Testing for Neglected Nonlinearity Using Extreme Learning Machines

KYU LEE SHIN
Educational Research Institute
Inha University
253 Inha-ro, Nam-gu, Incheon 402-751, Korea
Email: cecil004@hanmail.net

JIN SEO CHO
School of Economics
Yonsei University
50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, Korea
Email: jinseocho@yonsei.com

First version: September 2011 This version: June 2012

Abstract

In this study, we introduce statistics for testing neglected nonlinearity using the extreme leaning machines introduced by Huang, Zhu, and Siew (2006, *Neurocomputing*) and call them ELMNN tests. The ELMNN tests are very convenient and can be widely applied because they are obtained as by-products of estimating linear models, and they can serve as quick diagnostic test statistics complementing the computational burdens of other tests. For the proposed test statistics, we provide a set of regularity conditions under which they asymptotically follow a chi-squared distribution under the null and are consistent under the alternative. We conduct Monte Carlo experiments and examine how they behave when the sample size is finite. Our experiment shows that the tests exhibit the properties desired by the theory of this paper.

Key Words: Extreme learning machines, neglected nonlinearity, Wald test, single layer feedforward network, asymptotic distribution

Acknowledgements: The authors most grateful to the editors and three anonymous referees. We are also benefited from discussions with Halbert White and the authors most grateful to the editors and three anonymous referees. We are also benefited from discussions with Halbert White and the participants of ELM2012 held at Singapore. Jin Seo Cho acknowledges research support from the Korea Sanhak Foundation.

1 Introduction

Testing for neglected nonlinearity is an outstanding problem that has been examined by a number of studies. As an example, Bierens (1990) provides a test statistic for this problem using the exponential function and obtains its asymptotic null distribution as a function of a Gaussian stochastic process. As another example, Stinchcombe and White (1998) show that tests constructed by any non-polynomial analytic function are generically comprehensively revealing (GCR). The appealing feature of their tests is that they are omnibus, so that they can consistently detect any departure from the linearity.

Nevertheless, the existing test statistics are inconvenient for applications. Because they are mostly involved with a nuisance parameter identified only under the alternative as discussed in Davies (1977, 1987), their asymptotic null distributions are difficult to obtain. The existing testing procedures provide their own methodologies to overcome this, and this additional process requires more computational burdens than the simple tests such as the Wald test statistic obtained from the least squares (LS) estimation.

The goal of this study is, therefore, to provide a new statistic for testing for neglected nonlinearity, that is straightforward to compute and can be easily applied even to serially correlated data. We also desire the test statistic to have the asymptotic properties as a desirable test, so that it can be properly evaluated for data sets with large samples. That is, as the sample size tends to infinity, its empirical rejection rate should converge to the level of significance under the null; and it should tend to one under the alternative.

Recently, Huang, Zhu, and Siew (2006) introduce a new model estimation method called extreme learning machines (ELMs) and show that it can universally approximate the conditional mean equation. Readers can also refer to Huang, Wang, and Lan (2011). The breakthrough of this approach is that it can avoid many computational difficulties. Under some mild conditions, the only process involved is to compute the LS estimation. Due to this, the applications of ELMs continue to grow. For example, if we mention a few of them, Chacko, Vimal Krishnan, Raju, and Babu Anto (2012) and Mohammed, Minhas, Jonathan Wu, and Sid-Ahmed (2011) apply ELMs to recognize the characteristics of handwriting and human faces, respectively. As other examples, Wu, Wang, and Chung (2011) and Wang, Chen and Feng (2011) demonstrate that ELMs can be effectively used for data classifications.

We achieve the goal of this paper by applying ELMs. Using the efficiency property of ELMs, we provide test statistics which can effectively test for neglected nonlinearity and avoid computational burdens exhibited by most test statistics (e.g., Cho, Ishida, and White (2011, 2013)). In addition, using the properties of ELMs enables our tests to have standard asymptotic null distributions whereas most of the other tests do not have this property, so that applications of our tests are very straightforward. While developing our test statistics, we further extend the applicability of our test statistics. Theories on ELMs are mostly developed for independently and identically distributed (IID) data, and therefore they are not appropriate for dealing with serially correlated data. We extend the theory on ELMs to the one appropriate for stationary time-series data.

The test statistics we introduce in the current study are different from the existing tests in many respects. First, Cho and White (2011) consider questions similar to those in the current study and provide a test statistic by applying the functional regression in Cho, Huang, and White (2010). Nevertheless, our tests do not require the researcher to specify a particular alternative direction as for the test discussed in Cho and White (2011). Due to this, if an irrelevant direction is specified for the test in Cho and White (2011), our tests can perform better. Furthermore, our tests are computed more straightforwardly and can be applied more easily than the test in Cho and White (2011). As detailed below, our tests are computed as by-products of LS estimations. On the other hand, the testing procedure given by Cho and White (2011) requires to integrate the score with respect to unidentified parameter under the null. Therefore, as the dimension of the unidentified parameters gets bigger, the computation process becomes more complicated. That is why they let the dimension of unidentified parameters be the smallest number. Although the multi-dimensional problem may be resolved by using the distance and direction method in Cho (2012), the computational burdens can still be immense. The testing procedure provided in the current paper does not have this limitation, and this is an advantage over the test in Cho and White (2011) as well as other tests in the literature. Second, the asymptotic null distributions of our tests are standard chi-squared distributions, which implies that we can easily obtain their critical values and we can partially achieve the goal of this study. This is different from the test statistics having asymptotic null distributions represented as functions of a Gaussian stochastic process. Due to this difference, we do not have to go through additional computation processes like the weighted bootstrap in Hansen (1996). This is another advantage of our tests which apply the weighted bootstrap (e.g., Cho, Ishida, and White (2011, 2013)). Finally, our tests have respectable level and power properties. Although our tests are not the most powerful test, they can be properly exploited for testing linearity thanks to its convenience. That is, the researcher can use our tests as quick diagnostic tests complementing the computational burdens of other test statistics existing in the literature.

The plan of this paper is as follows. Section 2 expounds the environments for our test statistics, and we formally define them in the same section. In Section 3, we conduct Monte Carlo experiments and examine their finite sample properties. Section 4 provides concluding remarks. Finally, we present our mathematical proofs in the Appendix.

2 Testing for the Neglected Nonlinearity Using ELMs

2.1 Environments

We suppose the following data generating process (DGP) condition in order to proceed with our discussions in a manageable way.

Assumption 2.1 (DGP). $\{(Y_t, \mathbf{X}_t')' \in \mathbb{R}^{1+k} (k \in \mathbb{N}) : t = 1, 2, \cdots\}$ *is a strictly stationary and absolutely regular process defined on the complete probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*, with* $E(|Y_t|) < \infty$ *and mixing coefficients*

β_τ *such that for some* $\rho > 1$, $\sum_{\tau=1}^{\infty} \tau^{2\rho/(\rho-1)} \beta_\tau < \infty$.

This DGP condition is widely used for the analysis of stationary time-series data. For example, Cho and White (2007, 2011) and Cho, Ishida, and White (2011, 2013) also consider the same condition for their analysis of time-series data. Many popular time-series data such as ARMA and GARCH processes satisfy this condition.

The goal of this study is partially achieved by estimating the conditional mean equation of Y_t given $\mathbf{X}_t$, $E[Y_t|\mathbf{X}_t]$, and the most popular model for the conditional mean is a linear model. In other words, for some $(\alpha_*, \boldsymbol{\beta}_*')' \in \mathbb{R}^{1+k}$, it is popularly assumed that $E[Y_t|\mathbf{X}_t] = \alpha_* + \mathbf{X}_t'\boldsymbol{\beta}_*$.

Nevertheless, the linearity assumption may not be correct. That is, $E[Y_t|\mathbf{X}_t]$ may be a nonlinear function of $\mathbf{X}_t$. Because of this possibility, testing the following hypotheses has been a popular research topic:

$$\mathbb{H}_0: \text{ for some } (\alpha_*, \boldsymbol{\beta}_*),\ E[Y_t|\mathbf{X}_t] = \alpha_* + \mathbf{X}_t'\boldsymbol{\beta}_* \text{ with a probability of 1}$$

$$\mathbb{H}_1: \text{ for any } (\alpha, \boldsymbol{\beta}),\ E[Y_t|\mathbf{X}_t] = \alpha + \mathbf{X}_t'\boldsymbol{\beta} \text{ with a probability of less than 1.}$$

The hypotheses given above have been examined by many previous studies. The first and most widely used test is probably Ramsey's (1969) RESET statistic, which tests whether the coefficients of higher-order polynomial terms of $X_{t,j}$ are zero, where $X_{t,j}$ is the j-th element of $\mathbf{X}_t$. Bierens (1990) considers another model with an additional component constructed using the exponential function and shows that the standard test statistic has a non-standard asymptotic null distribution. It can be represented as a function of a Gaussian stochastic process because of Davies's (1977, 1987) identification problem. Stinchcombe and White (1998) look at this problem from another perspective and note that the exponential function advocated by Bierens (1990) is one of many functions having the omnibus power property. They show that any non-polynomial analytic function can be used for the same purpose and call this the "generically comprehensively revealing property (GCR)." Cho, Ishida, and White (2011) further note that the GCR tests cannot be analyzed in a manner to apply a second-order Taylor's expansion. They classify the set of analytic functions and provide an appropriate theory for analyzing their quasi-likelihood ratio (QLR) test using first-type analytic functions. They find that the QLR test requires a fourth-order Taylor's expansion. They also show that different analytic functions in other sets have to be analyzed in a manner different from that for the first-type analytic functions. For the same test statistic, White and Cho (2012) and Cho, Ishida, and White (2013) extend the analysis in Cho, Ishida, and White (2011) into second-type analytic functions. Although the analysis is more complicated, the QLR test using the second-type analytic functions has properties similar to those of the first-type analytic functions.

The common feature of the tests considered in the literature and mentioned above is that their asymptotic null distributions are not standard. As discussed in Davies (1977, 1987), the so-called nuisance parameter identified only under the alternative hypothesis is present in their contexts, and testing for neglected nonlinearity using the test statistics is more complicated than the process with the standard test statistics such as the Wald, the Lagrange multiplier (LM), and the likelihood-ratio (LR) test statistics.

We first fix our idea by considering a single layer feedforward network (SLFN) model and associate this with the ELM model.

Assumption 2.2 (SLFN). *For a non-polynomial analytic function* $\Psi : \mathbb{R} \mapsto \mathbb{R}$ *such that* $\Psi(0) \neq 0$*, we let* $f(\mathbf{X}_t; \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\delta}) := \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} + \sum_{j=1}^{m} \lambda_j \Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_j)$*, where* $\widetilde{\mathbf{X}}_t = (1, \mathbf{X}_t')'$ *and* $\boldsymbol{\delta} := (\boldsymbol{\delta}_1', \boldsymbol{\delta}_2', \ldots, \boldsymbol{\delta}_m')'$*, and let single hidden layer feedforward network model be defined as* $\mathcal{S} := \{f(\cdot; \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\delta}) : (\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\delta}) \in \boldsymbol{\Theta} \times \boldsymbol{\Lambda} \times \boldsymbol{\Delta}\}$*, where* $\boldsymbol{\Theta} \subset \mathbb{R}^{k+1}$*,* $\boldsymbol{\Lambda} \subset \mathbb{R}^m$*, and* $\boldsymbol{\Delta} \subset \mathbb{R}^{(k+1)\times m}$ *are non-empty compact and convex sets such that* $\mathbf{0} \in \mathrm{int}(\boldsymbol{\Lambda})$ *and* $\mathbf{0} \in \mathrm{int}(\boldsymbol{\Delta})$.

The SLFN has been applied to approximate the conditional mean equation. Indeed, Hornik, Stinchcombe, and White (1989, 1990) show that multilayer feedforward networks are universal approximations. In addition, the previously mentioned tests are constructed by using the SLFN. For example, when $m = 1$ and $\Psi(\cdot) = \exp(\cdot)$, testing $\lambda_* = 0$ reduces to Bierens's (1990) test statistic, where λ_* is the probability limit of the nonlinear least squares (NLS) estimator of λ. As another example, Cho, Ishida, and White (2011, 2013) test $\boldsymbol{\delta}_* = \mathbf{0}$ or $\lambda_* = 0$ by using the QLR test when $m = 1$. Although their asymptotic null distributions are not normal, their tests are asymptotically consistent against any departure from the null hypothesis. This omnibus power property has made them popular test statistics.

Nevertheless, the computational complexity of the SLFN has been pointed out as a drawback, and many efforts have been made to overcome this. ELMs are also developed for this purpose. ELMs modify the model assumption to the following model.

Assumption 2.3 (ELM). *(i)* $\{\boldsymbol{\delta}_j \in \mathbb{R}^{1+k} (k \in \mathbb{N}) : j = 1, 2, \cdots, m\}$ *is an IID process defined on the complete probability space* $(\boldsymbol{\Delta}, \mathcal{D}, \mathbb{Q})$ *such that* $\boldsymbol{\Delta}$ *is a compact subset of* $\mathbb{R}^{(1+k)\times m}$ *with* $\mathbf{0} \in \mathrm{int}(\boldsymbol{\Delta})$*,* $\mathcal{D}$ *is the Borel-sigma field on* $\boldsymbol{\Delta}$*, and* $\mathbb{Q}$ *is an absolutely continuous probability measure with respect to Lebesgue measure* μ*; (ii)* $(\Omega \times \boldsymbol{\Delta}, \mathcal{F} \otimes \mathcal{D}, \mathbb{P} \cdot \mathbb{Q})$ *is a complete probability space such that* $\mathbb{P}$ *is independent of* $\mathbb{Q}$*; and (iii) for a non-polynomial analytic function* $\Psi : \mathbb{R} \mapsto \mathbb{R}$ *with* $\Psi(0) \neq 0$*, we let* $f(\mathbf{X}_t, \boldsymbol{\delta}; \boldsymbol{\theta}, \boldsymbol{\lambda}) := \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} + \sum_{j=1}^{m} \lambda_j \Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_j)$ *and let ELM model be defined as* $\mathcal{E} := \{f(\cdot; \boldsymbol{\theta}, \boldsymbol{\lambda}) : (\boldsymbol{\theta}, \boldsymbol{\lambda}) \in \boldsymbol{\Theta} \times \boldsymbol{\Lambda}\}$*, where* $\boldsymbol{\Theta} \subset \mathbb{R}^{k+1}$*, and* $\boldsymbol{\Lambda} \subset \mathbb{R}^m$ *are non-empty compact and convex sets such that* $\mathbf{0} \in \mathrm{int}(\boldsymbol{\Lambda})$.

Thus, the model $\mathcal{E}$ plugs randomly generated $\{\boldsymbol{\delta}_j\}$ into $\mathcal{S}$ and estimates the other linear parameters $(\alpha, \boldsymbol{\beta}, \boldsymbol{\lambda})$. Thus, estimating unknown parameters is not so difficult. Furthermore, the compact parameter space assumption on $\boldsymbol{\Delta}$ and the IID assumption of $\{\boldsymbol{\delta}_j\}$ imply that $\boldsymbol{\Delta}$ can be considered to be a Cartesian product of $k + 1$ dimensional identical compact parameter spaces, say $\mathbf{T} \in \mathbb{R}^{1+k}$, so that $\boldsymbol{\Delta} \equiv \bigtimes_{j=1}^{m} \mathbf{T}$.

This model is first examined by Huang, Zhu, and Siew (2006),[1] and Theorem 2.2 of Huang, Zhu, and

[1] Huang, Wang, and Lan (2011) survey popular uses of ELMs in the literature. Furthermore, many variations of ELMs are developed. For example, to overcome the instability and over-fitting problems of ELMs, Zhai, Xu, and Wang (2012) provide another data classification algorithm called the dynamic ensemble ELM which is developed from the maximum ambiguity-based sample selection rule (e.g., Wang, Dong, and Yan (2012)).

Siew (2006) shows that $\mathcal{E}$ can also universally approximate the conditional mean under mild regularity conditions.

