrow 1$  as  $n \rightarrow \infty$ . Then  $\widehat{Q}_n = Q_* + o_{\mathbb{P}}(1)$ .

We note that  $\widehat{Q}_n$  has almost same formula as theorem 6.8 of Gallant and White [16], which generalizes the consistent covariance estimator of Newey and West [28] based on the Bartlett [2] kernel. Lemma 3 follows as a corollary to theorem 6.8 of Gallant and White [16]. The result follows by noting that the dataset  $\{(Y_t, \mathbf{X}_t)\}$  in A1 is also an  $\alpha$ -mixing sequence of size  $-2r/(r-2)$  for  $r > 2$ . This trivially holds by letting  $r = 2\rho$  and using the fact that  $2\alpha_{\tau} \leq \beta_{\tau}$  for all  $\tau$ .

The consistent estimator for  $Q_*$  can be further simplified if the disturbance term  $\{U_t := Y_t - \Phi(\mathbf{X}_t, \theta_*)\}$  is a martingale difference sequence (MDS). For this, we assume

**Assumption A12 (MDS):**  $\{U_t, \mathcal{F}_t\}$  is a martingale difference sequence, where for each  $t$ ,  $\mathcal{F}_{t-1}$  is the  $\sigma$ -field generated by  $\{\mathbf{X}_t, U_{t-1}, \mathbf{X}_{t-1}, U_{t-2}, \dots\}$ .

When A12 is imposed,  $Q_*$  can be consistently estimated using a simpler form. The following theorem states this.

**Corollary 2.** Given A1 - A12,  $\widetilde{Q}_n = Q_* + o_{\mathbb{P}}(1)$ , where  $\widetilde{Q}_n := (\widehat{\sigma}_n^2)^{-2} \widetilde{H}_n$  and

$$\widetilde{H}_n := n^{-1} \sum_{t=1}^n \widehat{\Gamma}_{n,t}^2.$$

Note that  $\widetilde{Q}_n$  does not take into account the serial correlation between time-series observations, as A12 ensures that this does not matter. The Wald statistic using

this estimator is

$$\widetilde{\mathcal{W}}_n := \frac{n}{\widetilde{Q}_n} \left[ \int_{\Delta} \widehat{\beta}_n(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) \right]^2.$$

Despite the promising features of these Wald tests, they are challenging to compute, due to need to compute quantities such as  $n^{-1} \sum_{t=1}^n \widehat{U}_t \int_{\Delta} \Psi(\mathbf{X}'_t \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta})$ . We overcome this difficulty by applying ELM. We suppose that IID random variables  $\{\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \dots, \boldsymbol{\delta}_m\}$  are drawn satisfying

**Assumption A13 (Extended Adjunct Probability Space):** (i)  $(\Delta_{\infty}, \mathcal{D}_{\infty}, \mathbb{Q}_{\infty})$  is a complete probability space, where  $\Delta_{\infty} := \times_{j=1}^{\infty} \Delta$ ,  $\mathcal{D}_{\infty}$  is the  $\sigma$ -field generated by  $\{\times_{j=1}^{\infty} D_j : D_j \in \mathcal{D}\}$ , and  $\mathbb{Q}_{\infty}$  is the measure on  $\mathcal{D}_{\infty}$  such that  $\mathbb{Q}_{\infty}(\times_{j=1}^{\infty} D_j) = \times_{j=1}^{\infty} \mathbb{Q}(D_j)$ ;  
(ii)  $(\Omega \times \Delta_{\infty}, \mathcal{F} \otimes \mathcal{D}_{\infty}, \mathbb{P} \cdot \mathbb{Q}_{\infty})$  is a complete probability space; and  
(iii)  $\{\boldsymbol{\delta}_i\}$  is a sequence of IID  $\mathcal{D}_{\infty}$ -measurable random variables, independent of  $\{Y_t, \mathbf{X}'_t\}$ .

The ELM condition enables computing the Wald statistics by applying the LLN. For this, for each  $t$ , we let  $\bar{\Psi}_{m,t} := m^{-1} \sum_{j=1}^m \Psi(\mathbf{X}'_t \boldsymbol{\delta}_j)$  and, with  $\widehat{\beta}_{n,m} := m^{-1} \sum_{j=1}^m \widehat{\beta}_n(\boldsymbol{\delta}_j)$ , we compute the WELM test statistics as:

$$\widehat{\mathcal{W}}_{n,m} := \frac{n(\widehat{\beta}_{n,m})^2}{\widehat{Q}_{n,m}}, \quad \text{and} \quad \widetilde{\mathcal{W}}_{n,m} := \frac{n(\widehat{\beta}_{n,m})^2}{\widetilde{Q}_{n,m}},$$

where  $\widehat{Q}_{n,m} := (\widehat{\sigma}_n^2)^{-2} \widehat{H}_{n,m}$  and  $\widetilde{Q}_{n,m} := (\widetilde{\sigma}_n^2)^{-2} \widetilde{H}_{n,m}$ , respectively. Here, for the same  $\omega_{nl}$  and  $k_n$ ,

$$\widehat{H}_{n,m} := \frac{1}{n} \left\{ \omega_{n0} \sum_{t=1}^n \widehat{\Gamma}_{n,m,t}^2 + 2 \sum_{\ell=1}^{k_n} \omega_{n\ell} \sum_{t=\ell+1}^n \widehat{\Gamma}_{n,m,t} \widehat{\Gamma}_{n,m,t-\ell} \right\};$$

$$\widetilde{H}_{n,m} := n^{-1} \sum_{t=1}^n \widehat{\Gamma}_{n,m,t}^2,$$

where  $\widehat{\Gamma}_{n,m,t} := \widehat{U}_t \bar{\Psi}_{m,t} - \widehat{U}_t \nabla'_{\boldsymbol{\theta}} \Phi(\mathbf{X}_t, \widehat{\boldsymbol{\theta}}_n) \widehat{\mathbf{D}}_n \widehat{\mathbf{F}}_{n,m}$ ; and

$$\widehat{\mathbf{F}}_{n,m} := n^{-1} \sum_{t=1}^n \bar{\Psi}_{m,t} \nabla_{\boldsymbol{\theta}} \Phi(\mathbf{X}_t, \widehat{\boldsymbol{\theta}}_n).$$

The following theorem establishes the properties of the WELM test statistics. To state the result, we write  $m, n \rightarrow \infty$  to denote the limit when first  $m \rightarrow \infty$  and then  $n \rightarrow \infty$ .

**Theorem 3.** *Suppose A1 - A11 and A13 hold. (a) If the conditions in Lemma 3 are further satisfied, then  $\widehat{Q}_{n,m} \rightarrow Q_* + o_{\mathbb{P}\mathbb{Q}_\infty}(1)$  as  $m, n \rightarrow \infty$ ; further,*

(i) *Under  $\mathcal{H}_0$ ,  $\widehat{\mathcal{W}}_{n,m} \Rightarrow \chi_1^2$  as  $m, n \rightarrow \infty$ ; and*

(ii) *Under  $\mathcal{H}'_1$ ,  $\mathbb{P}\mathbb{Q}_\infty[\widehat{\mathcal{W}}_{n,m} \geq c_n] \rightarrow 1$  as  $m, n \rightarrow \infty$  for any sequence  $\{c_n\}$  such that  $c_n = o(n)$ .*

(b) *If A12 also holds, then  $\widetilde{Q}_{n,m} \rightarrow Q_* + o_{\mathbb{P}\mathbb{Q}_\infty}(1)$  as  $m, n \rightarrow \infty$ ; further,*

(i) *Under  $\mathcal{H}_0$ ,  $\widetilde{\mathcal{W}}_{n,m} \Rightarrow \chi_1^2$  as  $m, n \rightarrow \infty$ ; and*

(ii) *Under  $\mathcal{H}'_1$ ,  $\mathbb{P}\mathbb{Q}_\infty[\widetilde{\mathcal{W}}_{n,m} \geq c_n] \rightarrow 1$  as  $m, n \rightarrow \infty$  for any sequence  $\{c_n\}$  such that  $c_n = o(n)$ .*

Therefore, the WELM tests and the Wald tests have essentially the same properties for  $m$  and  $n$  sufficiently large. By letting  $m$  tend to infinity first and then letting  $n$  go to infinity, we first approximate  $\int_{\Delta} \Psi(\mathbf{X}'_t \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta})$  by  $\bar{\Psi}_{m,t}$  and then apply the asymptotics with respect to  $n$ . This simplifies the analysis.

One can refine this result by examining the convergence rate of the WELM test. Letting  $m$  first tend to infinity turns out to be sufficient for obtaining the desired results in Theorem 3. For this, we may apply arguments used in the simulated method of moments literature (Gourieroux and Monfort [18]), which yield

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{t=1}^n \{U_t \bar{\Psi}_{m,t} - E[U_t \bar{\Psi}_t]\} \\ &= \frac{1}{\sqrt{n}} \sum_{t=1}^n \{U_t \bar{\Psi}_t - E[U_t \bar{\Psi}_t]\} + \frac{1}{\sqrt{n}} \sum_{t=1}^n \{U_t \bar{\Psi}_{m,t} - U_t \bar{\Psi}_t\} \\ &= O_{\mathbb{P}\cdot\mathbb{Q}_\infty}(1) + O_{\mathbb{P}\cdot\mathbb{Q}_\infty}(1/\sqrt{m}). \end{aligned}$$

Thus, just letting  $m$  tend to infinity eliminates the uncertainty introduced by using ELM, and this result applies directly to the WELM test. In obtaining our result, the ratio  $n/m$  does not matter, which differs from the result of Gourieroux and Monfort [18]. In their case, as their estimator is nonlinear, one needs to examine the limit of  $n/m$  to ensure the desired asymptotic behavior of their estimator.

In practice, we compute the WELM statistic as follows:

- Step 1: Estimate  $\widehat{\boldsymbol{\theta}}_n$  using NLS and compute  $\widehat{U}_t := Y_t - \Phi(\mathbf{X}_t, \widehat{\boldsymbol{\theta}}_n)$  for every  $t = 1, \dots, n$ ;

- Step 2: For sufficiently large  $m$ , generate a sequence  $\{\boldsymbol{\delta}_j\}$  satisfying A13,  $j = 1, 2, \dots, m$ ;
- Step 3: For each  $t$ , compute  $\bar{\Psi}_{m,t} := m^{-1} \sum_{j=1}^m \Psi(\mathbf{X}'_t \boldsymbol{\delta}_j)$  for sufficiently large  $m$ ;
- Step 4: Compute the integrated 2SFOLS estimator by

$$\begin{bmatrix} \hat{\alpha}_{n,m} \\ \hat{\beta}_{n,m} \end{bmatrix} = \begin{bmatrix} n & \sum_{t=1}^n \hat{U}_t \\ \sum_{t=1}^n \hat{U}_t & \sum_{t=1}^n \hat{U}_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^n \bar{\Psi}_{m,t} \\ \sum_{t=1}^n \hat{U}_t \bar{\Psi}_{m,t} \end{bmatrix};$$

- Step 5: Estimate the asymptotic variance by  $\hat{Q}_{n,m} := (\sigma_n^2)^{-2} \hat{H}_{n,m}$  or  $\tilde{Q}_{n,m} := (\sigma_n^2)^{-2} \tilde{H}_{n,m}$ , where  $\hat{H}_{n,m}$  and  $\tilde{H}_{n,m}$  are given above;
- Step 6: Compute the WELM statistics as in the definition of  $\widehat{\mathcal{W}}_{n,m}$  or  $\widetilde{\mathcal{W}}_{n,m}$ , respectively.

Although the WELM statistic is computationally convenient compared to its non-ELM analog, the terms  $\sum_{t=1}^n \bar{\Psi}_{m,t}$  and  $\sum_{t=1}^n \hat{U}_t \bar{\Psi}_{m,t}$  involve double sums, requiring  $O(mn)$  operations. When  $n$  is large, this can still represent a good deal of computation. An alternative method that requires only  $O(n)$  operations bases the test on  $\hat{\beta}_n^+$ , computed from

$$\begin{bmatrix} \hat{\alpha}_n^+ \\ \hat{\beta}_n^+ \end{bmatrix} = \begin{bmatrix} n & \sum_{t=1}^n \hat{U}_t \\ \sum_{t=1}^n \hat{U}_t & \sum_{t=1}^n \hat{U}_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^n \Psi(\mathbf{X}'_t \boldsymbol{\delta}_t) \\ \sum_{t=1}^n \hat{U}_t \Psi(\mathbf{X}'_t \boldsymbol{\delta}_t) \end{bmatrix}.$$

Note that only single sums appear in the last term on the RHS, where we see that each term uses a different random draw  $\boldsymbol{\delta}_t$ . Nevertheless, under our conditions, we have  $\hat{\beta}_n^+ \rightarrow E_{\mathbb{Q}}[\beta_*(\boldsymbol{\delta})]$  *a.s.*  $-\mathbb{P}\mathbb{Q}_\infty$  as  $n \rightarrow \infty$ , so tests based on  $\hat{\beta}_n^+$  should have some power.

We call tests based on  $\hat{\beta}_n^+$  WELM-plus or simply WELM<sup>+</sup> tests, given their minimal computing requirements. The WELM<sup>+</sup> test statistics are

$$\widehat{\mathcal{W}}_n^+ := \frac{n(\hat{\beta}_n^+)^2}{\hat{Q}_n^+} \quad \text{and} \quad \widetilde{\mathcal{W}}_n^+ := \frac{n(\tilde{\beta}_n^+)^2}{\tilde{Q}_n^+},$$

where  $\hat{Q}_n^+ := (\hat{\sigma}_n^2)^{-2} \hat{H}_n^+$  and  $\tilde{Q}_n^+ := (\tilde{\sigma}_n^2)^{-2} \tilde{H}_n^+$ , respectively. Here,  $\hat{H}_n^+$  and  $\tilde{H}_n^+$  are computed analogously to  $\hat{H}_{n,m}$  and  $\tilde{H}_{n,m}$ , but with  $\hat{\Gamma}_{n,m,t}$  replaced by  $\hat{\Gamma}_{n,t}^+ := \hat{U}_t \Psi(\mathbf{X}'_t \boldsymbol{\delta}_t) - \hat{U}_t \nabla'_\theta \Phi(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_n) \hat{\mathbf{D}}_n \hat{\mathbf{F}}_n^+$ , where

$$\hat{\mathbf{F}}_n^+ := n^{-1} \sum_{t=1}^n \Psi(\mathbf{X}'_t \boldsymbol{\delta}_t) \nabla_\theta \Phi(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_n).$$

The following theorem establishes the properties of the WELM<sup>+</sup> test statistics.

**Theorem 4.** *Suppose A1 - A11 and A13 hold. (a) If the conditions in Lemma 3 are further satisfied, then  $\widehat{Q}_n^+ \rightarrow Q_* + o_{\mathbb{P}\mathbb{Q}_\infty}(1)$  as  $n \rightarrow \infty$ ; further,*

(i) *Under  $\mathcal{H}_0$ ,  $\widehat{\mathcal{W}}_n^+ \Rightarrow \chi_1^2$  as  $n \rightarrow \infty$ ; and*

(ii) *Under  $\mathcal{H}'_1$ ,  $\mathbb{P}\mathbb{Q}_\infty[\widehat{\mathcal{W}}_n^+ \geq c_n] \rightarrow 1$  as  $n \rightarrow \infty$  for any sequence  $\{c_n\}$  such that  $c_n = o(n)$ .*

(b) *If A12 also holds, then  $\widetilde{Q}_n^+ \rightarrow Q_* + o_{\mathbb{P}\mathbb{Q}_\infty}(1)$  as  $n \rightarrow \infty$ ; further,*

(i) *Under  $\mathcal{H}_0$ ,  $\widetilde{\mathcal{W}}_{n,m} \Rightarrow \chi_1^2$  as  $n \rightarrow \infty$ ; and*

(ii) *Under  $\mathcal{H}'_1$ ,  $\mathbb{P}\mathbb{Q}_\infty[\widetilde{\mathcal{W}}_n^+ \geq c_n] \rightarrow 1$  as  $n \rightarrow \infty$  for any sequence  $\{c_n\}$  such that  $c_n = o(n)$ .*

Therefore, the WELM<sup>+</sup> tests and the Wald tests have essentially the same properties for  $n$  sufficiently large.

In practice, we compute the WELM<sup>+</sup> statistic as follows:

- Step 1: Estimate  $\widehat{\boldsymbol{\theta}}_n$  using NLS and compute  $\widehat{U}_t := Y_t - \Phi(\mathbf{X}_t, \widehat{\boldsymbol{\theta}}_n)$  for every  $t = 1, \dots, n$ ;

- Step 2: For sufficiently large  $m$ , generate a sequence  $\{\boldsymbol{\delta}_t\}$  satisfying A13,  $t = 1, \dots, n$ ;

- Step 3: Compute  $\widehat{\beta}_n^+$  as given above;

- Step 4: Estimate the asymptotic variance by  $\widehat{Q}_n^+$  or  $\widetilde{Q}_n^+$  as given above;

- Step 5: Compute the WELM<sup>+</sup> statistics  $\widehat{\mathcal{W}}_n^+$  or  $\widetilde{\mathcal{W}}_n^+$  as given above.

Although the WELM<sup>+</sup> statistic is an order of magnitude simpler to compute, it has the potential drawback that different researchers could reach different inferences based on different draws of  $\{\boldsymbol{\delta}_t\}$ . This is less likely to be an issue with the WELM statistic. Ultimately, whether to use WELM or WELM<sup>+</sup> must be determined by their level and power properties. We investigate these below.

## 5 A Model Exercise

We now illustrate use of our methods by deriving the objects need to compute a WELM statistic suitable for testing the correct specification of a linear model. This provides foundation for our Monte Carlo experiments in the next section.

We suppose data are generated according to the following stationary and ergodic first-order autoregressive process

$$Y_t = \theta_{1*} + \theta_{2*}Y_{t-1} + \varepsilon_t,$$

where  $\varepsilon_t \sim \text{IID } N(0, \sigma_*^2)$ . Given this, we let our explanatory variable  $\mathbf{X}_t$  be  $Y_{t-1}$  and suppose that  $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ , so that our null model is  $\mathcal{M}_0 := \{\Phi(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ , where  $\Phi(\mathbf{X}_t, \boldsymbol{\theta}) := \theta_1 + \theta_2 Y_{t-1}$ . We let  $\Theta = \mathbb{R}^2$ ; the fact that this is not compact will create no difficulties here.

Next, we define our alternative model using the exponential function. That is,  $\mathcal{M} := \{f(\cdot; \boldsymbol{\theta}, \lambda, \delta) : (\boldsymbol{\theta}, \lambda, \delta) \in \Theta \times \Lambda \times \Delta\}$ , where  $f(\mathbf{X}_t; \boldsymbol{\theta}, \lambda, \delta) := \theta_1 + \theta_2 Y_{t-1} + \lambda \exp(\delta Y_{t-1})$ ,  $\Lambda = [-\bar{\lambda}, \bar{\lambda}]$ ,  $\Delta := [\underline{\delta}, \bar{\delta}]$ .

With this choice, the quantities relevant to the Wald ELM test are:  $\mathbf{Z}_n(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) := n^{-1} \sum [\varepsilon_t, \varepsilon_t Y_{t-1}, \exp(\delta Y_{t-1}) - E[\exp(\delta Y_{t-1})], \varepsilon_t \exp(\delta Y_{t-1})]'$ ;  $\mathbf{A}_* = \text{diag}\{1, \sigma_*^2\}$ ;  $\mathbf{I}_* = \sigma_*^2 \mathbf{D}_*$ ;

$$\begin{aligned} \mathbf{K}_*(\delta) &:= \begin{bmatrix} E[\exp(\delta Y_t)] \\ E[Y_t \exp(\delta Y_t)] \end{bmatrix}; \quad \mathbf{D}_* := \begin{bmatrix} 1 & E[Y_t] \\ E[Y_t] & E[Y_t^2] \end{bmatrix}; \\ \boldsymbol{\varrho}_2(\delta) &:= \begin{bmatrix} \sigma_*^2 E[\exp(\delta Y_t)] \\ \sigma_*^2 E[Y_t \exp(\delta Y_t)] \end{bmatrix}; \quad \text{and} \\ \kappa_3(\delta_1, \delta_2) &:= \sigma_*^2 E[\exp\{(\delta_1 + \delta_2) Y_t\}]. \end{aligned}$$

It then follows that

$$\varrho_3(\delta_1, \delta_2) := \frac{1}{\sigma_*^2} M(\delta_1) M(\delta_2) \{\exp(\delta_1 \delta_2 \omega_*^2) - 1 - \delta_1 \delta_2 \omega_*^2\},$$

using the fact that  $Y_t \sim N(\mu_*, \omega_*^2)$ , where  $\mu_* := \theta_{1*}/(1 - \theta_{2*})$ , and  $\omega_*^2 := \sigma_*^2/(1 - \theta_{2*}^2)$ , where  $M(\delta)$  is the moment generating function of a normal random variable with population mean  $\mu_*$  and variance  $\omega_*^2$ . That is,  $M(\delta) := \exp[\mu_* \delta + \frac{1}{2} \omega_*^2 \delta^2]$ . This implies that

$$\begin{aligned} Q_* &= \frac{1}{\sigma_*^2} \int_{\Delta} \int_{\Delta} M(\delta_1) M(\delta_2) \\ &\quad \times \{\exp(\delta_1 \delta_2 \omega_*^2) - 1 - \delta_1 \delta_2 \omega_*^2\} d\mathbb{Q}(\delta_1) d\mathbb{Q}(\delta_2). \end{aligned}$$

Thus,  $\sqrt{n} \int_{\Delta} \widehat{\beta}_n(\delta) d\mathbb{Q}(\delta) \overset{A}{\sim} N(0, Q_*)$  by Corollary 1.