2.2 The ELMLS Estimator and Its Probability Limits under the Two Hypotheses

The goal of the current study is achieved more specifically by examining the objective function defined to estimate the linear coefficients. We may let the LS estimator be defined as

$$(\widehat{\theta}_n(\delta), \widehat{\lambda}_n(\delta)) := \underset{\theta,\lambda}{\arg\min} \sum_{t=1}^{n} \left(Y_t - \widetilde{\mathbf{X}}_t'\theta - \mathbf{H}_t(\delta)'\lambda \right)^2,$$

where $\mathbf{H}_t(\delta) := [H_t(\delta_1),\ldots,H_t(\delta_m)]' := [\Psi(\widetilde{\mathbf{X}}_t'\delta_1),\ldots,\Psi(\widetilde{\mathbf{X}}_t'\delta_m)]'$. Note that the LS estimator is indexed by δ, and we use this index to indicate that obtained estimates are different for different δ. We call it the ELM least squares (ELMLS) estimator. When the number of parameters $1 + k + m$ is less than the sample size n, its formula is as follows:

$$\widehat{\xi}_n(\delta) := \begin{bmatrix} \widehat{\theta}_n(\delta) \\ \widehat{\lambda}_n(\delta) \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} & \widetilde{\mathbf{X}}'\mathbf{H}(\delta) \\ \mathbf{H}(\delta)'\widetilde{\mathbf{X}} & \mathbf{H}(\delta)'\mathbf{H}(\delta) \end{bmatrix}^{-1} \begin{bmatrix} \widetilde{\mathbf{X}}'\mathbf{Y} \\ \mathbf{H}(\delta)'\mathbf{Y} \end{bmatrix}, \qquad (1)$$

where $\widetilde{\mathbf{X}} := [\widetilde{\mathbf{X}}_1,\ldots,\widetilde{\mathbf{X}}_n]'$, $\mathbf{H}(\delta) := [\mathbf{H}_1(\delta),\ldots,\mathbf{H}_n(\delta)]'$, and $\mathbf{Y} := [Y_1,\ldots,Y_n]'$. Here, n is larger than $m+k+1$, so that we can compute the ELMLS estimator by using the LS estimation. In case $m+k+1 \geq n$, the ELMLS estimator is computed in a different way. Huang, Zhu, and Siew (2006) and Huang, Wang, and Lan (2011) suggest computing this by Moore-Penrose's generalized inverse matrix, and Yuan, Wang, and Cao (2011) provide practical guidance for this purpose. The main reason to restrict our interests to $m + k + 1 < n$ is to see how the ELMLS estimator behaves when the sample size is large. When the number of parameters is larger than the sample size, the analysis of the ELMLS estimator is more challenging. We leave this as a future research topic.

Given the definition of the ELMLS estimator, we impose the following assumption for its regular large sample properties.

Assumption 2.4 (Regularity). *(i) $E[Y_t^2] < \infty$; (ii) for each $j = 1,2,\ldots,k$, $E[X_{t,j}^2] < \infty$; (iii) $E[\sup_{\tau\in\mathbf{T}} |\Psi(\widetilde{\mathbf{X}}_t\tau)|^2] < \infty$; (iv) for each $j = 1,2,\ldots,k+1$, $E[\sup_{\tau\in\mathbf{T}} |\partial/\partial\tau_j\ \Psi(\widetilde{\mathbf{X}}_t\tau)|^2] < \infty$; and (v) $\mathbf{A}(\delta)$ is positive definite with a probability of 1, where*

$$\mathbf{A}(\delta) := \begin{bmatrix} E[\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t'] & E[\widetilde{\mathbf{X}}_t\mathbf{H}_t(\delta)'|\delta] \\ E[\mathbf{H}_t(\delta)\widetilde{\mathbf{X}}_t'|\delta] & E[\mathbf{H}_t(\delta)\mathbf{H}_t(\delta)'|\delta] \end{bmatrix}.$$

This is a standard moment condition for applying the law of large numbers (LLN) although Assumption 2.4(v) is additionally added. Nevertheless, this additional condition is not difficult to satisfy. As $\boldsymbol{\delta}$ is randomly drawn from the continuous probability measure $\mathbb{Q}$, the chance for $\boldsymbol{\delta}_j$ and $\boldsymbol{\delta}_i$ to be identical is zero, where $\boldsymbol{\delta}_j$ and $\boldsymbol{\delta}_i$ are the j-th and i-th column of $\boldsymbol{\delta}$, respectively. Thanks to this, many popular DGPs easily satisfy Assumption 2.4(v).

Given these regularity conditions, we can interrelate the probability limit of the ELMLS estimator with the probability limit of the objective function. The following lemma formally states this.

Lemma 2.1. *Given Assumptions 2.1, 2.3, and 2.4, $\widehat{\boldsymbol{\xi}}_n(\boldsymbol{\delta}) \overset{\mathbb{P}\cdot\mathbb{Q}}{\to} \boldsymbol{\xi}_*(\boldsymbol{\delta})$, where*

$$\boldsymbol{\xi}_*(\boldsymbol{\delta}) := [\boldsymbol{\theta}_*(\boldsymbol{\delta})', \boldsymbol{\lambda}_*(\boldsymbol{\delta})']' := \underset{\boldsymbol{\theta},\boldsymbol{\lambda}}{\arg\min} E[(E[Y_t|\mathbf{X}_t] - \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})^2|\boldsymbol{\delta}].$$

Lemma 2.1 is not hard to prove. We note that $[\widehat{\boldsymbol{\theta}}_n(\boldsymbol{\delta})', \widehat{\boldsymbol{\lambda}}_n(\boldsymbol{\delta})']$ is the argument minimizing $n^{-1}\sum_{t=1}^n(Y_t - \mathbf{X}_t'\boldsymbol{\theta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})^2$, and applying the uniform law of large numbers (ULLN) yields

$$\frac{1}{n}\sum_{t=1}^n (Y_t - \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})^2 \overset{\mathbb{P}}{\to} E[(Y_t - \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})^2|\boldsymbol{\delta}]$$

from the fact that $\boldsymbol{\delta}$ is independent of $\{(Y_t, \mathbf{X}_t')'\}$. As proved in the Appendix (see the Proof of Lemma 2.1), it also follows that

$$\begin{aligned} & E[(Y_t - \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})^2|\boldsymbol{\delta}] \\ & \quad = E[(Y_t - E[Y_t|\mathbf{X}_t])^2] + E[(E[Y_t|\mathbf{X}_t] - \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})^2|\boldsymbol{\delta}], \end{aligned} \tag{2}$$

so that the argument minimizing the left-hand side (LHS) is also the argument minimizing the second term on the right-hand side (RHS). This is mainly because the first term does not involve itself with $(\boldsymbol{\theta}, \boldsymbol{\lambda})$. We can exploit this feature to test for neglected nonlinearity.

In addition, from Eq. (2), additional clues can be obtained for testing the hypotheses. First, the null hypothesis implies that for some $\boldsymbol{\theta}_* := (\alpha_*, \boldsymbol{\beta}_*')'$, $E[Y_t|\mathbf{X}_t] = \widetilde{\mathbf{X}}_t'\boldsymbol{\theta}_*$, so that we can minimize the RHS of Eq. (2) by letting $(\boldsymbol{\theta}_*(\boldsymbol{\delta}), \boldsymbol{\lambda}_*\ (\boldsymbol{\delta}))$ be $(\boldsymbol{\theta}_*, \mathbf{0})$, and

$$\min_{\boldsymbol{\theta},\boldsymbol{\lambda}} E[(Y_t - \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})^2|\boldsymbol{\delta}] = E[(Y_t - E[Y_t|\mathbf{X}_t])^2].$$

Furthermore, this argument uniquely minimizes the LHS of Eq. (2), given that $\mathbf{H}_t(\boldsymbol{\delta})$ is constructed by using the analytic functions, and this holds irrespective of how $\boldsymbol{\delta}$ is generated. Second, the ELMLS estimator behaves differently under the alternative. For this examination, we note that

$$\widehat{\lambda}_n(\delta) = (\mathbf{H}(\delta)'\mathbf{H}(\delta) - \mathbf{H}(\delta)'\widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\mathbf{H}(\delta))^{-1}(\mathbf{H}(\delta)'\mathbf{Y} - \mathbf{H}(\delta)'\widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\mathbf{Y})$$

from Eq. (1). We can further apply the LLN to this estimator and obtain the following probability limit:

$$\begin{aligned}\lambda_*(\delta) := (E[\mathbf{H}_t(\delta)\mathbf{H}_t(\delta)|\delta] - E[\mathbf{H}_t(\delta)\widetilde{\mathbf{X}}_t'|\delta]E[\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t']^{-1}E[\widetilde{\mathbf{X}}_t\mathbf{H}_t(\delta)'|\delta])^{-1} \\ \times E[\mathbf{H}_t(\delta)\{Y_t - \widetilde{\mathbf{X}}_t'E[\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t']^{-1}E[\widetilde{\mathbf{X}}_t Y_t]\}|\delta].\end{aligned}$$

We note that $\{Y_t - \widetilde{\mathbf{X}}_t'E[\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t']^{-1}E[\widetilde{\mathbf{X}}_t Y_t]\}$ is the linear projection error of Y_t on $\widetilde{\mathbf{X}}_t$. In addition, $\mathbf{A}(\delta)$ is positive definite with a probability of 1 according to Assumption 2.4(v). Therefore, $\lambda_*(\delta) = \mathbf{0}$ with a probability of 1 if and only if $E[\mathbf{H}_t(\delta)\{Y_t - \widetilde{\mathbf{X}}_t'E[\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t']^{-1}E[\widetilde{\mathbf{X}}_t Y_t]\}|\delta] = \mathbf{0}$ with a probability of 1, implying that if $E[\mathbf{H}_t(\delta)\{Y_t - \widetilde{\mathbf{X}}_t'E[\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t']^{-1}E[\widetilde{\mathbf{X}}_t Y_t]\}|\delta] \neq \mathbf{0}$ with a probability greater than zero, $\lambda_*(\delta) \neq \mathbf{0}$ with a probability greater than zero, which cannot arise under $\mathbb{H}_0$. Thus testing $\lambda_*(\delta) = \mathbf{0}$ with a probability greater than zero is equivalent to testing $\mathbb{H}_0$. The following lemma further strengthens this argument and provides a result key to the goal of this paper.

Lemma 2.2. *Given Assumptions 2.1, 2.3, and 2.4, (i)* $E[\mathbf{H}_t(\delta)\{Y_t - \widetilde{\mathbf{X}}_t'E[\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t']^{-1}\ E[\widetilde{\mathbf{X}}_t Y_t]\}|\delta] = \mathbf{0}$ *with probabilities equal to one and zero under* $\mathbb{H}_0$ *and* $\mathbb{H}_1$*, respectively; (ii) if for any* $\varepsilon > 0$*, there are* n_0 *and* $F \in \mathcal{D}$ *with* $\mathbb{Q}(F) = 1$ *and* 0 *under* $\mathbb{H}_0$ *and* $\mathbb{H}_1$*, respectively, such that for any* $n > n_0$*,* $\mathbb{P} \cdot \mathbb{Q}(\|\widehat{\lambda}_n(\delta)\| > \varepsilon|F) < \varepsilon$*, where* $\|\cdot\|$ *is the Euclidean norm.*

It is not difficult to prove Lemma 2.2(i). Indeed, similar results are available in the literature. By letting $\Psi(\cdot) = \exp(\cdot)$, Bierens's (1990) shows that $E[\exp(\widetilde{\mathbf{X}}_t'\boldsymbol{\tau})\{Y_t - \widetilde{\mathbf{X}}_t'E[\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t']^{-1}E[\widetilde{\mathbf{X}}_t Y_t]\}] = 0$ at $\boldsymbol{\tau} \in F$ such that $\mu(F) = 0$ under the alternative. Stinchcombe and White (1998) further extend this, and their corollary 3.9 shows that any other non-polynomial analytic function is GCR. In other words, $E[\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\tau})\{Y_t - \widetilde{\mathbf{X}}_t'E[\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t']^{-1}E[\widetilde{\mathbf{X}}_t Y_t]\}] \neq 0$ essentially for every $\boldsymbol{\tau}$. This fact leads to Lemma 2.2(i), and Lemma 2.2(ii) states its implication: the hypotheses can be tested by using the probability limit of $\widehat{\lambda}_n(\delta)$. If $\lambda_*(\delta) = \mathbf{0}$ with a probability 1, $\mathbb{H}_0$ holds and vice versa. From this aspect, we can construct test statistics with omnibus power by testing $\lambda_*(\delta) = \mathbf{0}$ with a probability 1.

2.3 Asymptotic Distribution of the ELMLS Estimator under $\mathbb{H}_0$

Given the implication of Lemma 2.2, we now examine the asymptotic null distribution of the ELMLS estimator to determine the asymptotic null distribution of our tests defined below. For this purpose, we first apply the functional central limit theorem (FCLT) to the ELMLS estimator. We first let $\mathbf{U}$ be the vector of projector errors. Using the definition of $\widehat{\xi}_n(\delta)$,

$$\widehat{\xi}_n(\delta) - \xi_*(\delta) = \begin{bmatrix} \widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} & \widetilde{\mathbf{X}}'\mathbf{H}(\delta) \\ \mathbf{H}(\delta)'\widetilde{\mathbf{X}} & \mathbf{H}(\delta)'\mathbf{H}(\delta) \end{bmatrix}^{-1} \begin{bmatrix} \widetilde{\mathbf{X}}'\mathbf{U} \\ \mathbf{H}(\delta)'\mathbf{U} \end{bmatrix}$$

under $\mathbb{H}_0$, and we can apply the FCLT to the RHS of this equation. For this examination, we let

$$\mathbf{W}_n(\delta) := \frac{1}{\sqrt{n}} \begin{bmatrix} \widetilde{\mathbf{X}}'\mathbf{U} \\ \mathbf{H}(\delta)'\mathbf{U} \end{bmatrix} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left[\widetilde{\mathbf{X}}_t' U_t, \Psi(\widetilde{\mathbf{X}}_t'\delta_1)U_t, \cdots, \Psi(\widetilde{\mathbf{X}}_t'\delta_m)U_t\right]'$$

and note that this can be interpreted as a vector constructed by plugging $\{\delta_j : j = 1, \ldots, m\}$ into $n^{-1/2}\sum_{t=1}^{n}$ $\Psi(\widetilde{\mathbf{X}}_t'(\cdot))U_t$ such that $\mathbb{Q}$ is independent of $\mathbb{P}$, and $\{\delta_j : j = 1, \ldots, m\}$ is a set of continuous random variables. We assume the following regularity conditions so that we can apply the FCLT.

Assumption 2.5 (Bounds and Covariance). *There is a stationary and ergodic process $\{M_t\}$ such that (i) for $\eta \geq 2(\rho-1)$, $E[M_t^{4+2\eta}] < \infty$; (ii) $|U_t| \leq M_t$; (iii) for each $j = 1, 2, \ldots, k$, $|X_{t,j}| \leq M_t$; (iv) $\sup_{\tau\in\mathbf{T}} |\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\tau})| \leq M_t$; and (v) for each $j = 1, 2, \ldots, k+1$, $\sup_{\tau\in\mathbf{T}} |\partial/\partial\tau_j \Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\tau})| \leq M_t$. Furthermore, for each $\boldsymbol{\tau} \in \mathbf{T}$, $\Sigma(\boldsymbol{\tau}, \boldsymbol{\tau})$ is positive definite, where for each $\boldsymbol{\tau}$ and $\widetilde{\boldsymbol{\tau}}$,*

$$\Sigma(\boldsymbol{\tau}, \widetilde{\boldsymbol{\tau}}) := acov\left[\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\begin{bmatrix} \widetilde{\mathbf{X}}_t U_t \\ \Psi(\widetilde{\mathbf{X}}_t'(\boldsymbol{\tau}))U_t \end{bmatrix}, \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\begin{bmatrix} \widetilde{\mathbf{X}}_t U_t \\ \Psi(\widetilde{\mathbf{X}}_t'(\widetilde{\boldsymbol{\tau}}))U_t \end{bmatrix}\right],$$

and $acov(\cdot,\cdot)$ indicates the asymptotic covariance matrix of given arguments.

Given our DGP condition in Assumption 2.1, the mixing coefficient condition combined with the moment coefficient condition in Assumption 2.5(*i*, *ii*, and *iii*) enables us to apply the FCLT discussed in Doukhan, Massrt, and Rio (1995). These conditions are popularly used in the previous studies. For example, Cho and White (2011) use similar conditions to obtain the asymptotic null distribution of their test statistic. In addition, Hansen (1996), and Cho, Ishida, and White (2011, 2013) employ similar conditions for their FCLTs. As many stationary time-series data satisfy this condition, we also use this condition. In addition to this, the positive-definite covariance matrix assumption is imposed for obtaining a regular Gaussian stochastic process as the limit of $\mathbf{Z}_n(\cdot)$, where

$$\mathbf{Z}_n(\cdot) := \begin{bmatrix} \mathbf{Z}_{n,0} \\ \mathbf{Z}_n(\cdot) \end{bmatrix} := \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\begin{bmatrix} \widetilde{\mathbf{X}}_t U_t \\ \Psi(\widetilde{\mathbf{X}}_t'(\cdot))U_t \end{bmatrix}.$$

Without satisfying this condition, we cannot apply the Carmér-Wold device, which is necessary to extend a finite dimensional CLT to an infinite dimensional CLT.

The following lemma provides the asymptotic distribution of $\mathbf{Z}_n(\cdot)$ by using the regularity conditions provided so far.

Lemma 2.3. *Given Assumptions 2.1, 2.3 –2.5, and* $\mathbb{H}_0$, *(i)* $\mathbf{Z}_n(\cdot) \Rightarrow \mathcal{Z}(\cdot) := [\mathcal{G}_0', \mathcal{G}(\cdot)]'$, *where* $\mathcal{Z}(\cdot)$ *is a zero-mean Gaussian stochastic process such that for each* τ *and* $\widetilde{\tau} \in \mathbf{T}$, $E[\mathcal{Z}(\tau)\mathcal{Z}(\widetilde{\tau})] = \Sigma(\tau, \widetilde{\tau})$; *(ii)* $\mathbf{W}_n(\delta)|\delta \Rightarrow \mathcal{W}(\delta)|\delta$, *where* $\mathcal{W}(\delta) := [\mathcal{G}_0', \mathcal{G}(\delta_1), \mathcal{G}(\delta_2), \ldots, \mathcal{G}(\delta_m)]'$; *and (iii)* $\sqrt{n}[\widehat{\xi}_n(\delta) - \xi_*(\delta)]|\delta \Rightarrow \mathcal{U}(\delta)|\delta$, *where* $\mathcal{U}(\delta) := \mathbf{A}(\delta)^{-1}\mathcal{W}(\delta)$.

Lemma 2.3(*i*) implies that for every $\tau \in \mathbf{T}$, the weak distribution of $\mathbf{Z}_n(\tau)$ is available and that the asymptotic conditional distribution of $\mathbf{Z}_n(\delta_j)$ on δ_j is derived as $\mathcal{Z}(\delta_j)$ conditional on δ_j. Even when multiple δ_j's are drawn randomly and independently, so that δ is defined as $[\delta_1, \delta_2, \ldots, \delta_m]$ as before, we can obtain similar results thanks to the FCLT and the independence property between $\mathbb{P}$ and $\mathbb{Q}$. Lemma 2.3(*ii* and *iii*) are obtained by exploiting this simple fact. Specifically, the asymptotic distribution $\mathcal{U}(\delta)$ conditional on δ is provided as

$$\mathcal{U}(\delta)|\delta \sim N\left(\mathbf{0}_{m\times 1}, \mathbf{A}(\delta)^{-1}\mathbf{B}(\delta)\mathbf{A}(\delta)^{-1}\right), \tag{3}$$

where

$$\mathbf{B}(\delta) := \text{acov}\left[\frac{1}{\sqrt{n}}\begin{bmatrix} \widetilde{\mathbf{X}}'\mathbf{U} \\ \mathbf{H}(\delta)'\mathbf{U} \end{bmatrix}, \frac{1}{\sqrt{n}}\begin{bmatrix} \widetilde{\mathbf{X}}'\mathbf{U} \\ \mathbf{H}(\delta)'\mathbf{U} \end{bmatrix} \middle| \delta\right].$$

Most statistics testing for neglected nonlinearity have asymptotic null distributions represented as functions of the given Gaussian stochastic process $\mathcal{U}(\cdot)$. For example, Baek and Cho (2013), Bierens (1990), Cho and Han (2009), Cho and Ishida (2012), Cho and White (2007, 2010, 2011), Cho, Ishida, and White (2011, 2013), Hansen (1996), among others, define their test statistics using the Gaussian stochastic process in Lemma 2.3. Although their test statistics have respectable level and power properties, it is not so convenient to exploit them due to their exploitation of the Gaussian stochastic process.

2.4 Testing Hypotheses

We define test statistics in this subsection to overcome the drawbacks of the existing tests. As Demsăr (2006) point out, there are many ways to define test statistics and their finite sample properties are not uniquely determined. We define our test statistics so that they are obtained as by-products of linear model estimations for the goal of this paper. Furthermore, we desire standard chi-squared distributions as their asymptotic null distributions, which implies that we can test for neglected nonlinearity by using standard statistical packages.

To achieve this goal, we obtain the asymptotic null distribution of $\widehat{\lambda}_n(\delta)|\delta$ using Lemma 2.3. If we let

$\mathbf{S} := [\mathbf{O}_{m\times(k+1)} \vdots \mathbf{I}_m]$, $\mathbf{S}(\widehat{\boldsymbol{\xi}}_n(\delta) - \boldsymbol{\xi}_*(\delta)) = (\widehat{\boldsymbol{\lambda}}_n(\delta) - \boldsymbol{\lambda}_*(\delta))$, so that Lemma 2.3(*iii*) implies that

$$\sqrt{n}(\widehat{\boldsymbol{\lambda}}_n(\delta) - \boldsymbol{\lambda}_*(\delta))|\delta = \sqrt{n}\mathbf{S}(\widehat{\boldsymbol{\xi}}_n(\delta) - \boldsymbol{\xi}_*(\delta))|\delta \Rightarrow \mathbf{S}\boldsymbol{\mathcal{U}}(\delta)|\delta,$$

and the asymptotic conditional null distribution of $\widehat{\boldsymbol{\lambda}}_n(\delta)$ on δ is easily determined by this.

Before converting this asymptotic conditional null distribution into the asymptotic null distribution of the tests defined below, we consider the following additional condition.

Assumption 2.6 (MDA). *(i) For some $\sigma_*^2 > 0$, $E[U_t^2|\mathbf{X}_t] = \sigma_*^2$; and (ii) $\{U_t, \mathcal{F}_t\}$ is a martingale difference array (MDA), where for each t, $\mathcal{F}_{t-1}$ is the σ-field generated by $\{\mathbf{X}_t, U_{t-1}, \mathbf{X}_{t-1}, U_{t-2}, \ldots\}$.*

Assumption 2.6 may be too strong a condition for general time-series data: it assumes the MDA and conditional homoskedasticity condition. Although this is more restrictive, we nevertheless consider this and examine its implications to the asymptotic conditional null distribution mainly because this condition is often assumed for analyses of empirical data (e.g., Demsăr (2006)). In addition, when data are IID observations, Assumption 2.6 is relevant under a conditional homoskedasticity condition.

We now define our test statistics. We note that $\mathbb{H}_0$ implies that $\boldsymbol{\lambda}_*(\delta) = \mathbf{0}$ with a probability 1 and vice versa by Lemma 2.2. Thus, if $\mathbb{H}_0$ is correct,

$$\sqrt{n}\widehat{\boldsymbol{\lambda}}_n(\delta)|\delta \overset{\text{A}}{\sim} N(\mathbf{0}, \mathbf{S}\mathbf{A}(\delta)^{-1}\mathbf{B}(\delta)\mathbf{A}(\delta)^{-1}\mathbf{S}'); \qquad (4)$$

otherwise, the the location parameter of $\sqrt{n}\widehat{\boldsymbol{\lambda}}_n(\delta)|\delta$ is different from $\mathbf{0}$ for whatever δ realizes. We exploit this to test the hypotheses using our test statistics defined as

$$\widetilde{T}_n(\delta) := n\widehat{\boldsymbol{\lambda}}_n(\delta)'[\mathbf{S}\mathbf{A}(\delta)^{-1}\mathbf{B}(\delta)\mathbf{A}(\delta)^{-1}\mathbf{S}']^{-1}\widehat{\boldsymbol{\lambda}}_n(\delta);$$

$$\ddot{T}_n(\delta) := \frac{n}{\sigma_*^2}\widehat{\boldsymbol{\lambda}}_n(\delta)'[\mathbf{S}\mathbf{A}(\delta)^{-1}\mathbf{S}']^{-1}\widehat{\boldsymbol{\lambda}}_n(\delta).$$

In particular, $\ddot{T}_n(\delta)$ is defined for data sets satisfying Assumption 2.6. Note that the tests are defined based on the Wald testing procedure.[2] The asymptotic conditional distributions of our tests are provided as chi-squared distributions under the null; otherwise, they are not bounded in probability due to the fact that the location parameter of $\sqrt{n}\widehat{\boldsymbol{\lambda}}_n(\delta)|\delta$ is different from $\mathbf{0}$.

Nevertheless, $\mathbf{A}(\delta)$ and $\mathbf{B}(\delta)$ are unknown, so that $\widetilde{T}_n$ and $\ddot{T}_n$ cannot be computed using the data set. We, therefore, replace them with their consistent estimators $\widehat{\mathbf{A}}_n(\delta)$ and $\widehat{\mathbf{B}}_n(\delta)$. We specifically define them

[2] Thus, our test statistic has a similar motivation to the paired t-test discussed by Demsăr (2006), who advocates using non-parametric tests more than the paired t-test under his circumstance. Demsăr (2006) explains that this is mainly due to the wrong assumption on the normality and the homoskedasticity. Nevertheless, our set of assumptions relaxes these conditions.

as follows:

$$\widehat{\mathbf{A}}_n(\delta) := \frac{1}{n}\sum_{t=1}^{n}\begin{bmatrix} \widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t & \widetilde{\mathbf{X}}_t\mathbf{H}_t(\delta)' \\ \mathbf{H}_t(\delta)\widetilde{\mathbf{X}}_t & \mathbf{H}_t(\delta)\mathbf{H}_t(\delta)' \end{bmatrix}; \tag{5}$$

$$\widehat{\mathbf{B}}_n(\delta) := \frac{\omega_{n0}}{n}\sum_{t=1}^{n}\widetilde{U}_t^2\mathbf{\Gamma}_t(\delta)\mathbf{\Gamma}_t(\delta)' + \sum_{j=1}^{q_n}\frac{\omega_{nj}}{n-j}\sum_{t=j+1}^{n}\widetilde{U}_t\widetilde{U}_{t-j}\left(\mathbf{\Gamma}_t(\delta)\mathbf{\Gamma}_{t-j}(\delta)' + \mathbf{\Gamma}_{t-j}(\delta)\mathbf{\Gamma}_t(\delta)'\right), \tag{6}$$

where $\mathbf{\Gamma}_t(\delta) := [1, Y_{t-1}, \Psi(\widetilde{\mathbf{X}}_t'\delta_1), \Psi(\widetilde{\mathbf{X}}_t'\delta_2), \ldots, \Psi(\widetilde{\mathbf{X}}_t'\delta_m)]'$; $\widetilde{U}_t := Y_t - \widetilde{\mathbf{X}}_t\widehat{\boldsymbol{\theta}}_n(\delta) - \mathbf{H}_t(\delta)'\widehat{\boldsymbol{\lambda}}(\delta)$; and q_n and ω_{nj} are functions of n that satisfy the conditions given below.

Similar estimators are used in the context of NLS estimation. For example, Gallant and White's (1988) theorem 6.8 provides a similar estimator. In addition, Newey and West's (1987) and Andrews' (1991) consistent covariance estimators have structures similar to ours. We modify their structures to fit the environments for our test statistics. For this modification, we let $Q_t(\boldsymbol{\theta}, \boldsymbol{\lambda}, \delta) := (Y_t - \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} - \mathbf{H}_t(\delta)'\boldsymbol{\lambda})^2$ and impose the following additional conditions for the consistency of $\widehat{\mathbf{A}}_n(\cdot)$ and $\widehat{\mathbf{B}}_n(\cdot)$:

Assumption 2.7 (Covariance). *(i) $\{Q_t(\cdot)\}$ is a near epoch dependent (NED) process on $\{Y_t, \mathbf{X}_t\}$ of size $-(2\rho - 1)/(\rho - 1)$ uniformly on $\mathbf{\Theta} \times \mathbf{\Lambda} \times \mathbf{\Delta}$; (ii) $\{\nabla_{(\boldsymbol{\theta},\boldsymbol{\lambda})}Q_t(\cdot)\}$ is a near epoch dependent (NED) process on $\{Y_t, \mathbf{X}_t\}$ of size $-(2\rho - 1)/(\rho - 1)$ uniformly on $\mathbf{\Theta} \times \mathbf{\Lambda} \times \mathbf{\Delta}$; and (iii) $\{q_n\}$ is a sequence of integers such that q_n tends to infinity as n tends to infinity with $q_n = O(n^{1/4})$, and $\{\omega_{nj}\}$ is a sequence of positive numbers uniformly bounded by a finite number such that for each $j = 1, 2, \ldots$, ω_{nj} converges to one as n tends to infinity.*

We note that Assumption 2.7(*i* and *ii*) further restricts the scope of DGPs. Nevertheless, many DGPs still satisfy Assumption 2.7. From this respect, Assumption 2.7 does not sacrifice too many DGPs for the applications of our test statistics. The following lemma shows that $\widehat{\mathbf{A}}_n(\delta)$ and $\widehat{\mathbf{B}}_n(\delta)$ are consistent for $\mathbf{A}(\delta)$ and $\mathbf{B}(\delta)$, respectively.

Lemma 2.4. *Given Assumptions 2.1, 2.3–2.5, 2.7, and $\mathbb{H}_0$, if for any $\varepsilon > 0$, there are n_0 and $F \in \mathcal{D}$ with $\mathbb{Q}(F) = 1$ such that for any $n > n_0$, $\mathbb{P}\cdot\mathbb{Q}(\|\widehat{\mathbf{A}}_n(\delta) - \mathbf{A}(\delta)\|_\infty > \varepsilon|F) < \varepsilon$ and $\mathbb{P}\cdot\mathbb{Q}(\|\widehat{\mathbf{B}}_n(\delta) - \mathbf{B}(\delta)\|_\infty > \varepsilon|F) < \varepsilon$, where $\|\cdot\|_\infty$ is the matrix norm.*

Using these covariance estimators, we may redefine our new test statistics as follows:

$$\widehat{T}_n := n\widehat{\boldsymbol{\lambda}}_n(\delta)'[\mathbf{S}\widehat{\mathbf{A}}_n(\delta)^{-1}\widehat{\mathbf{B}}_n(\delta)\widehat{\mathbf{A}}_n(\delta)^{-1}\mathbf{S}']^{-1}\widehat{\boldsymbol{\lambda}}_n(\delta), \tag{7}$$

and

$$\dot{T}_n := \frac{n}{\widehat{\sigma}_n^2}\widehat{\boldsymbol{\lambda}}_n(\delta)'[\mathbf{S}\widehat{\mathbf{A}}_n(\delta)^{-1}\mathbf{S}']^{-1}\widehat{\boldsymbol{\lambda}}_n(\delta), \tag{8}$$

where $\widehat{\sigma}_n^2 := n^{-1}\widehat{\mathbf{U}}'\widehat{\mathbf{U}}$ and $\widehat{\mathbf{U}} := \mathbf{Y} - \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}\mathbf{Y}$. Note that every unknown element in $\widetilde{T}_n(\boldsymbol{\delta})$ and $\dot{T}_n(\boldsymbol{\delta})$ is replaced by its consistent estimator. We call these test statistics the *extreme learning machine tests for neglected nonlinearity* and abbreviate this as the ELMNN tests.

The following theorem states the asymptotic behaviors of the ELMNN tests under the null hypothesis.