## 6 Monte Carlo Experiments

We report some Monte Carlo experiments in this section. First we consider the behavior of the WELM test under the null hypothesis. We then consider its behavior under the alternative. Finally, we discuss the properties of the WELM<sup>+</sup> test.

Table 1: NULL REJECTION RATES FOR WELM TEST  $\widetilde{\mathcal{W}}_{n,m}$  (IN PERCENT)

NUMBER OF REPLICATIONS: 10,000  
DGP:  $Y_t = U_t; U_t \sim \text{IID } N(0, 1)$   
MODEL:  $Y_t = \theta_1 + \theta_2 Y_{t-1} + \lambda \exp(\delta Y_{t-1}) + U_t$

$n$	50	100	200	500	1,000
Nominal Level	$\delta_j \sim \text{IID } U[-0.5, 0.5]$ and $m = 10,000$				
1.00 %	0.38	0.55	0.67	0.90	0.94
5.00 %	4.04	4.57	4.61	4.66	5.15
10.0 %	10.35	10.52	10.11	9.54	9.84
Nominal Level	$\delta_j \sim \text{IID } U[-1.0, 1.0]$ and $m = 10,000$				
1.00 %	0.35	0.66	0.55	0.77	0.83
5.00 %	4.08	4.27	4.66	4.63	4.62
10.0 %	10.33	9.49	9.90	10.28	9.63

## 6.1 WELM Tests under the Null

### 6.1.1 Examination of $\widetilde{\mathcal{W}}_{n,m}$

For  $\mathcal{H}_0$ , we use the environment described in Section 5, with  $\theta_{1*} = \theta_{2*} = 0$  and  $\sigma_*^2 = 1$ , so that  $\mu_* = 0$  and  $\omega_*^2 = 1$ . Accordingly, the observed data sequence  $\{Y_t\}$  is a sequence of IID standard normal random variables, implying that it is an MDS sequence. We use  $\widetilde{\mathcal{W}}_{n,m}$  for our Monte Carlo simulations, where  $\mathbb{Q}$  is selected as the uniform measure on  $\Delta = [-0.5, 0.5]$ , implying that  $Q_* \approx 0.0040$ , which we compute using MATHEMATICA<sup>®</sup>7.0. In particular, for each  $t$ , we approximate  $\bar{\Psi}_t := \int_{\Delta} \Psi(\mathbf{X}'_t \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta})$  by  $\bar{\Psi}_{m,t} := m^{-1} \sum_{i=1}^m \exp(\mathbf{X}'_t \delta_i)$ , so that  $\tilde{Q}_n := (\hat{\sigma}_n^2)^{-2} \tilde{H}_n$  can be approximated by  $\tilde{Q}_{n,m} := (\hat{\sigma}_n^2)^{-2} \tilde{H}_{n,m}$ , where  $\tilde{H}_{n,m} := n^{-1} \sum_{t=1}^n \hat{\Gamma}_{n,m}^2$ ;

$$\hat{\Gamma}_{n,m,t} := \hat{U}_t P_{m,t} - \hat{\mathbf{F}}'_{n,m} \hat{\mathbf{D}}_n^{-1} \mathbf{s}_t(\hat{\boldsymbol{\theta}}_n); \quad \hat{\mathbf{F}}_{n,m} := n^{-1} \sum_{t=1}^n \bar{\Psi}_{m,t} [1, Y_{t-1}]';$$

and  $\mathbf{s}_t(\hat{\boldsymbol{\theta}}_n) := \hat{U}_t [1, Y_{t-1}]'$ . We let  $m = 10,000$ .

Our simulation results for the null are presented in the first panel of Table 1. The total number of Monte Carlo replications is 10,000. The finite sample null distribution is well approximated by the standard chi-squared distribution with one degree of freedom. Even for small sample sizes, this approximation is successful. We can also affirm this conclusion using the empirical distribution functions. The first panel in Figure 1 shows that these are very close to the chi-

Table 2: NULL REJECTION RATES FOR WELM TEST  $\widehat{\mathcal{W}}_{n,m}$  (IN PERCENT)

NUMBER OF REPLICATIONS: 10,000  
DGP:  $Y_t = 0.75Y_{t-1} - 0.25Y_{t-2} + U_t; U_t \sim \text{IID } N(0, 1)$   
MODEL:  $Y_t = \theta_1 + \theta_2 Y_{t-1} + \lambda \exp(\delta Y_{t-1}) + U_t$

$n$	50	100	200	500	1,000
Nominal Level	$\delta_j \sim \text{IID } U[-0.5, 0.5]$ and $m = 10,000$				
1.00 %	0.54	0.72	0.79	1.12	1.16
5.00 %	5.29	5.81	5.22	6.05	6.22
10.0 %	11.87	11.96	11.17	11.45	11.80
Nominal Level	$\delta_j \sim \text{IID } U[-1.0, 1.0]$ and $m = 10,000$				
1.00 %	0.55	0.48	0.74	0.87	0.99
5.00 %	4.57	4.49	5.40	5.37	5.81
10.0 %	11.11	10.63	11.45	11.57	11.34

squared distribution with one degree of freedom, for sample sizes  $n = 50, 100, 200,$  and  $1,000$  respectively. This affirms Theorem 6.

As another experiment, we consider the case in which  $\Delta$  is larger:  $\Delta = [-1, 1]$ . The experiment is identical to the previous case, except that here we generate  $\delta_j$  from  $U[-1.0, 1.0]$ . The second panel of Table 1 reports the simulation results under  $\mathcal{H}_0$ , and the results are more or less similar to the previous case, although the convergence rate is a bit slower than the first case. This is also affirmed by the second panel in Figure 1, where we show the empirical distributions of the WELM tests along with the chi-squared distribution with one degree of freedom. The sample sizes are  $n = 50, 100, 200,$  and  $1,000$  respectively, as before. All are very close to each other, again affirming Theorem 6.

### 6.1.2 Examination of $\widehat{\mathcal{W}}_{n,m}$

We next examine the WELM test when  $Q_*$  is estimated by  $\widehat{Q}_{n,m}$ . Now data are generated as

$$Y_t = 0.75Y_{t-1} - 0.25Y_{t-2} + U_t,$$

where  $U_t \sim \text{IID } N(0, 1)$ , but the specification is still a first-order autoregression, as in Section 6.1.1. Thus, this model is correctly specified for  $E[Y_t|Y_{t-1}]$ , but it is dynamically misspecified because the specification omits  $Y_{t-2}$ . Hence, we test the correct model assumption using  $\widehat{\mathcal{W}}_{n,m}$ , together with the same simulation designs as in Section 6.1.1.

The null distribution of WELM statistic  $\widehat{\mathcal{W}}_{n,m}$  is also well approximated by the standard chi-squared distribution. Table 2 and Figure 2 contain the simulation results when  $\Delta = [-0.5, 0.5]$  and  $\Delta = [-1.0, 1.0]$ . The results are slightly inferior to those in Table 1 and Figure 1, but the difference is minor.

## 6.2 WELM Tests under the Alternative

We now examine the asymptotic power of the WELM tests. For this, we suppose that data are generated by a nonlinear AR(1) process  $Y_t = \pi_* \cos(Y_{t-1}) + U_t$  with  $U_t \sim \text{IID } N(0, 1)$ . We again assume a first order linear autoregression specification, as in Section 6.1.1. This model is misspecified when  $\pi_* \neq 0$ . Thus, the WELM statistic should reject the null whenever  $\pi_* \neq 0$ . The farther  $\pi_*$  is from zero or the larger is  $n$ , the more often the WELM statistic should reject the null.

The simulation results for  $\widetilde{\mathcal{W}}_{n,m}$  and  $\widehat{\mathcal{W}}_{n,m}$  at the 5% significance level are presented in Tables 3 and 4, respectively. We note that the power of the test is rather good when  $\Delta = [-0.5, 0.5]$  and  $\delta_j$  is randomly drawn from  $U[-0.5, 0.5]$ . As we expect, the empirical rejection rate increases as  $\pi_*$  departs from zero or  $n$  increases.

We also conduct the analogous experiment with  $\Delta = [-1.0, 1.0]$  and draw  $\delta_j$  from  $U[-1.0, 1.0]$ . The second panels of Tables 3 and 4 show the simulation results, and we observe the same encouraging behavior.

As a further experiment to examine power, we consider another distribution for  $\delta_j$ , drawing  $\delta_j$  from the mixture of two distributions. Specifically, let  $Z_j \sim U[0.0, 1.0]$ , and draw  $\delta_j = (Z_j^{50} - 1)/2$  with probability 0.5, and  $(1 - Z_j^{50})/2$  with probability 0.5. Thus, the support set of  $\delta_j$  is  $[-0.5, 0.5]$ ; we denote this as  $M_{50}[-0.5, 0.5]$ . We compare this to the WELM test where  $\delta_j \sim U[-0.5, 0.5]$ . Although these choices have the same support set, their densities are very different. Our simulation results are given in the third panel of Tables 3 and 4. From this, we can see that their power properties are not so different. As a further comparison, we draw  $\delta_j = Z_j^{50} - 1$  with probability 0.5, and  $1 - Z_j^{50}$  with probability 0.5. This distribution is denoted as  $M_{50}[-1.0, 1.0]$ , and the simulation results are provided in the final panel of Tables 3 and 4. We see that this test is slightly weaker than the WELM test with  $\delta_j \sim U[-1.0, 1.0]$ . This suggests that when the parameter space gets larger, selecting the probability measure  $\mathbb{Q}$  becomes more important.

### 6.3 Properties of WELM<sup>+</sup> Tests

Analogous simulations for the WELM<sup>+</sup> test show very similar patterns to those observed in Tables 1 and 2 for the levels of the WELM test, so we do not report these here. On the other hand, the WELM test appears to have much better power than the WELM<sup>+</sup> test. To illustrate, consider the results of Table 5 for the  $\widehat{\mathcal{W}}_n^+$  statistic. We considered the same DGP and model as in Table 4, and we test the hypotheses using the WELM<sup>+</sup> test.

Although this WELM<sup>+</sup> test does have power, its performance is much less impressive than that for  $\widehat{\mathcal{W}}_{n,m}$ . The results for  $\widehat{\mathcal{W}}_n^+$  are even more mediocre, compared to  $\widehat{\mathcal{W}}_{n,m}$ . The reason is simple: the sample variance of  $\widehat{\beta}_n^+$  is several orders of magnitude greater than that of  $\widehat{\beta}_{n,m}$ . Indeed, using  $\widehat{\beta}_n^+$  is comparable to  $\widehat{\beta}_{n,m}$  with  $m = 1$ . Taking  $m$  large stabilizes the test statistic.

We therefore recommend using the WELM test instead of WELM<sup>+</sup> unless  $n$  is quite large. Alternatively, one could perform a quick preliminary test using WELM<sup>+</sup>. If rejection of the null occurs, then one has very strong evidence of misspecification, and one can stop with WELM<sup>+</sup>. If not, one can proceed to conduct the more powerful WELM test. One further possibility is to consider a modified version of  $\widehat{\beta}_n^+$ , say  $\widehat{\beta}_{n,m}^+$ , that replaces  $\Psi(\mathbf{X}'_t \boldsymbol{\delta}_t)$  with  $m^{-1} \sum_{j=1}^m \Psi(\mathbf{X}'_t \boldsymbol{\delta}_{t,j})$  for some moderate value of  $m$ . For brevity, we leave this as a topic for future research.