Theorem 2.1. *(i) Given Assumptions 2.1, 2.3–2.5, and 2.7, if* $\mathbb{H}_0$ *holds,* $\widehat{T}_n|\boldsymbol{\delta} \overset{A}{\sim} \mathcal{X}_m^2$*; and (ii) Given Assumptions 2.1, 2.3–2.7 if* $\mathbb{H}_0$ *holds,* $\dot{T}_n|\boldsymbol{\delta} \overset{A}{\sim} \mathcal{X}_m^2$*.*

Therefore, the asymptotic null behaviors of $\widehat{T}_n(\boldsymbol{\delta})$ and $\dot{T}_n(\boldsymbol{\delta})$ are now obtained, although their finite sample size properties are expected to be different from Theorem 2.1. We affirm this expectation in the section for Monte Carlo experiments. Here, $\widehat{\sigma}_n^2$ used for $\dot{T}_n(\boldsymbol{\delta})$ is not the only estimator used to define the ELMNN tests. When $\widetilde{\sigma}_n^2 := n^{-1}\widetilde{\mathbf{U}}'\widetilde{\mathbf{U}}$, where $\widetilde{\mathbf{U}} := [\widetilde{U}_1, \widetilde{U}_2, \ldots, \widetilde{U}_n]'$, this statistic is also a consistent estimator for σ_*^2 under $\mathbb{H}_0$ because $\widehat{\boldsymbol{\xi}}_n(\boldsymbol{\delta})$ is consistent for $(\boldsymbol{\theta}_*', \mathbf{0}')'$. Thus, the test statistic associated with $\widetilde{\sigma}_n^2$ still follows the chi-squared distribution under $\mathbb{H}_0$ as well.

3 Monte Carlo Experiments

Testing neglected nonlinearity using ELMNN tests is practically straightforward. In this section, we provide a practical guide to ELMNN tests and conduct Monte Carlo experiments using the guide. The following steps summarize the key steps to apply the ELMNN tests:

- Step 1: Generate $\boldsymbol{\delta} \in \mathbb{R}^{(k+1)\times m}$ from $\boldsymbol{\Delta}$ to follow $\mathbb{Q}$, where m is a pre-specified number such that $k + m + 1 < n$;
- Step 2: Compute the additional regressors by $\mathbf{H}_t(\boldsymbol{\delta}) := [\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_1), \Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_2), \ldots, \Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_m)]$, where for $j = 1, 2, \ldots, m$, $\boldsymbol{\delta}_j$ is the j-th column of $\boldsymbol{\delta}$;
- Step 3: Estimate the unknown coefficient by $\widehat{\boldsymbol{\xi}}_n(\boldsymbol{\delta})$ in Eq. (1);
- Step 4: Estimate the Hessian and asymptotic covariance matrices $\mathbf{A}(\boldsymbol{\delta})$ and $\mathbf{B}(\boldsymbol{\delta})$ by $\widehat{\mathbf{A}}_n(\boldsymbol{\delta})$ and $\widehat{\mathbf{B}}_n(\boldsymbol{\delta})$ in Eqs. (5) and (6), respectively;
- Step 5: Compute the test statistic $\widehat{T}_n$ according to the formula in Eq. (7); if U_t further exhibits conditional homoskedasticity and MDA, compute $\dot{T}_n$ in Eq. (8);
- Step 6: If $\widehat{T}_n$ or $\dot{T}_n$ is greater than the critical value implied by $\mathcal{X}_m^2$, reject the null hypothesis; otherwise, do not reject the null.

As can be seen from this practical guide, the testing procedure using the ELMNN tests is quite straightforward. Only if $\boldsymbol{\delta}$ is generated and $\mathbf{H}_t(\boldsymbol{\delta})$ is defined accordingly, the other steps are computed by standard statistical packages.

The goal of our Monte Carlo experiments is twofold. First, we want to verify the level and power properties of the ELMNN tests. The previous theoretical results do not provide any guidance for the performances of the ELMNN tests particularly when the sample size is finite. Through our Monte Carlo experiments, we aim to validate the theoretical results in the previous section and further investigate the finite sample properties. Second, we aim to examine the properties of the ELMNN tests under different contexts and draw meaningful implications. More specifically, our theory in the previous section does not examine how the ELMNN tests behave as the number of activation functions m increases, although examining this through Monte Carlo experiments is straightforward. We examine this in this section.

Our Monte Carlo experiments examine $\dot{T}_n$ and $\widehat{T}_n$ separately. First, we examine $\dot{T}_n$ by assuming that the researcher has additional information that U_t exhibits conditional homoskedasticity and the MDA property. Second, we examine $\widehat{T}_n$ without any information on U_t.

The environments for our Monte Carlo experiments are specified as follows. First, we consider three DGPs. The first DGP is an MDA with conditional homoskedasticty:

$$Y_t = \varepsilon_t,$$

where $\varepsilon_t \sim$ IID $N(0,1)$. The second DGP we consider is a simple autoregressive process with an order 2, AR(2) process. In other words,

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

where ε_t is the same ε_t as for the first DGP. The third DGP is the following nonlinear process:

$$Y_t = \cos(Y_{t-1}) + \varepsilon_t,$$

where ε_t is the same ε_t as for the first DGP. This DGP is different from the previous DGPs as it has a nonlinear component, so that a linear model is not correctly specified.

Second, we consider the following AR(1) model for the above DGPs:

$$Y_t = \alpha_* + \beta_* Y_{t-1} + U_t.$$

Note that this model is correctly specified for the first DGP. When $\alpha_* = 0$ and $\beta_* = 0$, $\varepsilon_t = U_t$, implying that this linear model is correctly specified, and U_t is an MDA and does not exhibit conditional heteroskedasticity. Thus, we can use the first DGP for our null simulations. Furthermore, as U_t is an MDA and exhibits conditional homoskedasticity, we can use this to examine the null behavior of $\dot{T}_n$. For the second DGP, the AR(1) model is also correctly specified. Although the AR(1) model cannot estimate AR(2) DGP, $E[Y_t|Y_{t-1}]$ is still a linear function of Y_{t-1}, so that the AR(1) model is correctly specified for AR(2) DGP. On the other hand, we can use this to examine the null behavior of $\widehat{T}_n$ as it is dynamically misspecified. Finally, the third DGP is a nonlinear process, so that the AR(1) model is misspecified for the third DGP. We use this to examine the power properties of $\widehat{T}_n$ and $\dot{T}_n$.

There are additional components we specify for our Monte Carlo experiments. First, we consider three different numbers of activation functions. As discussed above, different number of activation functions can lead to different results. We let m be 1, 2, and 3 and examine how the ELMNN tests perform under different values of m. Second, we generate δ from independent uniform distributions. More specifically, $(k+1) \times m$ number of δ's are randomly generated. We let each δ_j be randomly drawn from $U[-10, 10]$ and be independent of another δ_i, where δ_j and δ_i are j-th and i-th random draw of δ. This random draw is exactly the same random draw as given in Huang, Zhu, and Siew (2006). In addition to this experiment, we also let δ_j be randomly drawn from $U[-0.5, 0.5]$ and $U[-1.0, 1.0]$ so that various DGPs for δ can be investigated. Third, according to theorem 2.2 of Huang, Zhu, and Siew (2006), any analytic function can be used to approximate $E[Y_t|Y_{t-1}]$ universally. We use the logistic function based on Huang, Zhu, and Siew (2006), which is also advocated by White (1989). Finally, we consider various sample sizes by letting $n = 50$, 100, 200, 500, and 1,000 for the null behaviors of the ELMNN tests and $n = 50$, 100, 200, 400, 600, and 1,000 for the alternative behaviors. The number of replications are 10,000 and 2,000 under $\mathbb{H}_0$ and $\mathbb{H}_1$, respectively. We iterate more times under $\mathbb{H}_0$ to obtain precise empirical rejection rates.

3.1 Examination of $\dot{T}_n$

We present our null simulation results in Table 1. As mentioned above, the simulations are conducted using different numbers of activation functions, $m = 1, 2,$ and 3; different distributions for δ_j, $U[-0.5, 0.5]$, $U[-1.0, 1.0]$, and $U[-10, 10]$; and 5 different sample sizes, $n = 50$, 100, 200, 500, and 1,000. Therefore, the simulations have to be conducted under 45 different environments. Although this may sound challenging, the required times are not so long. This is mainly because our ELMNN tests are constructed by using the ELMLS estimator, and this is quite a convenient aspect of the ELMNN test when compared with other tests available under similar contexts (e.g., Cho, Cheong, and White (2011), Cho, Ishida, and White (2011, 2013)). Computing the simple LS estimators does not require much time.

<<<<<< Insert Table 1 here. >>>>>>

We summarize the simulation results under $\mathbb{H}_0$ as follows. First, when the number of activation functions m is small as in our environment, few level distortions are observed for the ELMNN tests. In other words, when $m = 1$, 2, and 3, the nominal levels could be exactly delivered by our ELMNN tests. The distribution of δ_j does not matter in obtaining this result. This remarkable aspect is observed even when the sample size is only 50. Figures 1, 2, and 3 also show the P-P plots of the ELMNN tests when the sample size is 200 and $m = 1$, 2, and 3, respectively. In other words, we draw the empirical distribution of $F_m(\dot{T}_n)$, where $F_m(\cdot)$ is the cumulative distribution function of a chi-squared random variable with degrees of freedom m. If $\dot{T}_n \sim \mathcal{X}^2_m$, $F_m(\dot{T}_n)$ has to follow the standard uniform distribution by Rosenblatt (1950). As we can see from Figures 1, 2, and 3, the P-P plots exactly overlap with the 45-degree line, implying that there is no serious level distortion. This affirms the result in Theorem 2.1(i). Even for other sample sizes, we could obtain similar P-P plots, but we do not report them for brevity.

<<<<<< Insert Figure 1 here. >>>>>>

<<<<<< Insert Figure 2 here. >>>>>>

<<<<<< Insert Figure 3 here. >>>>>>

Second, although the empirical rejection rates are very close to the nominal levels as given in Tables 1, we also note that the empirical rejection rates begin to differ from the nominal levels as m increases. We can see that the empirical rejection rates in Figure 3 differ more greatly from the 45-degree line than those in Figure 1. Although they are not reported here, we observed that the empirical rejection rates are quite different from the nominal levels when the sample size is finite and $m = 5$. This level distortion disappears as the sample size increases but very slowly. For conservative researchers, choosing small m is recommended.

Next, we examine the power properties of the ELMNN tests. The simulation results under $\mathbb{H}_1$ are presented in Table 2. We conduct the same experiments as for the null simulations. The only difference is that the number of replications has reduced from 10,000 to 2,000 and the DGP is modified to the alternative DGP associated with the cosine function.

<<<<<< Insert Table 2 here. >>>>>>

We summarize the simulation results as follows. First, as the sample size increases to infinity, the empirical rejection rates tend to one, affirming that our test is a consistent test. Second, the powers of the ELMNN tests have a tendency to increase as m increases. In other words, as we can see from Table 2, the ELMNN tests are most powerful when $m = 3$ and the sample size is moderately large. This fact implies that the ELMNN tests exhibit more power when they are constructed by using more activation functions. Nevertheless, this does not necessarily imply that larger powers are necessarily gained without limit by increasing the number of activation functions. Although we do not report the results when $m = 10$, we could observe that the powers of the ELMNN tests are smaller than the ELMNN tests with $m = 1$.

We now sum up the results of our Monte Carlo experiments for $\dot{T}_n$ as follows. First, the empirical rejection rates are close to the nominal levels when $m = 1$, 2, or 3, but level distortions begin to increase as m increases. Second, the powers of the ELMNN tests tend to increase as m increases but not without limit. Thus, there is a trade-off between the level distortions and the powers of the ELMNN tests. The researcher needs to choose the number of activation functions m with caution and balance this trade-off.

3.2 Examination of $\widehat{T}_n$

We now examine the Monte Carlo experiments using $\widehat{T}_n$. The simulation environments are already expounded above. The only thing we add here is the construction of $\widehat{\mathbf{B}}_n(\delta)$. When estimating this, we use Bartletts's kernel (1950) based on Newey and West (1987). In addition, we truncate the lag at $q_n := \lfloor 2n^{2/9} \rfloor$, which is the truncation order same as that in Newey and West (1987). Specifically, we let

$$\widehat{\mathbf{B}}_n(\delta) := \frac{1}{n}\sum_{t=1}^{n-1} \widehat{U}_t^2 \mathbf{\Gamma}_t(\delta)\mathbf{\Gamma}_t(\delta)' + \sum_{j=1}^{q_n}\left(1 - \frac{j}{q_n+1}\right)\frac{1}{n-1-j}\sum_{t=j+1}^{n-1} \widehat{U}_t\widehat{U}_{t-j}\left(\mathbf{\Gamma}_t(\delta)\mathbf{\Gamma}_{t-j}(\delta)' + \mathbf{\Gamma}_{t-j}(\delta)\mathbf{\Gamma}_t(\delta)'\right),$$

where $\mathbf{\Gamma}_t(\delta) := [1, Y_{t-1}, \Psi([1, Y_{t-1}]\delta_1), \Psi([1, Y_{t-1}]\delta_2), \ldots, \Psi([1, Y_{t-1}]\delta_m)]'$. Under the given conditions, Lemma 2.4 implies that $\widehat{\mathbf{B}}_n(\delta)$ is consistent for $\mathbf{B}(\delta)$.

<<<<<< Insert Table 3 here. >>>>>>

The simulation results under $\mathbb{H}_0$ are presented in Table 3. As we can see from Table 3, the empirical level distortions are relatively larger than $\dot{T}_n$ when the sample size is small. This is not due to the serial correlation in U_t but due to the covariance estimation, which accommodates serial correlation and conditional heteroskedasticity. Even when $\widehat{T}_n$ is applied for the first DGP, we can observe similar patterns.

Nevertheless, the level distortions begin to disappear as the sample size increases. When the number of observations is close to 1,000, the level distortions almost disappear. This aspect reinforces the consequences in Theorem 2.1(*ii*). Although the finite sample size performances of $\widehat{T}_n$ is disappointing, the ELMNN test provides asymptotically precise level performances for a large sample size. Figures 4, 5, and 6 show the P-P plots of the ELMNN tests when δ_j is randomly drawn from $U[-1.0, 1.0]$. As we can from these figures, the ELMNN tests with larger sample sizes are closer to the 45-degree line than the other ELMNN tests.

<<<<<< Insert Figure 4 here. >>>>>>

<<<<<< Insert Figure 5 here. >>>>>>

<<<<<< Insert Figure 6 here. >>>>>>

Another aspect of the ELMNN tests is that the level distortion increases as the number of activation functions m increases. As we can see from Table 3 and Figures 4, 5, and 6, the empirical distributions are quite different when $m = 3$, even if level distortions decrease by increasing the sample size. This implies that the researcher needs to choose m carefully when applying the ELMNN tests. If the researcher wants to minimize the level distortion, choosing $m = 1$ is recommended.

<<<<<< Insert Table 4 here. >>>>>>

Next, we examine the power properties of the ELMNN tests. The simulation results are presented in Table 4. As we can see from Table 4, the ELMNN tests are consistent against neglected nonlinearity. In other words, they are consistent test statistics even when the prediction error U_t exhibits conditional heteroskedasticity and serial correlation as expected by our discussions. Furthermore, the power of the ELMNN test increases when the number of activation functions increases. This is the same trade-off observed from $\dot{T}_n$, and the researcher needs to balance this trade-off when choosing m as well.

4 Conclusion

In this study, we introduce test statistics testing for neglected nonlinearity. The primary motivation of introducing new tests is for the use of ELMs. According to theorem 2 of Huang, Zhu, and Siew (2006), the ELMs can estimate the conditional mean equation consistently and universally. We exploit this convenient feature to define our test statistic and call it the ELMNN test.

The main contribution of the ELMNN test is in its convenience and efficacy in computing unknown parameters. Because they are present as the linear coefficients of explanatory variables and activation functions, the only necessary computational process is the LS estimation. Thus, its application is straightforward and can be widely applied under the model and DGP conditions provided in the current study. In particular, we examine the ELMNN tests in the context of stationary time-series data, and the researcher can use our tests as quick diagnostic statistics complementing the computational burdens of other tests available in the literature.