## 7 Conclusion

In this study, we propose WELM tests for correct model specification, obtained by combining the theory of functional regression of Cho, Huang, and White [9] with that of Extreme Learning Machines. The goals of the WELM test are straightforward: we seek a test powerful against essentially arbitrary alternatives that is simple to compute and that has a standard asymptotic null distribution for the test statistic that is not model dependent. Computation of the WELM statistic is convenient because it computes the associated function integrals using ELM methods. The WELM statistic asymptotic null distribution is standard: it weakly converges to a chi-squared random variable. This makes the WELM test different from the others popularly considered in the literature, as their null distributions are typically functions of Gaussian processes. Further, these distributions are model dependent. The WELM statistic distribution is not model dependent.

We also examine the asymptotic power of the WELM test under the alternative; the test is consistent against essentially arbitrary alternatives.

Our Monte Carlo experiments affirm the asymptotic theory developed here. The WELM statistic's null distribution is well approximated by the chi-squared distribution even for small sample sizes, and it has useful power. The WELM statistic's features make it widely applicable to testing for model misspecification.

## A Proofs

Before proving the main claims in the text, we first provide some preliminary lemmas to facilitate our proofs. For notational implicity, we let  $\Psi_t(\boldsymbol{\delta})$  and  $\Phi(\mathbf{X}_t, \boldsymbol{\theta})$  denote  $\Psi(\mathbf{X}'_t \boldsymbol{\delta})$  and  $\Phi_t(\boldsymbol{\theta})$ , respectively.

**Lemma A1.** *Given A1, A2, and A4,*

- (i)  $\sup_{\boldsymbol{\theta} \in \Theta} |n^{-1} \sum_{t=1}^n U_t(\boldsymbol{\theta}) - E[U_t(\boldsymbol{\theta})]| \rightarrow 0 \text{ a.s. } -\mathbb{P};$
- (ii)  $\sup_{\boldsymbol{\theta} \in \Theta} |n^{-1} \sum_{t=1}^n U_t(\boldsymbol{\theta})^2 - E[U_t(\boldsymbol{\theta})^2]| \rightarrow 0 \text{ a.s. } -\mathbb{P}.$

*Proof of Lemma A1:* (i) We note that

$$\begin{aligned} & \sup_{\boldsymbol{\theta} \in \Theta} |n^{-1} \sum_{t=1}^n \{Y_t - \Phi_t(\boldsymbol{\theta})\} - \{E[Y_t] - E[\Phi_t(\boldsymbol{\theta})]\}| \\ & \leq |n^{-1} \sum_{t=1}^n \{Y_t - E[Y_t]\}| + \sup_{\boldsymbol{\theta} \in \Theta} |n^{-1} \sum_{t=1}^n \{\Phi_t(\boldsymbol{\theta}) - E[\Phi_t(\boldsymbol{\theta})]\}|. \end{aligned}$$

We can now apply the ergodic theorem to  $n^{-1} \sum_{t=1}^n Y_t$  and the ULLN of Ranga Rao (1962) to  $n^{-1} \sum_{t=1}^n \Phi_t(\cdot)$  by A4(i and ii). The desired result follows from this.

(ii) From the definition of  $U_t(\boldsymbol{\theta})$ ,

$$\begin{aligned} & \sup_{\boldsymbol{\theta} \in \Theta} |n^{-1} \sum_{t=1}^n \{Y_t - \Phi_t(\boldsymbol{\theta})\}^2 - E[\{Y_t - [\Phi_t(\boldsymbol{\theta})]\}^2]| \\ & \leq |n^{-1} \sum_{t=1}^n Y_t^2 - E[Y_t^2]| + 2|n^{-1} \sum_{t=1}^n Y_t \Phi_t(\boldsymbol{\theta}) - E[Y_t \Phi_t(\boldsymbol{\theta})]| \\ & \quad + \sup_{\boldsymbol{\theta} \in \Theta} |n^{-1} \sum_{t=1}^n \Phi_t(\boldsymbol{\theta})^2 - E[\Phi_t(\boldsymbol{\theta})^2]|. \end{aligned}$$

We can now apply the ergodic theorem to  $n^{-1} \sum_{t=1}^n Y_t^2$  and the ULLN of Ranga Rao (1962) to  $n^{-1} \sum_{t=1}^n \Phi_t(\cdot)^2$  by A4(i and ii). Further, we note that  $|Y_t \Phi_t(\cdot)| \leq$

$W_t^2$  and  $E[W_t^2] < \infty$ . Thus, we can apply the ULLN to  $n^{-1} \sum_{t=1}^n Y_t \Phi_t(\cdot)$  as well. This complete the proof.  $\blacksquare$

*Proof of Lemma 1:* (i) Given that  $\widehat{\boldsymbol{\theta}}_n \rightarrow \boldsymbol{\theta}_*$  a.s.- $\mathbb{P}$  and that  $n^{-1} \sum U_t(\cdot)$  and  $n^{-1} \sum U_t(\cdot)^2$  satisfy the ULLN as shown in Lemma A1, the desired consequence follows because  $E[U_t(\cdot)]$  and  $E[U_t(\cdot)^2]$  are continuous; for every  $t$ ,  $\widehat{U}_t = U_t(\widehat{\boldsymbol{\theta}}_n)$ ;  $E[U_t(\boldsymbol{\theta}_*)] = B_*$ ; and  $E[U_t(\boldsymbol{\theta}_*)^2] = C_*$ .

(ii) This follows by applying the ULLN of Ranga Rao (1962) and using A5(iii).

(iii) We note that  $\sup_{\boldsymbol{\delta} \in \Delta} |n^{-1} \sum Y_t \Psi_t(\boldsymbol{\delta}) - E[Y_t \Psi_t(\boldsymbol{\delta})]| \rightarrow 0$  a.s.- $\mathbb{P}$  by the ULLN because we can apply Cauchy-Schwarz inequality using A4(i) and A5(iii), where  $E[W_t^2] < \infty$  and  $E[M_t^2] < \infty$ . Also,  $\sup_{\boldsymbol{\delta} \in \Delta} |n^{-1} \sum \Phi_t(\boldsymbol{\theta}) \Psi_t(\boldsymbol{\delta}) - E[\Phi_t(\boldsymbol{\theta}) \Psi_t(\boldsymbol{\delta})]| \rightarrow 0$  a.s.- $\mathbb{P}$  for the same reason. Further,  $E[\Phi_t(\boldsymbol{\theta}) \Psi_t(\boldsymbol{\delta})]$  is a continuous function of  $(\boldsymbol{\theta}, \boldsymbol{\delta})$ . From these, the desired consequence follows.  $\blacksquare$

*Proof of Lemma 2:* As the first step, we prove the tightness of  $\{\sqrt{n}G_{1,n}(\cdot)\}$  and  $\{\sqrt{n}G_{2,n}(\cdot)\}$ . Then, given the covariance structure in A6, the asymptotic weak limits follow as a corollary of this tightness.

We note that  $|\Psi_t(\boldsymbol{\delta}) - \Psi_t(\widetilde{\boldsymbol{\delta}})| \leq M_t \cdot \|\boldsymbol{\delta} - \widetilde{\boldsymbol{\delta}}\|$  by A5(iv), so that

$$E \left[ \sup_{\|\boldsymbol{\delta} - \widetilde{\boldsymbol{\delta}}\| \leq \zeta} |\Psi_t(\boldsymbol{\delta}) - \Psi_t(\widetilde{\boldsymbol{\delta}})|^{2+\eta} \right]^{\frac{1}{2+\eta}} \leq E[M_t^{2+\eta}]^{\frac{1}{2+\eta}} \zeta.$$

Further,  $|E[\Psi_t(\boldsymbol{\delta}) - \Psi_t(\widetilde{\boldsymbol{\delta}})]| \leq |E[M_t]| \cdot \|\boldsymbol{\delta} - \widetilde{\boldsymbol{\delta}}\|$ , so that there is some  $B' > 0$  such that

$$E \left[ \sup_{\|\boldsymbol{\delta} - \widetilde{\boldsymbol{\delta}}\| \leq \zeta} |\Psi_t(\boldsymbol{\delta}) - \Psi_t(\widetilde{\boldsymbol{\delta}}) - E[\Psi_t(\boldsymbol{\delta}) - \Psi_t(\widetilde{\boldsymbol{\delta}})]|^{2+\eta} \right]^{\frac{1}{2+\eta}} \leq B' \zeta.$$

Thus, Ossiander's  $L^{2+\eta}$  entropy is finite, and  $\{\sqrt{n}G_{1,n}(\cdot)\}$  is tight by Theorem 1 of Doukhan, Massart, and Rio (1995), given that  $\{Y_t, \mathbf{X}_t\}$  is a  $\beta$ -mixing sequence with size  $-2\rho/(\rho-1)$ , implying that  $\sum_{\tau=1}^{\infty} \beta_{\tau} \tau^{1/(\rho-1)} < \infty$ .

Next,  $|U_t\{\Psi_t(\boldsymbol{\delta}) - \Psi_t(\widetilde{\boldsymbol{\delta}})\}| \leq M_t \cdot |U_t| \cdot \|\boldsymbol{\delta} - \widetilde{\boldsymbol{\delta}}\|$  by A2, A5(ii), and A5(iv). This implies that

$$E \left[ \sup_{\|\boldsymbol{\delta} - \widetilde{\boldsymbol{\delta}}\| \leq \zeta} |U_t[\Psi_t(\boldsymbol{\delta}) - \Psi_t(\widetilde{\boldsymbol{\delta}})]|^{2+\eta} \right]^{\frac{1}{2+\eta}} \leq E[M_t^{4+2\eta}]^{\frac{1}{2+\eta}} \zeta,$$

and also that  $|E[U_t\{\Psi_t(\boldsymbol{\delta}) - \Psi_t(\tilde{\boldsymbol{\delta}})\}]| \leq |E[M_t^2] \cdot \|\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}\|$ , so that there is  $B > 0$  such that

$$E \left[ \sup_{\|\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}\| \leq \zeta} |U_t[\Psi_t(\boldsymbol{\delta}) - \Psi_t(\tilde{\boldsymbol{\delta}})] - E[U_t[\Psi_t(\boldsymbol{\delta}) - \Psi_t(\tilde{\boldsymbol{\delta}})]]|^{2+\eta} \right]^{\frac{1}{2+\eta}} \leq B\zeta,$$

implying that Ossiander's  $L^{2+\eta}$  entropy is finite, and that  $\{\sqrt{n}G_{2,n}(\cdot)\}$  is tight by the same argument. Thus, the desired weak convergence follows from this tightness and the finite dimensional covariance structure, which we do not prove, as it is straightforward.  $\blacksquare$

*Proof of Theorem 1: (i) and (ii)* We note that

$$\begin{bmatrix} \hat{\alpha}_n - \alpha_* \\ \hat{\beta}_n - \beta_* \end{bmatrix} = \hat{\mathbf{A}}_n^{-1} \begin{bmatrix} G_{1,n} \\ G_{2,n} \end{bmatrix} + (\hat{\mathbf{A}}_n^{-1} - \mathbf{A}_*^{-1}) \begin{bmatrix} E[\Psi_t] \\ E[U_t\Psi_t] \end{bmatrix},$$

where

$$\hat{\mathbf{A}}_n := \begin{bmatrix} 1 & n^{-1} \sum \hat{U}_t \\ n^{-1} \sum \hat{U}_t & n^{-1} \sum \hat{U}_t^2 \end{bmatrix}.$$

Given this, Lemma 1(ii) and (iii) imply that  $G_{1,n}$  and  $G_{2,n}$  uniformly converge to 0 *a.s.*  $-\mathbb{P}$ ; and  $\mathbf{A}_n^{-1} - \mathbf{A}_*^{-1} \rightarrow \mathbf{0}$  *a.s.*  $-\mathbb{P}$  by Lemma 1(i) and A7, so that each element in the second term of the RHS converges to zero *a.s.*  $-\mathbb{P}$ . These two facts imply the desired results.

(iii) We note that  $\sup_{\boldsymbol{\delta} \in \Delta} |\Psi_t(\boldsymbol{\delta})| \leq M_t$  with  $E[M_t^2] < \infty$  by A5; and  $\sup_{\boldsymbol{\theta} \in \Theta} |\frac{\partial}{\partial \theta_j} \Phi_t(\boldsymbol{\theta})| \leq W_t$  with  $E[W_t^2] < \infty$  by A4. The desired result now follows by the ULLN of Ranga Rao (1962) and the Cauchy-Schwarz inequality.

(iv) We note that for some  $\bar{\boldsymbol{\theta}}_n$  between  $\boldsymbol{\theta}_*$  and  $\hat{\boldsymbol{\theta}}_n$ ,

$$\sum \hat{U}_t \Psi_t(\boldsymbol{\delta}) = \sum U_t \Psi_t(\boldsymbol{\delta}) - \sum \nabla'_{\boldsymbol{\theta}} \Phi_t(\bar{\boldsymbol{\theta}}_n) \Psi_t(\boldsymbol{\delta}) (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*)$$

by the mean-value theorem. Thus, we can write  $\sqrt{n}[\hat{\alpha}_n - \alpha_*, \hat{\beta}_n - \beta_*]' = \hat{\mathbf{A}}_n^{-1} \hat{\mathbf{R}}_n \mathbf{Z}_n + o_{\mathbb{P}}(1)$ , where for each  $\boldsymbol{\delta}$ ,

$$\hat{\mathbf{R}}_n(\boldsymbol{\delta}) := \begin{bmatrix} \mathbf{0}' & 1 & 0 \\ -\bar{\mathbf{K}}_n(\boldsymbol{\delta})' \mathbf{D}_*^{-1} & 0 & 1 \end{bmatrix},$$

and  $\bar{\mathbf{K}}_n(\boldsymbol{\delta}) := n^{-1} \sum \nabla_{\boldsymbol{\theta}} \Phi_t(\bar{\boldsymbol{\theta}}_n) \Psi_t(\boldsymbol{\delta})$ . Given this, Theorem 1(iii) shows that  $n^{-1} \sum \nabla_{\boldsymbol{\theta}} \Phi(\cdot) \Psi_t(\cdot)$  obeys the ULLN, so that  $\bar{\mathbf{K}}_n(\cdot) \rightarrow \mathbf{K}_*(\cdot)$  *a.s.*  $-\mathbb{P}$  because  $\bar{\boldsymbol{\theta}}_n \rightarrow \boldsymbol{\theta}_*$  *a.s.*  $-\mathbb{P}$  by A3. Also, applying Lemma 2 implies that  $\{\sqrt{n}(\hat{\boldsymbol{\theta}}_n -$

$\boldsymbol{\theta}_*$ ),  $n^{-1/2} \sum U_t \Psi_t \Rightarrow \{\mathbf{D}_*^{-1} \boldsymbol{\mathcal{S}}_0, \mathcal{G}_2\}$ . From this,  $n^{-1/2} \sum \widehat{U}_t \Psi_t \Rightarrow \mathcal{G}_2 - \mathbf{K}'_* \mathbf{D}_*^{-1} \boldsymbol{\mathcal{S}}_0$  by the continuous mapping theorem. Combining these implies that  $\sqrt{n}[\widehat{\alpha}_n - \alpha_*, \widehat{\beta}_n - \beta_*] \Rightarrow \mathbf{A}_*^{-1} \mathbf{R}_* \boldsymbol{\mathcal{Z}} =: \boldsymbol{\mathcal{U}}$ . ■

*Proof of Corollary 1:* By the continuous mapping theorem and Theorem 1,

$$\sqrt{n} \begin{bmatrix} \int_{\Delta} \{\widehat{\alpha}_n(\boldsymbol{\delta}) - \alpha_*(\boldsymbol{\delta})\} d\mathbb{Q}(\boldsymbol{\delta}) \\ \int_{\Delta} \{\widehat{\beta}_n(\boldsymbol{\delta}) - \beta_*(\boldsymbol{\delta})\} d\mathbb{Q}(\boldsymbol{\delta}) \end{bmatrix} \Rightarrow \begin{bmatrix} \int_{\Delta} \mathcal{U}_1(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) \\ \int_{\Delta} \mathcal{U}_2(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) \end{bmatrix},$$

where for each  $\boldsymbol{\delta}$ ,  $\boldsymbol{\mathcal{U}}(\boldsymbol{\delta})$  is partitioned into  $[\mathcal{U}_1(\boldsymbol{\delta}), \mathcal{U}_2(\boldsymbol{\delta})]'$ . We note that the RHS is a vector of integrated Gaussian processes, so that it is jointly normal, implying that  $\int_{\Delta} \boldsymbol{\mathcal{U}}(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) \sim N[\mathbf{0}, \mathbf{C}_*]$ . This is the desired result. ■

*Proof Theorem 2:* (i) Under  $\mathcal{H}_0$ ,  $\sqrt{n} \int_{\Delta} \widehat{\beta}_n(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) \overset{\Delta}{\rightsquigarrow} N(0, Q_*)$ . Thus,  $\widehat{\mathcal{W}}_n(\mathbb{Q}) \overset{\Delta}{\rightsquigarrow} \chi_1^2$  given that  $\widehat{Q}_n = Q_* + o_{\mathbb{P}}(1)$ , as is assumed.

(ii) Under  $\mathcal{H}_1$ ,  $n \{ \int_{\Delta} \widehat{\beta}_n(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) - \int_{\Delta} \beta_*(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) \}^2 = O_{\mathbb{P}}(1)$ . Thus,  $n \{ \int_{\Delta} \widehat{\beta}_n(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) \}^2 = O_{\mathbb{P}}(n)$ , if  $\int_{\Delta} \beta_*(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) \neq 0$ , as is assumed. ■

*Proof of Lemma 3:* Proving Lemma 3 is completed by verifying the regularity conditions for theorem 6.8 of Gallant and White [16] denoted as DG, OP', MX', SM, DM'', NE''', ID', TL, and WT. We verify these conditions one by one. First, condition DG is satisfied by A1. Second, the compact parameter space  $\Theta$  is assumed in A2; but their  $Q_n(\cdot)$  is not formed from our  $q_t(\cdot)$ . We therefore simply let their  $S_{nt}(\cdot)$  be our  $q_t(\cdot)$ , which is equivalent to assuming that their  $g_n(x)$  and  $\nabla_{\theta} q_{nt}(\cdot)$  are our  $x$  and  $q_t(\cdot)$  respectively. Their OP' then holds. Third, condition MX' is satisfied by our  $\beta$ -mixing condition of size  $-2\rho/(\rho-1)$ . If we let  $r := 2\rho$ , then our dataset  $\{Y_t, \mathbf{X}_t\}$  is also an  $\alpha$ -mixing sequence of size  $-2r/(r-2)$  because  $\alpha_{\tau} \leq \frac{1}{2}\beta_{\tau}$  for all  $\tau$ , which is required by MX'. Fourth, their smoothness (SM) condition requires that our  $q_t(\cdot)$  and  $\nabla_{\theta} q_t(\cdot)$  are  $L_1$ -Lipschitz continuous. By our A2,  $\Phi_t(\cdot)$  is twice continuously differentiable, and A10(i and iii) provide sufficient conditions for the  $L_1$ -dominating Lipschitz constant. This implies that  $q_t(\cdot)$  is  $L_1$ -Lipschitz continuous. Also,  $\nabla_{\theta} q_t(\cdot)$  are  $L_1$ -Lipschitz continuous by A10(v). Fifth, their domination condition (DM) requires that  $q_t(\cdot)$  and  $\nabla_{\theta} q_t(\cdot)$  are  $4\rho$ -dominated in our context. This is already assumed by A10(i, ii, and iii). Sixth, their near epoch (NE) and model identification (ID) conditions are given in A11 and A3 respectively. Seventh, the truncation lag (TL) condition is directly assumed as an additional condition to Lemma 3. Eighth, the covariance matrix in our case is a scalar. Thus, the weight condition (WT) is not necessary under our

consideration. Finally, the condition in Lemma 3 assumes that  $\widehat{\mathbf{D}}_n \rightarrow \mathbf{D}_*$  a.s., and we can apply Ranga Rao [31] to  $\widehat{\mathbf{F}}_n$  using A10(iv), so that  $\widehat{\mathbf{J}}_n \rightarrow \mathbf{J}_*$  a.s. This completes the proof.  $\blacksquare$

*Proof of Corollary 2:* Given A12, this follows trivially by noting that  $\{U_t, \mathcal{F}_t\}$  and  $\{\mathbf{s}_t(\boldsymbol{\theta}_*), \mathcal{F}_t\}$  are martingale difference sequences, so that  $E[U_t | \mathcal{F}_{t-1}] = 0$  and  $E[\mathbf{s}_t(\boldsymbol{\theta}_*) | \mathcal{F}_{t-1}] = \mathbf{0}$ , implying that  $\{\Gamma_t := U_t P_t - \mathbf{F}' \mathbf{D}_* \mathbf{s}_t(\boldsymbol{\theta}_*), \mathcal{F}_t\}$  is an MDS, and  $E[\Gamma_t \Gamma_{t-\ell}] = 0$  for  $\ell = 1, 2, \dots$ , where  $\mathbf{F} := \int_{\Delta} \mathbf{K}(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta})$ . Thus, the off-diagonal elements in  $\widehat{H}_n$  must estimate zero, and  $\frac{\omega_{n0}}{\widehat{\sigma}_n^2} \sum_{t=1}^n \widehat{\Gamma}_{n,t}^2$  of  $\widehat{H}_n$  can consistently estimate  $Q_*$  a.s.  $-\mathbb{P}$ . Letting  $\omega_{n0}$  be one satisfies the condition for  $\omega_{n\ell}$ . Thus,  $\widetilde{H}_n - \widehat{H}_n \rightarrow 0$  a.s.  $-\mathbb{P}$ . This completes the proof.  $\blacksquare$

We provide a supplementary lemma before proving Theorem 3.