Another contribution of the ELMNN test is that it has a standard asymptotic null distribution, a chi-squared distribution, so that it can also be easily applied for testing the linearity. This aspect is different from most popular statistics defined for the same goal, and this fact makes them inconvenient to apply.

Finally, the ELMNN test does not requires the researcher to choose a particular alternative direction as for the Wald ELM test of Cho and White (2011). This aspect is mainly due to the fact that the ELMNN test is based upon the universal approximation feature.

Nevertheless, applying the ELMNN tests requires cautions from the researcher. Finite sample level distortions can be relatively large when a large number of activation functions are employed. In particular, when the serial correlation and/or conditional heteroskedasticity structure of prediction errors is unknown, the level distortions can be immense. For reducing level distortions, choosing a small number of activation functions is recommended. On the other hand, the power of the ELMNN test is relatively large when multiple activation functions are employed.

5 Appendix: Proofs

We first provide preliminary lemmas and their proof before proving the main claims. For notational simplicity, we let $H_t(\delta_j) := \Psi(\widetilde{\mathbf{X}}_t'\delta_j)$.

Lemma 5.1. *Given Assumptions 2.1, 2.3, and 2.4,*

$$\sup_{(\theta,\delta,\lambda)\in\mathbf{\Theta}\times\mathbf{\Delta}\times\mathbf{\Lambda}}\left|n^{-1}\sum_{t=1}^{n}(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta}-\mathbf{H}_t(\delta)'\boldsymbol{\lambda})^2-E[(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta}-\mathbf{H}_t(\delta)'\boldsymbol{\lambda})^2]\right|\overset{\mathbb{P}}{\to}0$$

as n tends to infinity, where $\mathbf{\Theta} := \mathbf{A}\times\mathbf{B}$.

Lemma 5.2. *Given Assumptions 2.1, 2.3, and 2.4, (i)*

$$\sup_{(\delta_i,\delta_j)\in\mathbf{T}\times\mathbf{T}}\left|\frac{1}{n}\sum_{t=1}^{n}H_t(\delta_i)H_t(\delta_j)-E[H_t(\delta_i)H_t(\delta_j)]\right|\overset{\mathbb{P}}{\to}0;$$

and (ii) for $i=1,2,\ldots,k+1$,

$$\sup_{\delta_j\in\mathbf{T}}\left|\frac{1}{n}\sum_{t=1}^{n}\widetilde{X}_{t,i}H_t(\delta_j)-E[\widetilde{X}_{t,i}H_t(\delta_j)]\right|\overset{\mathbb{P}}{\to}0.$$

Proof of Lemma 5.1: We now note that

$$\begin{aligned}
&\sup_{(\theta,\delta,\lambda)}\left|n^{-1}\sum_{t=1}^{n}(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta}-\mathbf{H}_t(\delta)'\boldsymbol{\lambda})^2-E[(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta}-\mathbf{H}_t(\delta)'\boldsymbol{\lambda})^2]\right| \qquad (9)\\
&\leq\sup_{\theta}\left|\frac{1}{n}\sum_{t=1}^{n}(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})^2-E[(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})^2]\right|\\
&\quad+\sup_{(\delta,\lambda)}\left|\frac{1}{n}\sum_{t=1}^{n}(\mathbf{H}_t(\delta)'\boldsymbol{\lambda})^2-E[(\mathbf{H}_t(\delta)'\boldsymbol{\lambda})^2]\right|\\
&\quad+\sup_{(\theta,\delta,\lambda)}2\left|\frac{1}{n}\sum_{t=1}^{n}(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})(\mathbf{H}_t(\delta)'\boldsymbol{\lambda})-E[(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})(\mathbf{H}_t(\delta)'\boldsymbol{\lambda})]\right|.
\end{aligned}$$

We examine each element in the RHS one-by-one. We note that

$$\sup_{\boldsymbol{\theta}}\left|\frac{1}{n}\sum_{t=1}^{n}(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})^2-E[(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})^2]\right| \leq \left|\frac{1}{n}\sum_{t=1}^{n}Y_t^2-E[Y_t^2]\right|$$
$$+\sup_{\boldsymbol{\theta}}\left|\frac{1}{n}\sum_{t=1}^{n}(\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})^2-E[(\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})^2]\right|+2\sup_{\boldsymbol{\theta}}\left|\frac{1}{n}\sum_{t=1}^{n}Y_t(\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})-E[Y_t(\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})]\right|.$$

Given this, applying the ergodic theorem implies that $\left|\frac{1}{n}\sum_{t=1}^{n}Y_t^2-E[Y_t^2]\right| \xrightarrow{\mathbb{P}} 0$, and $|\frac{1}{n}\sum_{t=1}^{n}(\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})^2 - E[(\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})^2]| = |\sum_{j=1}^{k+1}\sum_{i=1}^{k+1}\frac{1}{n}\sum_{t=1}^{n}(\widetilde{X}_{t,j}\widetilde{X}_{t,i}-E[\widetilde{X}_{t,j}\widetilde{X}_{t,i}])\theta_j\theta_i|$, where $n^{-1}\sum_{t=1}^{n}(\widetilde{X}_{t,j}\widetilde{X}_{t,i}-E[\widetilde{X}_{t,j}\widetilde{X}_{t,i}]) \xrightarrow{\mathbb{P}} 0$ by the ergodic theorem for each $j,i=1,2,\ldots,k+1$. Furthermore, θ_j and θ_i are the elements of the compact parameter space $\boldsymbol{\Theta}$, so that $\sup_{\boldsymbol{\theta}}|n^{-1}\sum_{t=1}^{n}(\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})^2-E[(\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})^2]| \xrightarrow{\mathbb{P}} 0$. Finally, $|\frac{1}{n}\sum_{t=1}^{n}Y_t(\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})-E[Y_t(\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})]|$ $=|\sum_{j=1}^{k+1}(n^{-1}\sum_{t=1}^{n}\widetilde{X}_{t,j}Y_t-E[\widetilde{X}_{t,j}Y_t])\theta_j|$. Using the moment condition and Cauchy-Schwarz's inequality, we can obtain $n^{-1}\sum_{t=1}^{n}\widetilde{X}_{t,j}Y_t \xrightarrow{\mathbb{P}} E[\widetilde{X}_{t,j}Y_t]$. In addition, θ_j is the element of the compact parameter space, so that $|\frac{1}{n}\sum_{t=1}^{n}Y_t(\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})-E[Y_t(\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})]| \xrightarrow{\mathbb{P}} 0$. Therefore, it now follows that

$$\sup_{\boldsymbol{\theta}}\left|\frac{1}{n}\sum_{t=1}^{n}(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})^2-E[(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})^2]\right| \xrightarrow{\mathbb{P}} 0. \tag{10}$$

Next, we examine the second term in the RHS of Eq. (9). We note that $\sum_{i=1}^{m}\sum_{j=1}^{m} n^{-1}\sum_{t=1}^{n}\{H_t(\delta_i)$ $H_t(\delta_j)-E[H_t(\delta_i)H_t(\delta_j)]\}\lambda_i\lambda_j$, and applying Ranga-Rao's ULLN shows that for each $i,j=1,2,\ldots,m$,

$$\sup_{\delta_i,\delta_j}|n^{-1}\sum_{t=1}^{n}\{H_t(\delta_i)H_t(\delta_j)-E[H_t(\delta_i)H_t(\delta_j)]\}| \xrightarrow{\mathbb{P}} 0$$

using Assumption 2.4(*iii*). Furthermore, λ_i and λ_j are the elements of the compact parameter space $\boldsymbol{\Lambda}$. Thus, it follows that

$$\sup_{(\delta,\lambda)}\left|\frac{1}{n}\sum_{t=1}^{n}(\mathbf{H}_t(\delta)'\boldsymbol{\lambda})^2-E[(\mathbf{H}_t(\delta)'\boldsymbol{\lambda})^2]\right| \xrightarrow{\mathbb{P}} 0. \tag{11}$$

Finally, we examine the final term in the RHS of Eq. (9). We note that $|\frac{1}{n}\sum_{t=1}^{n}(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})(\mathbf{H}_t(\delta)'\boldsymbol{\lambda}) - E[(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})(\mathbf{H}_t(\delta)'\boldsymbol{\lambda})]| \leq |\sum_{j=1}^{m}\{n^{-1}\sum_{t=1}^{n}Y_tH_t(\boldsymbol{\delta}_j)-E[Y_tH_t(\boldsymbol{\delta}_j)]\}\,\lambda_j|+|\sum_{j=1}^{m}\sum_{i=1}^{k+1}\{n^{-1}\sum_{t=1}^{n}\widetilde{X}_{t,i}H_t(\boldsymbol{\delta}_j)- E[\widetilde{X}_{t,i}H_t(\boldsymbol{\delta}_j)]\}\lambda_j\theta_i|$. Here, we note that for each $i=1,2,\ldots,m$ and $j=1,2,\ldots,k+1$, $\sup_{\delta_j}|n^{-1}\sum_{t=1}^{n}Y_t$ $H_t(\boldsymbol{\delta}_j)-E[Y_tH_t(\boldsymbol{\delta}_j)]| \xrightarrow{\mathbb{P}} 0$, and $\sup_{\delta_j}|\sum_{j=1}^{m}\sum_{i=1}^{k+1}\{n^{-1}\sum_{t=1}^{n}\widetilde{X}_{t,i}H_t(\boldsymbol{\delta}_j)-E[\widetilde{X}_{t,i}H_t(\boldsymbol{\delta}_j)]\}\lambda_j\theta_i| \xrightarrow{\mathbb{P}} 0$ according to Assumption 2.4(*iii*). In addition, λ_j and θ_i are the elements of the compact parameter spaces. Thus, it follows that

$$\sup_{(\theta,\delta,\lambda)} 2\left|\frac{1}{n}\sum_{t=1}^{n}(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})(\mathbf{H}_t(\delta)'\boldsymbol{\lambda})-E[(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta})(\mathbf{H}_t(\delta)'\boldsymbol{\lambda})]\right| \xrightarrow{\mathbb{P}} 0. \tag{12}$$

We now combine the given inequalities in (10), (11), and (12), and this shows that

$$\sup_{(\boldsymbol{\theta},\delta,\boldsymbol{\lambda})}\left|n^{-1}\sum_{t=1}^{n}(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta}-\mathbf{H}_t(\delta)'\boldsymbol{\lambda})^2-E[(Y_t-\widetilde{\mathbf{X}}_t'\boldsymbol{\theta}-\mathbf{H}_t(\delta)'\boldsymbol{\lambda})^2]\right|\xrightarrow{\mathbb{P}}0,$$

as desired. ■

Proof of Lemma 5.2: (*i*) We already proved this while proving Eq. (11). (*ii*) This was also proved while proving Eq. (12), and this completes the proof. ■

We now prove the main claims of this paper.

Proof of Lemma 2.1: For this proof, we note that

$$(\widehat{\alpha}_n(\delta),\widehat{\boldsymbol{\beta}}_n(\delta),\widehat{\boldsymbol{\lambda}}_n(\delta)) := \arg\min_{\alpha,\beta,\lambda} n^{-1}\sum_{t=1}^{n}(Y_t-\alpha-\mathbf{X}_t'\boldsymbol{\beta}-\mathbf{H}_t(\delta)'\boldsymbol{\lambda})^2,$$

and Lemma 5.1 implies that the ULLN holds. Further, $\mathbb{P}$ is independent of $\mathbb{Q}$, so that $(\widehat{\alpha}_n(\delta),\widehat{\boldsymbol{\beta}}_n(\delta),\widehat{\boldsymbol{\lambda}}_n(\delta))\xrightarrow{\mathbb{P}}$ $\arg\min_{\alpha,\beta,\lambda}E[(Y_t-\alpha-\mathbf{X}_t'\boldsymbol{\beta}-\mathbf{H}_t(\delta)'\boldsymbol{\lambda})^2|\delta]$. We can further simplify the RHS. We note that

$$\begin{aligned}&E[(Y_t-\alpha-\mathbf{X}_t'\boldsymbol{\beta}-\mathbf{H}_t(\delta)'\boldsymbol{\lambda})^2|\delta]\\&\quad=E[(Y_t-E[Y_t|\mathbf{X}_t])^2]|\delta]+E[(E[Y_t|\mathbf{X}_t]-\alpha-\mathbf{X}_t'\boldsymbol{\beta}-\mathbf{H}_t(\delta)'\boldsymbol{\lambda})^2|\delta]\\&\qquad+2E[(Y_t-E[Y_t|\mathbf{X}_t])(E[Y_t|\mathbf{X}_t]-\alpha-\mathbf{X}_t'\boldsymbol{\beta}-\mathbf{H}_t(\delta)'\boldsymbol{\lambda})|\delta],\end{aligned}$$

and applying the law of iterated expectation yields that

$$\begin{aligned}&E[(Y_t-E[Y_t|\mathbf{X}_t])(E[Y_t|\mathbf{X}_t]-\alpha-\mathbf{X}_t'\boldsymbol{\beta}-\mathbf{H}_t(\delta)'\boldsymbol{\lambda})|\delta]\\&\quad=E[E[Y_t-E[Y_t|\mathbf{X}_t]|\mathbf{X}_t,\delta](E[Y_t|\mathbf{X}_t]-\alpha-\mathbf{X}_t'\boldsymbol{\beta}-\mathbf{H}_t(\delta)'\boldsymbol{\lambda})|\delta],\end{aligned}$$

and $E[Y_t-E[Y_t|\mathbf{X}_t]|\mathbf{X}_t,\delta]=E[Y_t-E[Y_t|\mathbf{X}_t]|\mathbf{X}_t]$ because $\mathbf{X}_t$ and δ are independent, so that $E[Y_t-E[Y_t|\mathbf{X}_t]|$ $\mathbf{X}_t,\delta]=0$. This now implies that

$$\begin{aligned}&E[(Y_t-\alpha-\mathbf{X}_t'\boldsymbol{\beta}-\mathbf{H}_t(\delta)'\boldsymbol{\lambda})^2|\delta]\\&\quad=E[(Y_t-E[Y_t|\mathbf{X}_t])^2]|\delta]+E[(E[Y_t|\mathbf{X}_t]-\alpha-\mathbf{X}_t'\boldsymbol{\beta}-\mathbf{H}_t(\delta)'\boldsymbol{\lambda})^2|\delta],\end{aligned}$$

so that it follows that

$$\begin{aligned}&\underset{\alpha,\boldsymbol{\beta},\boldsymbol{\lambda}}{\arg\min} E[\,(Y_t - \alpha - \mathbf{X}_t'\boldsymbol{\beta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})^2\,|\boldsymbol{\delta}]\\ &\qquad\qquad = \underset{\alpha,\boldsymbol{\beta},\boldsymbol{\lambda}}{\arg\min} E[(E[Y_t|\mathbf{X}_t] - \alpha - \mathbf{X}_t'\boldsymbol{\beta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})^2|\boldsymbol{\delta}].\end{aligned}$$

The desired result now follows from this. ■

Proof of Lemma 2.2: (*i*) We first consider the null behavior of $\widehat{\boldsymbol{\lambda}}_n(\boldsymbol{\delta})$. If we decompose Y_t into $E[Y_t|\mathbf{X}_t] + U_t$, it trivially follows that

$$\begin{aligned}&E[\mathbf{H}_t(\boldsymbol{\delta})\{Y_t - \widetilde{\mathbf{X}}_t'E[\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t']^{-1}E[\widetilde{\mathbf{X}}_tY_t]\}|\boldsymbol{\delta}]\\ &\qquad =E[\mathbf{H}_t(\boldsymbol{\delta})\{E[Y_t|\mathbf{X}_t] - \widetilde{\mathbf{X}}_t'E[\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t']^{-1}E[\widetilde{\mathbf{X}}_tE[Y_t|\mathbf{X}_t]]\}|\boldsymbol{\delta}]\\ &\qquad =E[\mathbf{H}_t(\boldsymbol{\delta})\{\widetilde{\mathbf{X}}_t'\boldsymbol{\theta}_* - \widetilde{\mathbf{X}}_t'E[\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t']^{-1}E[\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t'\boldsymbol{\theta}_*]\}|\boldsymbol{\delta}] = \mathbf{0}\end{aligned}$$

by noting that $E[Y_t|\mathbf{X}_t] = \widetilde{\mathbf{X}}_t'\boldsymbol{\theta}_*$ under $\mathbb{H}_0$. We next consider the probability limit of $\widehat{\boldsymbol{\lambda}}_n(\boldsymbol{\delta})$ under $\mathbb{H}_1$. By corollary 3.9 of Stinchcombe and White (Stinchcombe 1998), if we let $F \in \mathcal{D}$ such that for each $\boldsymbol{\tau} \in F$, $E[\mathbf{H}_t(\boldsymbol{\delta})\{Y_t - \widetilde{\mathbf{X}}_t'E[\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t']^{-1}E[\widetilde{\mathbf{X}}_tY_t]\}|\boldsymbol{\delta}] = \mathbf{0}$, then $\mu(F) = 0$. We also note that $\mathbb{Q}$ is absolutely continuous with respect to μ, implying that $\mathbb{Q}(F) = 0$.