**Lemma A2.** *Given A1 - A11 and A13, as  $n, m \rightarrow \infty$ ,*

- (i)  $m^{-1} \sum_{j=1}^m \sqrt{n} [\widehat{\alpha}_n(\boldsymbol{\delta}_j) - \alpha_*(\boldsymbol{\delta}_j), \widehat{\beta}_n(\boldsymbol{\delta}_j) - \beta_*(\boldsymbol{\delta}_j)]' \Rightarrow \int_{\Delta} \boldsymbol{\mathcal{U}}(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta});$  and
- (ii)  $\sqrt{n} [\widehat{\beta}_{n,m} - \int_{\Delta} \beta_*(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta})] \overset{A}{\sim} N(0, Q_*).$

*Proof of Lemma A2:* (i) By the definition of  $(\widehat{\alpha}_n, \widehat{\beta}_n)$ ,

$$\begin{aligned} & \mathbf{A}_* \frac{1}{m} \sum_{j=1}^m \sqrt{n} \begin{bmatrix} \widehat{\alpha}_n(\boldsymbol{\delta}_j) - \alpha_*(\boldsymbol{\delta}_j) \\ \widehat{\beta}_n(\boldsymbol{\delta}_j) - \beta_*(\boldsymbol{\delta}_j) \end{bmatrix} \\ &= \frac{1}{\sqrt{n}} \sum_{t=1}^n \frac{1}{m} \sum_{j=1}^m \begin{bmatrix} \Psi(\mathbf{X}'_t \boldsymbol{\delta}_j) - E_{\mathbb{P}}[\Psi(\mathbf{X}'_t \boldsymbol{\delta}_j)] \\ -\mathbf{K}_*(\boldsymbol{\delta}_j)' \mathbf{D}_*^{-1} \mathbf{s}_t(\boldsymbol{\theta}_*) + U_t \Psi(\mathbf{X}'_t \boldsymbol{\delta}_j) \end{bmatrix} + o_{\mathbb{P}}(1), \end{aligned}$$

where  $E_{\mathbb{P}}[\cdot]$  denotes the expectation operator with respect to  $\mathbb{P}$ . Applying Kolmogorov's LLN implies that  $m^{-1} \sum_{j=1}^m \mathbf{K}_*(\boldsymbol{\delta}_j) \rightarrow \int_{\Delta} \mathbf{K}_*(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta})$  a.s.  $-\mathbb{Q}_{\infty}$ , so that it follows that for each  $\omega \in \Omega$ ,  $m^{-1} \sum_{j=1}^m \mathbf{K}_*(\boldsymbol{\delta}_j)' \mathbf{D}_*^{-1} \mathbf{s}_t(\omega, \boldsymbol{\theta}_*) \rightarrow \int_{\Delta} \mathbf{K}_*(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) \mathbf{D}_*^{-1} \mathbf{s}_t(\omega, \boldsymbol{\theta}_*)$  a.s.  $-\mathbb{Q}_{\infty}$ . It also follows that  $\int_{\Delta} \mathbf{K}_*(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) = \int_{\Omega} \int_{\Delta} \Psi(\mathbf{X}(\omega)' \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) \nabla_{\boldsymbol{\theta}} \Phi(\mathbf{X}(\omega), \boldsymbol{\theta}_*) d\mathbb{P}(\omega)$ , so that  $\int_{\Delta} E_{\mathbb{P}}[\Psi(\mathbf{X}' \boldsymbol{\delta})] d\mathbb{Q}(\boldsymbol{\delta}) = \int_{\Delta} \int_{\Omega} \Psi(\mathbf{X}(\omega)' \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) d\mathbb{P}(\omega)$  by Fubini's theorem, A4, and A5. Next, for each  $\omega \in \Omega$ , we have that

$$\frac{1}{m} \sum_{j=1}^m \Psi(\mathbf{X}_t(\omega)' \boldsymbol{\delta}_j) = \int_{\Delta} \Psi(\mathbf{X}_t(\omega)' \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) + o_{\text{a.s.}-\mathbb{Q}_{\infty}}(1), \quad (5)$$

and similarly

$$\begin{aligned} & \frac{1}{m} \sum_{j=1}^m U_t(\omega) \Psi(\mathbf{X}_t(\omega)' \boldsymbol{\delta}_j) \\ &= \int_{\Delta} U_t(\omega) \Psi(\mathbf{X}_t(\omega)' \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) + o_{\text{a.s.}-\mathbb{Q}_{\infty}}(1) \end{aligned} \quad (6)$$

by Kolmogorov's LLN.

Thus, we have that

$$\begin{aligned} & \mathbf{A}_* \frac{1}{m} \sum_{j=1}^m \sqrt{n} \begin{bmatrix} \widehat{\alpha}_n(\boldsymbol{\delta}_j) - \alpha_*(\boldsymbol{\delta}_j) \\ \widehat{\beta}_n(\boldsymbol{\delta}_j) - \beta_*(\boldsymbol{\delta}_j) \end{bmatrix} \\ &= \frac{1}{\sqrt{n}} \sum_{t=1}^n \left[ \int_{\Delta} \Psi(\mathbf{X}'_t \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) - E_{\mathbb{P}}[\int_{\Delta} \Psi(\mathbf{X}'_t \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta})] \right. \\ & \quad \left. - \int_{\Delta} U_t \Psi(\mathbf{X}'_t \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) - \int_{\Delta} \mathbf{K}_*(\boldsymbol{\delta})' d\mathbb{Q}(\boldsymbol{\delta}) \mathbf{D}_*^{-1} \mathbf{s}_t(\boldsymbol{\theta}_*) \right] \\ & \quad + \frac{1}{\sqrt{n}} \sum_{t=1}^n \{\boldsymbol{\eta}_{m,t}\} + o_{\mathbb{P}}(1), \end{aligned}$$

where  $\boldsymbol{\eta}_{m,t}(\omega, \cdot)$  is  $o_{\text{a.s.}-\mathbb{Q}_{\infty}}(1)$  for each  $\omega \in \Omega$  and  $t$ . That is, for each  $t, \omega$  and each  $\xi \in D \in \mathcal{D}_{\infty}$  with  $\mathbb{Q}_{\infty}(D) = 1$ , for every  $\epsilon > 0$  there exists  $M(\epsilon; t, \omega, \xi)$  such that if  $m > M(\epsilon; t, \omega, \xi)$ , then for  $j = 1, 2$ ,  $|\eta_{m,t,j}(\omega, \xi)| < \epsilon$ , where we let  $\boldsymbol{\eta}_{m,t}(\omega, \xi) = [\eta_{m,t,1}(\omega, \xi), \eta_{m,t,2}(\omega, \xi)]'$ . We further note that for  $j = 1, 2$ ,  $n^{-1/2} \sum_{t=1}^n \eta_{m,t,j}(\omega, \xi) = o_{\mathbb{P}\mathbb{Q}_{\infty}}(1)$  because when  $\epsilon = \delta n^{-(1/2)-a}$  for some  $a > 0$  and  $\delta > 0$ , we can let  $m$  be greater than  $M(\epsilon; t, \omega, \xi)$ , so that  $|\eta_{m,t,j}(\omega, \xi)| < \delta n^{-(1/2)-a}$ .

Finally, we note that Theorem 1(iv) implies that

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{t=1}^n \left[ \int_{\Delta} \Psi(\mathbf{X}'_t \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) - E_{\mathbb{P}}[\int_{\Delta} \Psi(\mathbf{X}'_t \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta})] \right. \\ & \quad \left. - \int_{\Delta} \mathbf{K}_*(\boldsymbol{\delta})' d\mathbb{Q}(\boldsymbol{\delta}) \mathbf{D}_*^{-1} \mathbf{s}_t(\boldsymbol{\theta}_*) + \int_{\Delta} U_t \Psi(\mathbf{X}'_t \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) \right] \\ & \Rightarrow \left[ \int_{\Delta} -\mathbf{K}_*(\boldsymbol{\delta})' \mathbf{D}_*^{-1} \mathbf{s}_0 + \mathcal{G}_2(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) \right] = \mathbf{A}_* \int_{\Delta} \boldsymbol{\mathcal{U}}(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}), \end{aligned}$$

where the equality holds by the definition of  $\boldsymbol{\mathcal{U}}$ , so that

$$\frac{1}{m} \sum_{j=1}^m \sqrt{n} \begin{bmatrix} \widehat{\alpha}_n(\boldsymbol{\delta}_j) - \alpha_*(\boldsymbol{\delta}_j) \\ \widehat{\beta}_n(\boldsymbol{\delta}_j) - \beta_*(\boldsymbol{\delta}_j) \end{bmatrix} \Rightarrow \int_{\Delta} \boldsymbol{\mathcal{U}}(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}),$$

as desired.

(ii) We note that  $\widehat{\beta}_{n,m} := m^{-1} \sum_{j=1}^m \widehat{\beta}_n(\boldsymbol{\delta}_j)$  and that Kolmogorov's LLN implies that  $m^{-1} \sum_{j=1}^m \beta_*(\boldsymbol{\delta}_j) \rightarrow \int_{\Delta} \beta_*(\boldsymbol{\delta}) \mathbb{Q}(\boldsymbol{\delta})$  a.s.- $\mathbb{P} \cdot \mathbb{Q}_{\infty}$ , so that  $m^{-1} \sum_{j=1}^m \sqrt{n}[\widehat{\beta}_n(\boldsymbol{\delta}_j) - \beta_*(\boldsymbol{\delta}_j)] - \sqrt{n}[\widehat{\beta}_{n,m} - \int_{\Delta} \beta_*(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta})] = o_{\mathbb{P}\mathbb{Q}_{\infty}}(1)$ . The desired result follows from Lemma A2(i) and Corollary 1.  $\blacksquare$