(*ii*) The desired result is directly implied by Lemma 2.2(*i*). ■

Proof of Lemma 2.3: (*i*) We partition our proof into three parts. First, for each $j = 1, 2, \ldots, k+1$, we derive the weak limit of the first $k+1$ elements of $\mathbf{Z}_n(\cdot)$, $n^{-1/2}\sum_{t=1}^n U_t\widetilde{X}_{t,j}$. We note that $E[\widetilde{X}_{t,j}U_t] = E[\widetilde{X}_{t,j}E[U_t|\widetilde{\mathbf{X}}_t]] = 0$. Given this, Bradley's (Bradley 1985) CLT can be applied, which is traced from Ibragimov. By Assumption 2.5, it trivially follows that $E[|U_t\widetilde{X}_{t,j}|^{2+\eta}] \leq E[|M_t^2|^{2+\eta}] \leq E[M_t^{4+2\eta}] < \infty$. We also note that for each τ, $2\alpha_\tau \leq \beta_\tau$, where α_τ is the strong-mixing coefficient. Thus, Assumption 2.1 implies that $2\sum_{\tau=1}^{\infty}\tau^{2\rho/(\rho-1)}\alpha_\tau < \infty$. This inequality also implies that for large τ, there are $\gamma > 0$ and $\varepsilon > 0$ such that $\alpha_\tau \leq \gamma\tau^{(-3\rho+1)/(\rho-1)-\varepsilon}$, implying further that

$$\sum_{\tau=1}^{\infty}\alpha_\tau^{\frac{\eta}{2+\eta}} \leq \kappa\sum_{\tau=1}^{\infty}\tau^{\frac{6-\varepsilon\eta-\frac{2\eta}{\rho-1}}{2+\eta}-3},$$

where $\kappa := \gamma^{\eta/(2+\eta)}$. Furthermore, $\eta \geq 2(\rho-1)$ by Assumption 2.5, and using this shows that $\{6 - \varepsilon\eta - 2\eta/(\rho-1)\}/\{2+\eta\} - 3 < -1$, so that $\sum_{\tau=1}^{\infty}\alpha_\tau^{\frac{\eta}{2+\eta}} < \infty$. It now follows that for each $j = 1, \ldots, k+1$, $n^{-1/2}\sum_{t=1}^n U_t\widetilde{X}_{t,j} \overset{\text{A}}{\sim} N(0, \text{avar}(n^{-1/2}\sum_{t=1}^n U_t\widetilde{X}_{t,j}))$ by Bradley's (Bradley 1985) theorem 0.

We next derive the weak limit of $n^{-1/2}\sum_{t=1}^n \Psi(\mathbf{X}_t'(\cdot))U_t$. Indeed, we can apply the proof of lemma 2 in Cho and White's (ChoWhite 2011). As $\Psi(\cdot)$ is an analytic function, we can apply the Lipschitz's condition, so that

$$|U_t\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\tau}) - U_t\Psi(\widetilde{\mathbf{X}}_t'\widetilde{\boldsymbol{\tau}})| \leq |U_t|M_t\|\boldsymbol{\tau}-\widetilde{\boldsymbol{\tau}}\| \leq M_t^2\|\boldsymbol{\tau}-\widetilde{\boldsymbol{\tau}}\|$$

by Assumption 2.5, and this implies that

$$E\left[\sup_{\|\boldsymbol{\tau}-\widetilde{\boldsymbol{\tau}}\|<\nu} |U_t\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\tau}) - U_t\Psi(\widetilde{\mathbf{X}}_t'\widetilde{\boldsymbol{\tau}})|^{2+\eta}\right]^{\frac{1}{2+\eta}} \leq E[M_t^{4+2\eta}]^{\frac{1}{2+\eta}}\nu.$$

This inequality and theorem 1 of Doukhan, Massrt, and Rio (Doukhan 1995) now show that Ossiander's $L^{2+\eta}$ entropy is finite, and $\sqrt{n}n^{-1/2}\sum_{t=1}^{n}\Psi(\widetilde{\mathbf{X}}_t'(\cdot))U_t$ is tight from the beta-mixing condition in Assumption 2.1: $\sum_{\tau=1}^{\infty}\tau^{1/(\rho-1)}\beta_t < \infty$. Thus, it now follows that $\{\sqrt{n}n^{-1/2}\sum_{t=1}^{n}\Psi(\widetilde{\mathbf{X}}_t'(\cdot))U_t\}$ is tight.

Finally, we note that the positive-definite covariance matrix condition in Assumption 2.5 imposes that for each $\boldsymbol{\tau}$ and $\widetilde{\boldsymbol{\tau}}$, the asymptotic covariance matrix of $\mathcal{Z}(\boldsymbol{\tau})$ and $\mathcal{Z}(\widetilde{\boldsymbol{\tau}})$ is positive definite. Therefore, for any $\ell \in \mathbb{N}$, the finite dimensional multivariate CLT holds for $n^{-1/2}\sum_{t=1}^{n}[U_t\widetilde{\mathbf{X}}', U_t\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\tau}_1), U_t\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\tau}_2), \ldots, U_t \Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\tau}_\ell)]'$ by the Cramèr-Wold device. The multivariate FCLT for this process follows from this fact and the two facts proved above.

(*ii*) By Assumption 2.3, $\mathbb{P}$ and $\mathbb{Q}$ are independent, so that the desired result follows from this and the result in (*i*).

(*iii*) This holds by the continuous mapping theorem given the result in (*ii*). ■

Proof of Lemma 2.4: We first prove the consistence of $\widehat{\mathbf{A}}_n(\delta)$. We note that $n^{-1}\sum_{t=1}^{n}\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t' \overset{\mathbb{P}}{\to} E[\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t']$ by the ergodic theorem. In addition, Lemma 5.2(*i*) implies that for each $i = 1, 2, \ldots, k+1$, $\sup_{\delta_j\in\mathbf{T}}|n^{-1}\sum_{t=1}^{n}\widetilde{X}_{t,j} H_t(\delta_j) - E[\widetilde{X}_{t,j}H_t(\delta_j)]| \overset{\mathbb{P}}{\to} 0$, so that $\sup_{\delta_j\in\mathbf{T}}|n^{-1}\sum_{t=1}^{n}\widetilde{\mathbf{X}}_t'\mathbf{H}_t(\delta_j) - E[\widetilde{\mathbf{X}}_t'\mathbf{H}_t(\delta_j)]| \overset{\mathbb{P}}{\to} 0$. Finally, Lemma 5.2(*ii*) implies that $\sup_{(\delta_j,\delta_i)\in\mathbf{T}\times\mathbf{T}}|n^{-1}\sum_{t=1}^{n}H_t(\delta_j)H_t(\delta_i) - E[H_t(\delta_j)H_t(\delta_i)]| \overset{\mathbb{P}}{\to} 0$, so that $\widehat{\mathbf{A}}_n(\cdot)$ obeys the ULLN. Given this, as δ is drawn from $\mathbb{Q}$ independent of $\mathbb{P}$, we can consider $\mathbf{A}(\delta)$ to be the limit of $\widehat{\mathbf{A}}_n(\delta)$. Thus, for any $\varepsilon > 0$, there are $F \in \mathcal{D}$ with $\mathbb{Q}(F) = 1$ and n_0 such that if $n > n_0$, $\mathbb{P}\cdot\mathbb{Q}(\|\widehat{\mathbf{A}}_n(\delta) - \mathbf{A}(\delta)\|_\infty > \varepsilon|F) < \varepsilon$.

Next, we prove the consistence of $\widehat{\mathbf{B}}_n(\delta)$. We show this by verifying the sufficient conditions for theorem 6.8 in Gallant and White (Gallant 1988): DG, OP′, MX′, SM, DM″, NE‴, ID′, TL, and WT. Then, the desired result follows by their theorem 6.8.

First, DG trivially holds by Assumption 2.1.

Second, given the definition of $Q_t(\cdot)$, if we let x and $Q_t(\cdot)$ be $g_n(x)$ and $q_t(\cdot)$ of Gallant and White (Gallant 1988), respectively, then the OP′ also holds.

Third, Assumption 2.1 implies that for large τ, there are π and ϵ such that $\beta_\tau \leq \pi\tau^{(-3\rho+1)/(\rho-1)-\epsilon}$. Therefore,

$$\beta_\tau\tau^{\frac{2r}{r-2}+\varepsilon} \leq \pi\tau^{-1+\varepsilon-\epsilon}$$

by letting $r = 2\rho$. Therefore, if we let ε be a number between 0 and $1+\epsilon$,

$$\beta_\tau = O(\tau^{-\frac{2r}{r-2}-\varepsilon}).$$

We also note that $2\alpha_\tau \leq \beta_\tau$ for every τ, so that

$$\alpha_\tau = O(\tau^{-\frac{2r}{r-2}-\varepsilon}).$$

In other words, for $r = 2\rho$ and $\rho > 1$, $\{Y_t, \mathbf{X}_t\}$ is an α-mixing sequence of size $-2r/(r-2)$. This now satisfies the MX$'$ of Gallant and White (Gallant 1988).

Fourth, we note that $Q_t(\cdot)$, $\nabla_{(\theta,\lambda)} Q_t(\cdot)$, $\nabla^2_{(\theta,\lambda)} Q_t(\cdot)$ are Lipschitz-L_1 on $\mathbf{\Theta} \times \mathbf{\Lambda} \times \mathbf{\Delta}$ by Assumptions 2.4 and 2.5. More specifically, we note that

$$Q_t(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\delta}) = (Y_t - \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})^2,$$

$$\nabla_{(\theta,\lambda)} Q_t(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\delta}) = -2 \begin{bmatrix} (Y_t - \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})\widetilde{\mathbf{X}}_t \\ (Y_t - \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})\widetilde{\mathbf{H}}_t(\boldsymbol{\delta}) \end{bmatrix},$$

and

$$\nabla^2_{(\theta,\lambda)} Q_t(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\delta}) = 2 \begin{bmatrix} \widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t' & \widetilde{\mathbf{X}}_t\mathbf{H}_t(\boldsymbol{\delta})' \\ \mathbf{H}_t(\boldsymbol{\delta})\widetilde{\mathbf{X}}_t' & \mathbf{H}_t(\boldsymbol{\delta})\mathbf{H}_t(\boldsymbol{\delta})' \end{bmatrix}.$$

As we show below, the modulus of each element of $\nabla_{(\theta,\lambda)} Q_t(\cdot)$ and $\nabla^2_{(\theta,\lambda)}\ Q_t(\cdot)$ is dominated by cM_t^2 for some finite $c > 0$. Therefore, Assumption 2.5(i) implies that $Q_t(\cdot)$ and each element of $\nabla^2_{(\theta,\lambda)} Q_t(\cdot)$ is Lipschitz-L_1 on $\mathbf{\Theta} \times \mathbf{\Lambda} \times \mathbf{\Delta}$.

In addition, we verify that each element of $\nabla^2_{(\theta,\lambda)} Q_t(\cdot)$ is Lipschitz-L_1 on $\mathbf{\Theta} \times \mathbf{\Lambda} \times \mathbf{\Delta}$. ($i$) We examine a representative element of $\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t'$: $\widetilde{X}_{t,j}\widetilde{X}_{t,i}$, where $\widetilde{X}_{t,j}$ and $\widetilde{X}_{t,i}$ are the j-th and i-th element of $\widetilde{\mathbf{X}}_t\widetilde{\mathbf{X}}_t'$, respectively. We note that this is not a function of $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\delta})$, so that it is trivially Lipschitz-L_1 on $\mathbf{\Theta} \times \mathbf{\Lambda} \times \mathbf{\Delta}$. ($ii$) We next examine a representative element of $\widetilde{\mathbf{X}}_t\mathbf{H}_t(\boldsymbol{\delta})'$: $\widetilde{X}_{t,j}\Psi(\widetilde{\mathbf{X}}'\boldsymbol{\delta}_h)$, where $h \in \{1, 2, \ldots, m\}$. Here, we note that

$$|\widetilde{X}_{t,j}\Psi(\widetilde{\mathbf{X}}'\boldsymbol{\delta}_h) - \widetilde{X}_{t,j}\Psi(\widetilde{\mathbf{X}}'\bar{\boldsymbol{\delta}}_h)| \leq |\widetilde{X}_{t,j}| \sum_{i=1}^{k+1} \sup_{\boldsymbol{\tau}\in\mathbf{T}} \left| \frac{\partial}{\partial \tau_i} \Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\tau}) \right| \|\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}\| \leq (k+1)M_t^2 \|\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}\|$$

by Assumption 2.5(v). Assumption 2.5(i) also implies that $E[M_t^2] < \infty$, so that $\widetilde{X}_{t,j}\Psi(\widetilde{\mathbf{X}}'\boldsymbol{\delta}_h)$ as a function of $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\delta})$ is Lipschitz-L_1 on $\mathbf{\Theta} \times \mathbf{\Lambda} \times \mathbf{\Delta}$. ($iii$) We finally examine a representative element of $\mathbf{H}_t(\boldsymbol{\delta})\mathbf{H}_t(\boldsymbol{\delta})'$:

$\Psi(\widetilde{\mathbf{X}}'\boldsymbol{\delta}_h)\Psi(\widetilde{\mathbf{X}}'\boldsymbol{\delta}_\ell)$, where $h,\ell \in \{1,2,\ldots,m\}$. We note that

$$|\Psi(\widetilde{\mathbf{X}}'\boldsymbol{\delta}_h)\Psi(\widetilde{\mathbf{X}}'\boldsymbol{\delta}_\ell) - \Psi(\widetilde{\mathbf{X}}'\bar{\boldsymbol{\delta}}_h)\Psi(\widetilde{\mathbf{X}}'\bar{\boldsymbol{\delta}}_\ell)| \le \sum_{i=1}^{k+1}\sup_{\tau\in\mathbf{T}}\left|\frac{\partial}{\partial\tau_i}\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\tau})\right|\sup_{\tau\in\mathbf{T}}\left|\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\tau})\right|(\|\boldsymbol{\delta}_i-\bar{\boldsymbol{\delta}}_i\|+\|\boldsymbol{\delta}_h-\bar{\boldsymbol{\delta}}_h\|)$$

which is bounded by $(k+1)M_t^2(\|\boldsymbol{\delta}_i-\bar{\boldsymbol{\delta}}_i\|+\|\boldsymbol{\delta}_h-\bar{\boldsymbol{\delta}}_h\|)$ according to Assumption 2.5(*iv* and v). Assumption 2.5(*i*) also implies that $E[M_t^2] < \infty$, so that $\Psi(\widetilde{\mathbf{X}}'\boldsymbol{\delta}_i)\Psi(\widetilde{\mathbf{X}}'\boldsymbol{\delta}_h)$ as a function of $(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\delta})$ is Lipschitz-L_1 on $\boldsymbol{\Theta}\times\boldsymbol{\Lambda}\times\boldsymbol{\Delta}$. From these verifications, we can conclude that each element of $\nabla^2_{(\theta,\lambda)}Q_t(\cdot)$ is Lipschitz-L_1 on $\boldsymbol{\Theta}\times\boldsymbol{\Lambda}\times\boldsymbol{\Delta}$. This verifies the SM of Gallant and White (Gallant 1988).