*Proof of Theorem 3:* (a) We first show that  $\widehat{Q}_{n,m} = Q_* + o_{\mathbb{P}\mathbb{Q}_{\infty}}(1)$ . First, for each  $\omega \in \Omega$ ,  $m^{-1} \sum_{j=1}^m \Psi(\mathbf{X}_t(\omega)' \boldsymbol{\delta}_j) = \int_{\Delta} \Psi(\mathbf{X}_t(\omega)' \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) + o_{\text{a.s.}-\mathbb{Q}_{\infty}}(1)$  as we saw in (5). This implies that

$$\begin{aligned} \widehat{\mathbf{F}}_{n,m} &:= \frac{1}{nm} \sum_{t=1}^n \sum_{j=1}^m \Psi(\mathbf{X}_t' \boldsymbol{\delta}_j) \nabla_{\boldsymbol{\theta}} \Phi(\mathbf{X}_t, \widehat{\boldsymbol{\theta}}_n) \\ &= \frac{1}{n} \sum_{t=1}^n \int_{\Delta} \Psi(\mathbf{X}_t' \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) \nabla_{\boldsymbol{\theta}} \Phi(\mathbf{X}_t, \widehat{\boldsymbol{\theta}}_n) + o_{\mathbb{P}\mathbb{Q}_{\infty}}(1) = \widehat{\mathbf{F}}_n + o_{\mathbb{P}\mathbb{Q}_{\infty}}(1), \end{aligned}$$

where the first equality holds by A10(iv), and the last equality holds by the definition of  $\widehat{\mathbf{F}}_n$ . Next, by the definition of  $\widehat{\Gamma}_{n,m,t} := \widehat{U}_t[\widehat{\Psi}_{m,t} - \nabla'_{\boldsymbol{\theta}} \Phi(\mathbf{X}_t, \widehat{\boldsymbol{\theta}}_n) \widehat{\mathbf{D}}_n \widehat{\mathbf{F}}_{n,m}]$ , it also follows that, uniformly in  $t$ ,

$$\begin{aligned} \widehat{\Gamma}_{n,m,t} &= \widehat{U}_t[m^{-1} \sum_{j=1}^m \Psi(\mathbf{X}_t' \boldsymbol{\delta}_j) - \nabla'_{\boldsymbol{\theta}} \Phi(\mathbf{X}_t, \widehat{\boldsymbol{\theta}}_n) \widehat{\mathbf{D}}_n \widehat{\mathbf{F}}_{n,m}] \\ &= \widehat{U}_t \left[ \int_{\Delta} \Psi(\mathbf{X}_t' \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) - \nabla'_{\boldsymbol{\theta}} \Phi(\mathbf{X}_t, \widehat{\boldsymbol{\theta}}_n) \widehat{\mathbf{D}}_n \widehat{\mathbf{F}}_n \right] + o_{\mathbb{P}\mathbb{Q}_{\infty}}(1) \\ &= \widehat{\Gamma}_{n,t} + o_{\mathbb{P}\mathbb{Q}_{\infty}}(1), \end{aligned}$$

where the second last equality holds by (5) and  $\widehat{\mathbf{F}}_{n,m} = \widehat{\mathbf{F}}_n + o_{\mathbb{P}\mathbb{Q}_{\infty}}(1)$ , and the last equality holds by the definition of  $\widehat{\Gamma}_{n,t}$ . This fact implies that

$$\begin{aligned} \widehat{H}_{n,m} &= \frac{1}{n} \left\{ \omega_{n0} \sum_{t=1}^n \{ \widehat{\Gamma}_{n,t} + o_{\mathbb{P}\mathbb{Q}_{\infty}}(1) \}^2 \right. \\ &\quad \left. + 2 \sum_{\ell=1}^{k_n} \omega_{n\ell} \sum_{t=\ell+1}^n \{ \widehat{\Gamma}_{n,t} + o_{\mathbb{P}\mathbb{Q}_{\infty}}(1) \} \{ \widehat{\Gamma}_{n,t-\ell} + o_{\mathbb{P}\mathbb{Q}_{\infty}}(1) \} \right\} \\ &= \frac{1}{n} \left\{ \omega_{n0} \sum_{t=1}^n \widehat{\Gamma}_{n,t}^2 + 2 \sum_{\ell=1}^{k_n} \omega_{n\ell} \sum_{t=\ell+1}^n \widehat{\Gamma}_{n,t} \widehat{\Gamma}_{n,t-\ell} \right\} + o_{\mathbb{P}\mathbb{Q}_{\infty}}(1) \\ &= \widehat{H}_n + o_{\mathbb{P}\mathbb{Q}_{\infty}}(1), \end{aligned}$$

where the second last equality holds because  $\sum_{t=1}^n \widehat{\Gamma}_{n,t} = O_{\mathbb{P}Q_\infty}(n)$ ,  $\sum_{\ell=1}^{k_n} \omega_{n\ell} \sum_{t=\ell+1}^n \widehat{\Gamma}_{n,t} = O_{\mathbb{P}Q_\infty}(n)$ , and  $\sum_{\ell=1}^{k_n} \omega_{n\ell} \sum_{t=\ell+1}^n \widehat{\Gamma}_{n,t-\ell} = O_{\mathbb{P}Q_\infty}(n)$ . The desired result now holds as corollary of Lemma 3.

(i) Under  $\mathcal{H}_0$ ,  $\int_{\Delta} \beta_*(\boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta}) = 0$ , so that Lemma A2 (ii) implies that  $\sqrt{n} \widehat{\beta}_{n,m} \overset{A}{\rightsquigarrow} N(0, Q_*)$  and that  $n(\widehat{\beta}_{n,m})^2/Q_* \Rightarrow \chi_1^2$  as  $m, n \rightarrow \infty$ . We already shown that  $\widehat{Q}_{n,m} = Q_* + o_{\mathbb{P}Q_\infty}(1)$ , so that  $\widehat{W}_{n,m} := n(\widehat{\beta}_{n,m})^2/\widehat{Q}_{n,m} = n(\widehat{\beta}_{n,m})^2/Q_* + o_{\mathbb{P}Q_\infty}(1) \Rightarrow \chi_1^2$ .

(ii) Under  $\mathcal{H}_1$ ,  $\int_{\Delta} \beta_*(\boldsymbol{\delta}) \mathbb{Q}(\boldsymbol{\delta}) \neq 0$ , so that  $n(\widehat{\beta}_{n,m})^2/Q_* = O_{\mathbb{P}Q_\infty}(n)$  but not  $o_{\mathbb{P}Q_\infty}(n)$ , and this implies the desired result.

(b) We note that  $E[\Gamma_t \Gamma_{t-\ell}] = 0$  by A12 for  $\ell = 1, 2, \dots$ , implying that  $Q_* = C_*^{-2} E[\Gamma_t^2]$ . We already saw that  $\widetilde{H}_{n,m} = E[\Gamma_t^2] + o_{\mathbb{P}Q_\infty}(1)$  and  $\widehat{\sigma}_n^2 = C_* + o_{\mathbb{P}}(1)$ , while proving that  $\widehat{Q}_{n,m} = Q_* + o_{\mathbb{P}Q_\infty}(1)$  in (a). Thus,  $\widetilde{Q}_{n,m} := (\widehat{\sigma}_n^2)^{-2} \widetilde{H}_{n,m} = Q_* + o_{\mathbb{P}Q_\infty}(1)$ , and the proofs of (i and ii) follow from this and (a).

*Proof of Theorem 4:* The results follow by standard arguments, similar to those previously given, applying the LLN and CLT for  $\beta$ -mixing processes. ■

## References

- [1]
- [2] BARTLETT, M. (1950): “Periodogram Analysis and Continuous Spectra,” *Biometrika*, 37, 1–16.
- [3] BARTLETT, M. (1953a): “Approximate Confidence Intervals,” *Biometrika*, 40, 12–19.
- [4] BARTLETT, M. (1953b): “Approximate Confidence Intervals II. More Than One Unknown Parameter,” *Biometrika*, 40, 306–317.
- [5] BIERENS, H. (1990): “A Consistent Conditional Moment Test of Functional Form,” *Econometrica*, 58, 1443–1458.
- [6] BILLINGSLEY, P. (1999): *Convergence of Probability Measures*. New York: Wiley.
- [7] CANDÈS, (1999): “Ridgelets: Estimating with Ridge Functions,” *Annals of Statistics*, 31, 1561–1599.

- [8] CHO, J. AND WHITE, H. (2007): “Testing for Regime Switching,” *Econometrica*, 75, 1671–1720.
- [9] CHO, J., HUANG, M., AND WHITE, H. (2008): “Testing the Constant Mean Function Using Functional Regression,” Yonsei University Discussion Paper.
- [10] CHO, J., ISHIDA, I., AND WHITE, H. (2010): “Revisiting Tests for Neglected Nonlinearity Using Artificial Neural Networks,” Yonsei University Discussion Paper.
- [11] CHO, J. AND WHITE, H. (2010): “Testing for Unobserved Heterogeneity in Exponential and Weibull Duration Models,” forthcoming, *Journal of Econometrics*.
- [12] DAVIES, R. (1977): “Hypothesis Testing When a Nuisance Parameter is Present Only under the Alternative,” *Biometrika*, 64, 247–254.
- [13] DAVIES, R. (1987): “Hypothesis Testing When a Nuisance Parameter is Present Only under the Alternative,” *Biometrika*, 74, 33–43.
- [14] DELGADO, M. AND STUTE, W. (2008): “Distribution-free Specification Tests of Conditional Models,” *Journal of Econometrics*, 143, 37–55.
- [15] DOUKHAN, P., MASSART, P., AND RIO, E. (1995): “Invariance Principles for Absolutely Regular Empirical Processes,” *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, 31, 393–427.
- [16] GALLANT, R. AND WHITE, H. (1988): *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Oxford: Basil Blackwell.
- [17] GINE, E. AND ZINN, J. (1990): “Bootstrapping General Empirical Measures,” *The Annals of Statistics*, 18, 851–869.
- [18] GOURIEROUX, C. AND MONFORT, A. (1991): “Simulation Based Econometrics in Models with Heterogeneity,” *Annales d’Econometrie et de Statistique*, 20/1, 69-107.
- [19] HANSEN, B. (1996): “Inference When a Nuisance Parameter is Not Identified under the Null Hypothesis,” *Econometrica*, 64, 413–430.

- [20] HANSEN, B. (2006): “Interval Forecasts and Parameter Uncertainty,” *Journal of Econometrics*, 135, 377–398.
- [21] HORNİK, K., STINCHCOMBE, M., AND WHITE, H. (1989): “Multilayer Feedforward Networks are Universal Approximators,” *Neural Networks*, 2, 359–366.
- [22] HORNİK, K., STINCHCOMBE, M., AND WHITE, H. (1990): “Universal Approximation of an Unknown Mapping and Its Derivatives Using Multilayer Feedforward Networks,” *Neural Networks*, 3, 551–560.
- [23] HUANG, G.B., ZHU, Q.-Y., AND SIEW, C.-K. (2006): “Extreme Learning Machine: Theory and Applications,” *Neurocomputing*, 70, 489–501.
- [24] KUAN, C.-M. AND WHITE, H. (1994): “Artificial Neural Networks: An Econometric Perspective,” *Econometric Reviews*, 13, 1–92.
- [25] LEE, T.-H., WHITE, H. AND GRANGER, C. (1993): “Testing for Neglected Nonlinearity in Time Series Models: A Comparison of Neural Network Methods and Alternative Tests,” *Journal of Econometrics*, 56, 269–290.
- [26] MCCULLAGH, P. (1987): *Tensor Methods in Statistics*. London: Chapman and Hall.
- [27] NEYMAN, J. (1959): “Optimal Asymptotic Tests of Composite Statistical Hypotheses,” *Probability and Statistics, The Harald Cramér Volume*. Ed. U. Grenander, 213–234. New York, NY: Wiley.
- [28] NEWEY, W. AND WEST, K. (1987): “A Simple Positive-definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix,” *Econometrica*, 55, 703–708.
- [29] NEYMAN, J. AND SCOTT, E. (1965): “Asymptotically Optimal Tests of Composite Hypotheses for Randomized Experiments with Noncontrolled Predictor Variables,” *Journal of the American Statistical Association*, 60, 699–721.
- [30] NEYMAN, J. AND SCOTT, E. (1966): “On the Use of  $C(\alpha)$  Optimal Tests of Composite Hypothesis,” *Bulletin of the International Statistics Institute*, 41, 477–497.

- [31] RANGA RAO, R. (1962): “Relations Between Weak and Uniform Convergence of Measures with Applications,” *Annals of Mathematical Statistics*, 33, 659–680.
- [32] STINCHCOMBE, M. AND WHITE, H. (1998): “Consistent Specification Testing with Nuisance Parameters Present Only under the Alternative,” *Econometric Theory*, 14, 295–324.
- [33] VAN DER VAART, A. (1998): *Asymptotic Statistics*. New York: Cambridge University Press.
- [34] WALD, A. (1943): “Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large,” *Transactions of the American Mathematical Society*, 54, 426–482.
- [35] WHITE, H. (1982): “Maximum Likelihood Estimation of Misspecified Models,” *Econometrica*, 50, 1–25.
- [36] WHITE, H. (1989a): “An Additional Hidden Unit Test for Neglected Nonlinearity in Multilayer Feedforward Networks,” *Proceedings of the International Joint Conference on Neural Networks*, II, New York, NY: IEEE Press, 451–455.
- [37] WHITE, H. (1989b): “Some Asymptotic Results for Learning in Single Hidden Layer Feedforward Network Models,” *Journal of the American Statistical Association*, 84, 1003–1013.
- [38] WHITE, H. (1990): “Connectionist Nonparametric Regression: Multilayer Feedforward Networks Can Learn Arbitrary Mappings,” *Neural Networks*, 3, 535–549.
- [39] WHITE, H. (1992): “Nonparametric Estimation of Conditional Quantiles Using Neural Networks,” in *Proceedings of the Symposium on the Interface*. New York: Springer-Verlag, 190–199.
- [40] WHITE, H. (1994): *Estimation, Inference, and Specification Analysis*. New York: Cambridge University Press.
- [41] WHITE, H. (1996): “Parametric Statistical Estimation Using Artificial Neural Networks,” in P. Smolensky, M.C. Mozer, and D.E. Rumelhart, eds., *Mathematical Perspectives on Neural Networks*. HillDale, NJ: L. Erlbaum Associates, 719–775.

- [42] WHITE, H. (2001): *Asymptotic Theory for Econometricians*. Orlando: Academic Press.
- [43] WHITE, H. (2006): "Approximate Nonlinear Forecasting Methods," in G. Elliott, C.W.J. Granger, and A. Timmermann, eds., *Handbook of Economic Forecasting*. New York: Elsevier, 460-512.

Figure 1: Empirical Distributions of the WELM Test Statistic  $\widetilde{\mathcal{W}}_{n,m}(\mathbb{Q})$

Number of Replications: 10,000, and  $m = 10,000$

$\Delta = [-0.5, 0.5]$

$\Delta = [-1.0, 1.0]$

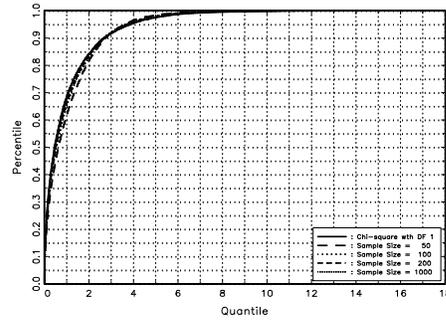
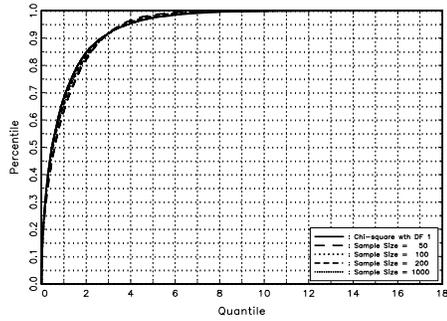


Figure 2: Empirical Distributions of the WELM Test Statistic  $\widehat{\mathcal{W}}_{n,m}(\mathbb{Q})$

Number of Replications: 10,000, and  $m = 10,000$

$\Delta = [-0.5, 0.5]$

$\Delta = [-1.0, 1.0]$

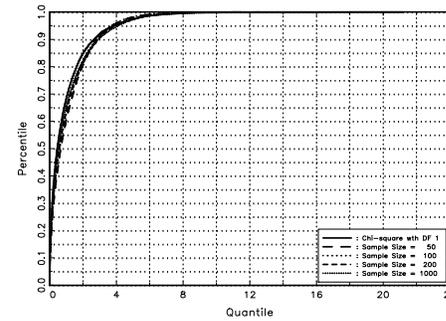
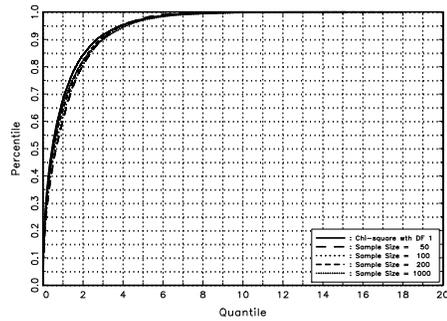


Table 3: ALTERNATIVE REJECTION RATES FOR WELM TEST  $\widetilde{\mathcal{W}}_{n,m}$

LEVEL OF SIGNIFICANCE: 5%

NUMBER OF REPLICATIONS: 2,000

DGP:  $Y_t = \pi_* \cos(Y_{t-1}) + U_t$ ;  $U_t \sim \text{IID } N(0, 1)$

$n$	50	100	200	400	600	800
$\pi_*$	$\delta_j \sim \text{IID } U[-0.5, 0.5]$ and $m = 10,000$					
0.1	5.80	6.70	8.96	13.05	17.65	19.70
0.2	7.70	12.15	22.35	37.80	53.10	65.55
0.3	11.80	24.60	42.30	67.05	86.15	94.50
0.4	16.85	34.75	64.00	90.60	98.30	99.50
0.5	24.80	48.35	79.60	98.35	99.90	100.0
$\pi_*$	$\delta_j \sim \text{IID } U[-1.0, 1.0]$ and $m = 10,000$					
0.1	4.65	6.05	8.00	11.75	15.75	21.05
0.2	6.45	11.60	21.85	36.30	51.35	63.10
0.3	11.35	21.45	38.35	64.60	83.45	90.10
0.4	16.95	32.50	59.75	87.05	96.35	98.90
0.5	23.50	45.40	76.60	95.75	99.30	100.0
$\pi_*$	$\delta_j \sim \text{IID } M_{50}[-0.5, 0.5]$ and $m = 10,000$					
0.1	3.95	7.00	9.30	12.55	16.55	22.50
0.2	8.15	11.55	21.60	37.30	53.70	65.85
0.3	12.32	21.65	40.95	67.85	86.20	93.55
0.4	17.55	35.00	62.20	89.05	97.00	99.40
0.5	24.65	49.00	79.70	97.30	99.70	100.0
$\pi_*$	$\delta_j \sim \text{IID } M_{50}[-1.0, 1.0]$ and $m = 10,000$					
0.1	4.75	5.60	8.75	12.80	16.60	20.30
0.2	5.15	10.95	19.25	32.90	47.00	58.10
0.3	9.25	20.30	36.00	63.30	80.35	89.00
0.4	13.40	28.15	54.15	83.95	94.45	98.00
0.5	21.55	43.75	74.05	93.55	98.45	99.15

Table 4: ALTERNATIVE REJECTION RATES FOR WELM TEST  $\widehat{\mathcal{W}}_{n,m}$

LEVEL OF SIGNIFICANCE: 5%

NUMBER OF REPLICATIONS: 2,000

DGP:  $Y_t = \pi_* \cos(Y_{t-1}) + U_t; U_t \sim \text{IID } N(0, 1)$

$n$	50	100	200	400	600	800
$\pi_*$	$\delta_j \sim \text{IID } U[-0.5, 0.5]$ and $m = 10,000$					
0.1	5.25	5.75	9.05	12.60	17.75	34.21
0.2	7.35	11.65	22.50	39.35	54.30	64.00
0.3	12.45	20.90	41.95	70.40	86.90	93.55
0.4	17.40	34.50	62.90	89.80	97.30	99.55
0.5	23.70	51.10	81.60	98.00	99.70	100.0
$\pi_*$	$\delta_j \sim \text{IID } U[-1.0, 1.0]$ and $m = 10,000$					
0.1	5.50	6.15	7.70	12.55	16.40	21.80
0.2	7.50	11.70	20.35	35.50	51.05	63.00
0.3	10.80	20.10	39.45	66.50	80.90	90.30
0.4	15.75	31.70	59.65	87.00	96.60	98.55
0.5	22.60	44.49	77.70	95.75	99.15	100.0
$\pi_*$	$\delta_j \sim \text{IID } M_{50}[-0.5, 0.5]$ and $m = 10,000$					
0.1	4.95	6.55	9.75	12.30	18.15	22.05
0.2	8.10	12.15	22.10	39.05	54.25	66.30
0.3	11.75	20.70	40.95	68.75	85.65	93.45
0.4	17.35	33.85	62.60	88.70	97.70	99.55
0.5	26.50	47.70	79.30	97.35	99.80	99.95
$\pi_*$	$\delta_j \sim \text{IID } M_{50}[-1.0, 1.0]$ and $m = 10,000$					
0.1	4.85	5.20	8.25	11.90	17.05	19.40
0.2	6.50	11.90	18.65	35.10	46.45	58.20
0.3	9.75	18.25	35.90	64.20	79.70	88.60
0.4	14.35	29.35	56.40	83.45	93.70	97.60
0.5	18.85	43.15	72.50	93.50	97.75	99.10

Table 5: ALTERNATIVE REJECTION RATES FOR WELM TEST  $\widehat{\mathcal{W}}_n^+$

LEVEL OF SIGNIFICANCE: 5%

NUMBER OF REPLICATIONS: 4,000

DGP:  $Y_t = \pi_* \cos(Y_{t-1}) + U_t; U_t \sim \text{IID } N(0, 1)$

$n$	50	100	200	400	600	800
$\pi_*$	$\delta_j \sim \text{IID } U[-0.5, 0.5]$					
0.1	4.47	3.67	4.62	4.75	5.80	6.15
0.2	4.25	4.40	5.32	6.47	6.52	7.22
0.3	4.87	4.82	6.17	7.52	8.50	10.05
0.4	4.77	5.07	6.92	9.97	12.62	17.95
0.5	5.15	5.05	6.67	11.72	14.80	20.45
$\pi_*$	$\delta_j \sim \text{IID } U[-1.0, 1.0]$					
0.1	3.85	3.70	4.77	4.95	5.25	6.70
0.2	3.82	4.05	5.27	7.72	9.05	11.02
0.3	4.17	4.92	7.25	11.92	13.95	18.57
0.4	4.82	6.30	8.42	14.72	23.15	29.42
0.5	4.72	6.92	11.02	21.00	30.77	39.70