Fifth, we examine each element of $Q_t(\cdot)$, $\nabla_{(\theta,\lambda)}Q_t(\cdot)$, and $\nabla^2_{(\theta,\lambda)}Q_t(\cdot)$ to show the DM$''$ of Gallant and White (Gallant 1988). (*i*) We now note that the null hypothesis implies that $|Y_t| = |\widetilde{\mathbf{X}}_t\boldsymbol{\theta}_* + U_t| \le |\widetilde{\mathbf{X}}_t\boldsymbol{\theta}_*| + |U_t|$, so that for some c_1, $|Y_t| \le c_1 M_t$ by Assumption 2.5(*ii* and *iii*). Furthermore,

$$(Y_t - \sum_{i=1}^{k+1}\theta_i\widetilde{X}_{t,i} - \sum_{i=1}^{m}\lambda_i\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_i))^2 \le (|Y_t| + \sum_{i=1}^{k+1}|\theta_i||\widetilde{X}_{t,i}| + \sum_{i=1}^{m}|\lambda_i||\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_i)|)^2$$

uniformly on $\boldsymbol{\Theta}\times\boldsymbol{\Lambda}\times\boldsymbol{\Delta}$, where $\widetilde{X}_{t,i}$ is the i-th element of $\widetilde{\mathbf{X}}_t$. We also note that

$$(|Y_t| + \sum_{i=1}^{k+1}|\theta_i||\widetilde{X}_{t,i}| + \sum_{i=1}^{m}|\lambda_i||\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_i)|)^2 \le \sup_{\theta,\lambda,\delta}(c_1 + \sum_{i=1}^{k+1}|\theta_i| + \sum_{i=1}^{m}|\lambda_i|)M_t^2.$$

We note that for some $c_2 > 0$, the RHS is bounded by $c_2M_t^2$ from the fact that $\boldsymbol{\Theta}\times\boldsymbol{\Lambda}\times\boldsymbol{\Delta}$ is a compact parameter space. Thus, $Q_t(\cdot)$ is $2r$-dominated on $\boldsymbol{\Theta}\times\boldsymbol{\Lambda}\times\boldsymbol{\Delta}$ for some $r > 1$ by Assumption 2.5(*i*). (*ii*) We first examine a representative element of the first-row block of $\nabla_{(\theta,\lambda)}Q_t(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\delta})$: $(Y_t - \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})\widetilde{X}_{t,j}$, where $\widetilde{X}_{t,j}$ is the j-th element of $\widetilde{\mathbf{X}}_t$. We note that

$$\begin{aligned}\sup_{\theta,\lambda,\delta}|(Y_t - \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})\widetilde{X}_{t,j}| &\le |Y_t\widetilde{X}_{t,j}| + \sup_{\theta}\sum_{i=1}^{k+1}|\theta_i\widetilde{X}_{t,i}\widetilde{X}_{t,j}| + \sup_{\lambda}\sum_{i=1}^{m}|\lambda_i\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_i)\widetilde{X}_{t,j}|\\ &\le c_1M_t^2 + \sup_{\theta}\sum_{i=1}^{k+1}|\theta_i|M_t^2 + \sup_{\lambda}\sum_{i=1}^{m}|\lambda_i|\sup_{\tau\in\mathbf{T}}|\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\tau})\widetilde{X}_{t,j}|\\ &\le (c_1 + \sup_{\theta}\sum_{i=1}^{k+1}|\theta_i| + \sup_{\lambda}\sum_{i=1}^{m}|\lambda_i|)M_t^2,\end{aligned}$$

where the last inequality holds by Assumption 2.5. We further note that $\boldsymbol{\Theta}$ and $\boldsymbol{\Lambda}$ are compact parameter spaces, so that for some $c_3 > 0$, $|(Y_t - \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})\widetilde{X}_{t,j}| \le c_3M_t^2$ uniformly on $\boldsymbol{\Theta}\times\boldsymbol{\Lambda}\times\boldsymbol{\Delta}$, implying that $(Y_t - \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})\widetilde{X}_{t,j}$ is $2r$-dominated on $\boldsymbol{\Theta}\times\boldsymbol{\Lambda}\times\boldsymbol{\Delta}$ for some $r > 1$ by Assumption 2.5(*i*). Next, we examine a representative element of the second-row block of $\nabla_{(\theta,\lambda)}Q_t(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\delta})$: $(Y_t - \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_i)$, where

$\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_h)$ is the h-th element of $\mathbf{H}_t(\boldsymbol{\delta})$. We note that

$$\begin{aligned}
&\sup_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\delta}} |(Y_t - \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_h)| \\
&\leq \sup_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\delta}} \left\{ |Y_t\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_h)| + \sum_{i=1}^{k+1} |\theta_i \widetilde{X}_{t,i}\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_h)| + \sum_{i=1}^{m} |\lambda_i \Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_i)\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_h)| \right\} \\
&\leq \sup_{\boldsymbol{\tau}\in\mathbf{T}} |Y_t\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\tau})| + \sup_{\boldsymbol{\theta}} \sum_{i=1}^{k+1} |\theta_i \widetilde{X}_{t,i}| \sup_{\boldsymbol{\tau}\in\mathbf{T}} |\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\tau})| + \sup_{\boldsymbol{\lambda}} \sum_{i=1}^{m} |\lambda_i| \{\sup_{\boldsymbol{\tau}\in\mathbf{T}} |\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_i)|\}^2,
\end{aligned}$$

which is also bounded by $(c_1 + \sum_{i=1}^{k+1}|\theta_i| + \sum_{i=1}^{m}|\lambda_i|)M_t^2$. This implies that for some $r > 1$, $(Y_t - \widetilde{\mathbf{X}}_t'\boldsymbol{\theta} - \mathbf{H}_t(\boldsymbol{\delta})'\boldsymbol{\lambda})\widetilde{X}_{t,j}$ is $2r$-dominated on $\boldsymbol{\Theta} \times \boldsymbol{\Lambda} \times \boldsymbol{\Delta}$. Therefore, for some $r > 1$, each element of $\nabla_{(\boldsymbol{\theta},\boldsymbol{\lambda})}Q_t(\cdot)$ is $2r$-dominated on $\boldsymbol{\Theta} \times \boldsymbol{\Lambda} \times \boldsymbol{\Delta}$. (*iii*) We first examine a representative element of the first-row and first-column block of $\nabla^2_{(\boldsymbol{\theta},\boldsymbol{\lambda})}Q_t(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\delta})$: $\widetilde{X}_{t,j}\widetilde{X}_{t,i}$. We note that $|\widetilde{X}_{t,j}\widetilde{X}_{t,i}| \leq M_t^2$, so that Assumption 2.5(*iii*) implies that for some $r > 1$, $\widetilde{X}_{t,j}\widetilde{X}_{t,i}$ is $2r$-dominated on $\boldsymbol{\Theta} \times \boldsymbol{\Lambda} \times \boldsymbol{\Delta}$. Next, we consider a presentative element of the first-row and second-column block of $\nabla^2_{(\boldsymbol{\theta},\boldsymbol{\lambda})}Q_t(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\delta})$: $\widetilde{X}_{t,j}\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_h)$. We note that $\sup_{\boldsymbol{\tau}\in\mathbf{T}}|\widetilde{X}_{t,j}\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\tau})| \leq M_t^2$, so that Assumption 2.5(*iv*) implies that for some $r > 1$, $\widetilde{X}_{t,j}\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\delta}_h)$ as a function of $(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\delta})$ is $2r$-dominated on $\boldsymbol{\Theta} \times \boldsymbol{\Lambda} \times \boldsymbol{\Delta}$, where $h \in \{1, 2, \ldots, m\}$. Similarly, for some $r > 1$, the representative element of the the second-row and first-column block of $\nabla^2_{(\boldsymbol{\theta},\boldsymbol{\lambda})}Q_t(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\delta})$ is $2r$-dominated on $\boldsymbol{\Theta} \times \boldsymbol{\Lambda} \times \boldsymbol{\Delta}$ from the symmetry of a Hessian matrix. Finally, for some $r > 1$, the second-row and second-column block of $\nabla^2_{(\boldsymbol{\theta},\boldsymbol{\lambda})}Q_t(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\delta})$ is $2r$-dominated on $\boldsymbol{\Theta} \times \boldsymbol{\Lambda} \times \boldsymbol{\Delta}$ because $\{\sup_{\boldsymbol{\tau}\in\mathbf{T}}|\Psi(\widetilde{\mathbf{X}}_t'\boldsymbol{\tau})|\}^2 \leq M_t^2$, where $h, \ell \in \{1, 2, \ldots, m\}$. This shows that for some $r > 1$, $\nabla^2_{(\boldsymbol{\theta},\boldsymbol{\lambda})}Q_t(\cdot)$ is a matrix of elements that are $2r$-dominated on $\boldsymbol{\Theta} \times \boldsymbol{\Lambda} \times \boldsymbol{\Delta}$. This shows that for some $r > 1$, each element of $Q_t(\cdot)$, $\nabla_{(\boldsymbol{\theta},\boldsymbol{\lambda})}Q_t(\cdot)$, and $\nabla^2_{(\boldsymbol{\theta},\boldsymbol{\lambda})}Q_t(\cdot)$ is $2r$-dominated on $\boldsymbol{\Theta} \times \boldsymbol{\Lambda} \times \boldsymbol{\Delta}$ and verifies the DM$''$ of Gallant and White (Gallant 1988).

Sixth, Assumption 2.7(*i* and *ii*) also imposes that $\{Q_t(\cdot)\}$ and $\{\nabla_{(\boldsymbol{\theta},\boldsymbol{\lambda})}Q_t(\cdot)\}$ are near epoch processes of size $-(2\rho - 1)/(\rho - 1)$. As we have let $\rho = r/2$ above, this implies that they are near epoch processes of size $-2(r - 1)/(r - 1)$. This verifies the NE$'''$ of Gallant and White (Gallant 1988).

Seventh, for each $\boldsymbol{\delta}$, $\sum_{t=1}^{n} Q_t(\cdot, \boldsymbol{\delta})$ is a sum of quadratic functions, so that $E[Q_t(\cdot, \boldsymbol{\delta})]$ has a unique minimizer from the ergodic theorem and Assumptions 2.1 and 2.4. This satisfies the ID$'$ of Gallant and White (Gallant 1988).

Finally, the TL and WT of Gallant and White (Gallant 1988) are directly imposed by Assumption 2.7(*iii*).

Thus, applying theorem 6.8 of Gallant and White (Gallant 1988) implies that $\widehat{\mathbf{B}}_n(\cdot)$ is a consistent estimator for $\mathbf{B}(\cdot)$ uniformly on $\boldsymbol{\Delta}$. Given this, as $\boldsymbol{\delta}$ is drawn from $\mathbb{Q}$ independent of $\mathbb{P}$, for any $\varepsilon > 0$, there are $F \in \mathcal{D}$ with $\mathbb{Q}(F) = 1$ and n_0 such that if $n > n_0$, $\mathbb{P} \cdot \mathbb{Q}(\|\widehat{\mathbf{B}}_n(\boldsymbol{\delta}) - \mathbf{B}(\boldsymbol{\delta})\|_\infty > \varepsilon | F) < \varepsilon$. This completes the proof. ■

Proof of Theorem 2.1: Given that $\widehat{\mathbf{A}}_n(\boldsymbol{\delta})$ and $\widehat{\mathbf{B}}_n(\boldsymbol{\delta})$ are consistent for $\mathbf{A}(\boldsymbol{\delta})$ and $\mathbf{B}(\boldsymbol{\delta})$, respectively, by Lemma 2.4, the results for $\widehat{T}_n|\boldsymbol{\delta}$ follow from Eq. (4).

Furthermore, the result for $\dot{T}_n|\delta$ also trivially follows if $\widehat{\sigma}_n^2 \overset{\mathbb{P}}{\to} \sigma_*^2$ under $\mathbb{H}_0$. We note that

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{t=1}^{n} U_t^2 - \left(\frac{1}{n}\sum_{t=1}^{n} U_t \widetilde{\mathbf{X}}_t'\right)\left(\frac{1}{n}\sum_{t=1}^{n} \widetilde{\mathbf{X}}_t \widetilde{\mathbf{X}}_t'\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n} U_t \widetilde{\mathbf{X}}_t\right).$$

We now apply the ergodic theorem and obtain

$$\widehat{\sigma}_n^2 \overset{\mathbb{P}}{\to} E[U_t^2] - \mathbf{0}' E[\widetilde{\mathbf{X}}_t \widetilde{\mathbf{X}}_t']\mathbf{0} = E[U_t^2] = \sigma_*^2,$$

where necessary moment conditions for the ergodic theorem are satisfied by Assumption 2.4. ■

References

ANDREWS, D.W.K. (1991): "Heteroskedasticity and Autocorreltaion Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

BAEK, Y. AND CHO, J. (2013): "Testing the Linearity Hypothesis Using Power Transformations," Discussion Paper, School of Economics, Yonsei University.

BARTLETT1950 (1950): "Peridogram Analysis and Continuous Spectra," *Biometrika*, 37, 1–16.

BIERENS, H. (1990): "A Consistent Conditional Moment Test of Functional Form," *Econometrica*, 58, 1443–1458.

BRADLEY, R. (1985): "On the Central Limit Theorem under Absolute Regularity," *Annals of Probability*, 13, 1314–1325.

CHACKO, B., V., VIMAL KRISHNAN, V., RAJU, G., AND BABU ANTO, P. (2012): "Handwritten Character Recognition Using Wavelet Energy and Extreme Learning Machine," *International Journal of Machine Learning and Cybernetics*, 3, 149-161.

CHO, J. (2012): "Quasi-maximum likelihood estimation revisited using the distance and direction method," *Journal of Economic Theory and Econometrics*, 23:2, 89-112.

CHO, J., CHEONG, T., AND WHITE, H. (2011): "Experience with the Weighted Bootstrap in Testing for Unobserved Heterogeneity in Exponential and Weibull Duration Models," *Journal of Economic Theory and Econometrics*, 22:2, 60–91.

CHO, J., AND HAN, C. (2009): "Testing for the Mixture Hypothesis of Geometric Distributions," *Journal of Economic Theory and Econometrics*, 20:3, 31–55.

CHO, J., HUANG, M., AND WHITE, H. (2010): "Testing the Constant Mean Function Using Functional Regression," Discussion Paper, School of Economics, Yonsei University.

CHO, J., AND ISHIDA, I. (2012): "Testing for the Effects of Omitted Power Transformations," *Economics Letters*, 117, 287–290.

CHO, J., ISHIDA, I., AND WHITE, H. (2011): "Revisiting Tests for Neglected Nonlinearity Using Artificial Neural Networks," *Neural Computation*, 23, 1133–1186.

CHO, J., ISHIDA, I., AND WHITE, H. (2013): "Testing for Neglected Nonlinearity Using Twofold Unidentified Models under the Null and Hexic Expansions," *Essays in Nonlinear Time Series Econometrics, A Festschrift in Honor of Timo Terasvirta*. Eds. Niels Haldrup, Mika Meitz, and Pentti Saikkonen. Oxford: Oxford University Press, forthcoming.

CHO, J. AND WHITE, H. (2007): "Testing for Regime Switching," *Econometrica*, 75, 1671–1720.

CHO, J. AND WHITE, H. (2010): "Testing for Unobserved Heterogeneity in Exponential and Weibull Duration Models," *Journal of Econometrics*, 157, 458–480.

CHO, J. AND WHITE, H. (2011): "Testing Correct Model Specification Using Extreme Leaning Machines," *Neurocomputing*, 74, 2552–2565.

DAVIES, R. (1977): "Hypothesis Testing When a Nuisance Parameter is Present Only under the Alternative," *Biometrika*, 64, 247–254.

DAVIES, R. (1987): "Hypothesis Testing When a Nuisance Parameter is Present Only under the Alternative," *Biometrika*, 74, 33–43.

DEMSĂR, J. "Statistical Comparisons of Classifiers over Multiple Data Sets," *Journal of Machine Learning Research*, 7, 1–30.

DOUKHAN, P., MASSART, P., AND RIO, E. (1995): "Invariance Principles for Absolutely Regular Empirical Processes," *Annales de l'Institut Henri Poincaré, Probabilites et Statistiques*, 31 393–427.

GALLANT, R., AND WHITE, H. (1988): *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Oxford: Basil Blackwell.

HANSEN, B. (1996): "Inference When a Nuisance Parameter Is Not Identified under the Null Hypothesism," *Econometrica*, 64, 413–430.

HORNIK, K., STINCHCOMBE, M., AND WHITE, H. (1989): "Multilayer Feedforward Networks are Universal Approximators," *Neural Networks*, 2, 359–366.

HORNIK, K., STINCHCOMBE, M., AND WHITE, H. (1990): "Universal Approximation of an Unknown Mapping and Its Derivatives Using Multi-layer Feedforward Networks," *Neural Networks*, 3, 551–560.

HUANG, G., ZHU, Q., AND SIEW, C. (2006): "Extreme Learning Machine: Theory and Applications," *Neurocomputing*, 70, 489–501.

G. B. HUANG, D. H. WANG, Y. LAN. (2011): "Extreme Learning Machines: A Survey," *International Journal of Machine Learning and Cybernetics*, 2, 107–122.

MOHAMMED, A., MINHAS, R., JONATHAN WU Q., AND SID-AHMED, M. (2011): "Human Facerecognition Based on Multidimensional PCA and Extreme Learning Machine," *Pattern Recognition*, 44, 2588–2597.

NEWEY, W. AND WEST, K. (1987): "A Simple Positive Definite Heteroskedasticity and Autocorrelation Consistent covariance Matrix," *Econometrica*, 55, 703–708.

RAMSEY, J. (1969): "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis," *Journal of the Royal Statistical Society, Series B.*, 31, 350–371.

ROSENBLATT, M. (1950): "Remarks on a Multivariate Transformation," Annals of Mathematical Statistics, 23, 470–472.

STINCHCOMBE, M. AND WHITE, H. (1998): "Consistent Specification Testing with Nuisance Parameters Present Only under the Alternative," *Econometric Theory*, 14, 295–324.

YUAN, Y., WANG, U., CAO, F. (2011): "Optimization Approximation Solution for Regression Problem Based on Extreme Learning Machine," *Neurocomputing*, 74, 2475–2482.

WANG, X., CHEN, A., FENG, H. (2011): "Upper Integral Network with Extreme Learning Mechanism," *Neurocomputing* 74, 2520–2525.

WANG, X., DONG, L., AND YAN, J. (2012): "Maximum Ambiguity-Based Sample Selection in Fuzzy Decision Tree Induction," *IEEE Transactions on Knowledge and Data Engineering*, 24, 491–1505.

WHITE, H. (1989): "An Additional Hidden Unit Test for Neglected Nonlienarity," *Proceedings of the International Joint Conference on Neural Networks*, 2, 451–455. New York: IEEE Press.

WHITE, H. (2001): *Asymptotic Theory for Econometricians*. Orlando: Academic Press.

WHITE, H. AND CHO, J. (2012): "Higher-Order Approximations for Testing Neglected Nonlinearity," *Neural Computation*, 24, 273–287.

WU, J., WANG, S., AND CHUNG, F. (2011): “Positive and Negative Fuzzy Rule System, Extreme Learning Machine and Image Classification,” *International Journal of Machine Learning and Cybernetics*, 2, 261-271.

ZHAI, J., XU, H., AND WANG, X. (2012): “Dynamic Ensemble Extreme Learning Machine Based on Sample Entropy,” *Soft Computing*, 16, 1493-1502.

Table 1: Empirical Rejection Rates of $\dot{T}_n$ under $\mathbb{H}_0$ (in Percent). Number of Replications: 10,000. DGP: $Y_t = \varepsilon_t$ and $\varepsilon_t \sim$ IID $N(0,1)$. Null Model: $Y_t = \alpha_* + \beta_* Y_{t-1} + U_t$.

m	Dist. of δ_j	Nominal Level \ n	50	100	200	500	1,000
1	$U[-0.5, 0.5]$	1.00 %	0.75	0.97	0.96	1.05	1.08
		5.00 %	4.73	5.02	5.07	4.80	5.30
		10.0 %	9.93	9.94	10.50	9.95	10.51
1	$U[-1.0, 1.0]$	1.00 %	0.78	0.92	1.03	1.13	0.94
		5.00 %	4.43	4.74	5.19	4.97	4.77
		10.0 %	9.68	9.63	10.16	10.38	9.82
1	$U[-10, 10]$	1.00 %	0.85	0.92	1.11	0.81	1.08
		5.00 %	5.19	4.61	4.86	4.50	5.20
		10.0 %	10.26	9.78	10.19	9.80	10.20
2	$U[-0.5, 0.5]$	1.00 %	0.53	0.84	0.83	0.80	0.94
		5.00 %	4.15	4.69	4.59	4.86	4.65
		10.0 %	9.19	9.61	9.02	10.00	9.96
2	$U[-1.0, 1.0]$	1.00 %	0.69	0.90	0.86	0.89	0.89
		5.00 %	4.51	4.91	4.48	4.55	4.95
		10.0 %	9.57	10.02	9.51	9.42	9.73
2	$U[-10, 10]$	1.00 %	0.75	0.74	0.94	1.00	0.99
		5.00 %	4.34	4.66	4.63	4.79	4.97
		10.0 %	9.27	10.00	9.76	10.12	9.75
3	$U[-0.5, 0.5]$	1.00 %	0.76	0.88	0.80	0.93	0.98
		5.00 %	4.34	4.57	4.78	4.64	4.61
		10.0 %	9.27	9.25	9.59	9.43	9.64
3	$U[-1.0, 1.0]$	1.00 %	0.53	0.84	0.85	0.93	0.94
		5.00 %	4.78	4.44	4.61	4.84	4.71
		10.0 %	9.70	9.47	9.40	9.59	9.71
3	$U[-10, 10]$	1.00 %	0.59	0.86	1.01	0.98	1.18
		5.00 %	4.42	4.77	5.05	5.08	3.97
		10.0 %	9.87	10.09	10.02	10.12	8.63

Table 2: Empirical Rejection Rates of $\dot{T}_n$ under $\mathbb{H}_1$ (in Percent): Level of Significance = 5%. Number of Replications: 2,000. DGP: $Y_t = \cos(Y_{t-1}) + \varepsilon_t$ and $\varepsilon_t \sim$ IID $N(0,1)$. Null Model: $Y_t = \alpha_* + \beta_* Y_{t-1} + U_t$.

m	Dist. of δ_j \ n	50	100	200	400	600	800	1,000
1	$U[-0.5, 0.5]$	61.55	79.55	87.50	91.20	93.10	93.96	94.90
	$U[-1.0, 1.0]$	61.70	81.10	88.45	93.00	93.45	94.55	95.55
	$U[-10, 10]$	55.65	75.95	86.25	91.75	92.85	94.90	95.25
2	$U[-0.5, 0.5]$	75.65	96.15	99.55	99.90	99.90	99.80	99.90
	$U[-1.0, 1.0]$	72.90	95.65	99.75	99.90	99.90	99.80	99.95
	$U[-10, 10]$	65.90	90.85	98.35	99.80	99.95	99.95	99.80
3	$U[-0.5, 0.5]$	69.85	95.20	99.85	99.95	99.95	99.90	99.95
	$U[-1.0, 1.0]$	70.20	96.20	99.95	99.95	100.0	100.0	100.0
	$U[-10, 10]$	66.4	92.35	99.75	100.0	100.0	100.0	100.0

Table 3: Empirical Rejection Rates of $\widehat{T}_n$ under $\mathbb{H}_0$ (in Percent). Number of Replications: 10,000. DGP: $Y_t = 0.75Y_{t-1} - 0.25Y_{t-2} + \varepsilon_t$ and $\varepsilon_t \sim$ IID $N(0,1)$. Null Model: $Y_t = \alpha_* + \beta_* Y_{t-1} + U_t$.

m	Dist. of δ_j	Nominal Level $\setminus$ n	50	100	200	500	1,000
		1.00 %	4.25	3.32	2.84	1.72	1.49
1	$U[-0.5, 0.5]$	5.00 %	10.77	9.71	8.49	6.47	6.14
		10.0 %	17.34	15.74	14.44	12.11	11.21
		1.00 %	3.58	3.06	2.21	1.54	1.26
1	$U[-1.0, 1.0]$	5.00 %	10.17	8.81	7.64	6.16	5.74
		10.0 %	16.45	14.71	13.62	11.23	10.66
		1.00 %	4.17	2.93	2.44	1.65	1.40
1	$U[-10, 10]$	5.00 %	10.70	8.66	7.44	6.24	6.13
		10.0 %	16.81	14.53	12.41	10.91	10.86
		1.00 %	8.25	6.08	4.61	2.85	2.22
2	$U[-0.5, 0.5]$	5.00 %	17.48	14.82	12.28	8.77	7.55
		10.0 %	25.28	22.15	19.07	15.01	12.90
		1.00 %	7.96	5.51	4.11	2.39	2.02
2	$U[-1.0, 1.0]$	5.00 %	16.60	14.20	10.76	8.20	6.91
		10.0 %	24.43	21.41	17.28	14.31	12.63
		1.00 %	8.11	5.84	3.82	2.42	1.77
2	$U[-10, 10]$	5.00 %	17.17	13.80	11.04	8.10	6.68
		10.0 %	24.65	20.93	17.14	13.37	11.74
		1.00 %	18.97	14.53	11.83	7.77	5.90
3	$U[-0.5, 0.5]$	5.00 %	30.01	25.57	22.09	15.82	12.65
		10.0 %	37.58	32.89	29.76	23.08	19.83
		1.00 %	18.05	13.00	9.62	5.90	3.91
3	$U[-1.0, 1.0]$	5.00 %	29.91	23.59	19.60	14.07	10.54
		10.0 %	37.65	31.39	27.12	20.97	16.88
		1.00 %	15.71	9.28	5.86	3.58	2.60
3	$U[-10, 10]$	5.00 %	27.02	18.49	13.31	8.86	8.23
		10.0 %	34.81	25.85	19.92	14.78	13.64

Table 4: Empirical Rejection Rates of $\widehat{T}_n$ under $\mathbb{H}_1$ (in Percent): Level of Significance = 5%. Number of Replications: 2,000. DGP: $Y_t = \cos(Y_{t-1}) + \varepsilon_t$ and $\varepsilon_t \sim$ IID $N(0,1)$. Null Model: $Y_t = \alpha_* + \beta_* Y_{t-1} + U_t$.

m	Dist. of δ_j $\setminus$ n	50	100	200	400	600	800	1,000
	$U[-0.5, 0.5]$	68.95	79.60	86.65	90.10	91.80	92.55	93.75
1	$U[-1.0, 1.0]$	71.65	80.55	87.60	90.35	93.15	93.30	93.80
	$U[-10, 10]$	63.85	78.55	86.65	90.35	91.65	94.05	93.15
	$U[-0.5, 0.5]$	88.90	97.60	99.50	99.70	99.85	99.85	99.85
2	$U[-1.0, 1.0]$	87.55	97.75	99.85	99.60	99.95	99.95	99.90
	$U[-10, 10]$	80.60	93.65	98.45	99.60	99.50	99.70	99.65
	$U[-0.5, 0.5]$	87.25	94.95	98.35	97.95	98.30	98.15	98.65
3	$U[-1.0, 1.0]$	91.20	98.00	99.70	99.60	99.80	99.40	99.75
	$U[-10, 10]$	86.50	96.45	99.80	99.95	99.95	100.0	100.0

Figure 1: EMPIRICAL P-P PLOTS OF $\dot{T}_n$. Number of Replications: 10,000. DGP: $Y_t = \varepsilon_t$ and $\varepsilon_t \sim$ IID $N(0,1)$. Null Model: $Y_t = \alpha_* + \beta_* Y_{t-1} + U_t$. $m = 1$.

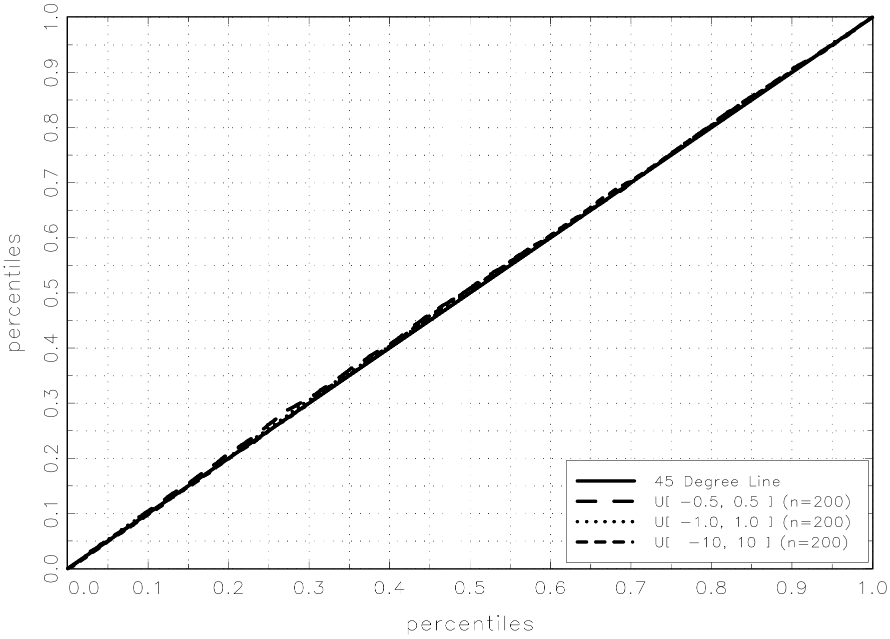


Figure 2: EMPIRICAL P-P PLOTS OF $\dot{T}_n$. Number of Replications: 10,000. DGP: $Y_t = \varepsilon_t$ and $\varepsilon_t \sim$ IID $N(0,1)$. Null Model: $Y_t = \alpha_* + \beta_* Y_{t-1} + U_t$. $m = 2$.

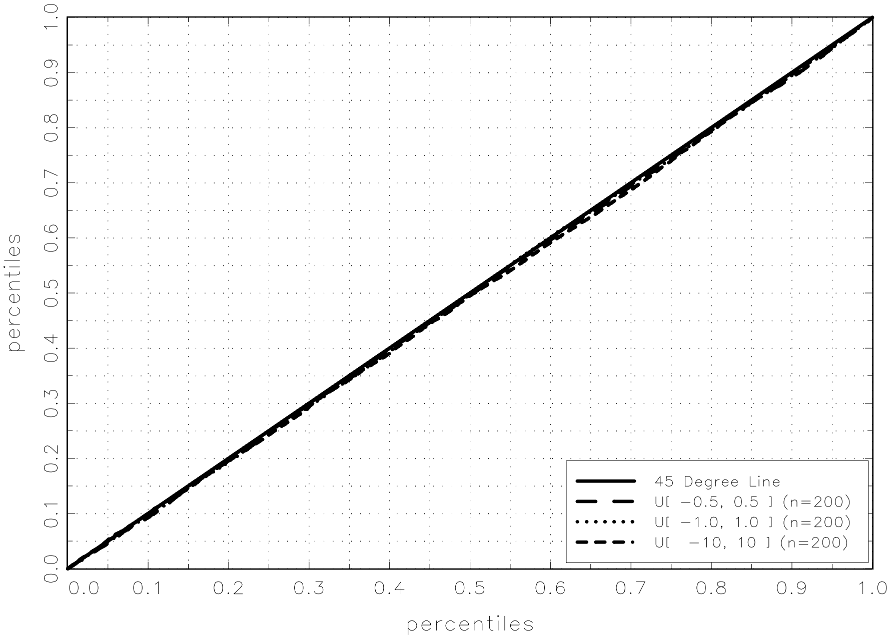


Figure 3: Empirical P-P Plots of $\dot{T}_n$. Number of Replications: 10,000. DGP: $Y_t = \varepsilon_t$ and $\varepsilon_t \sim$ IID $N(0,1)$. Null Model: $Y_t = \alpha_* + \beta_* Y_{t-1} + U_t$. $m = 3$.

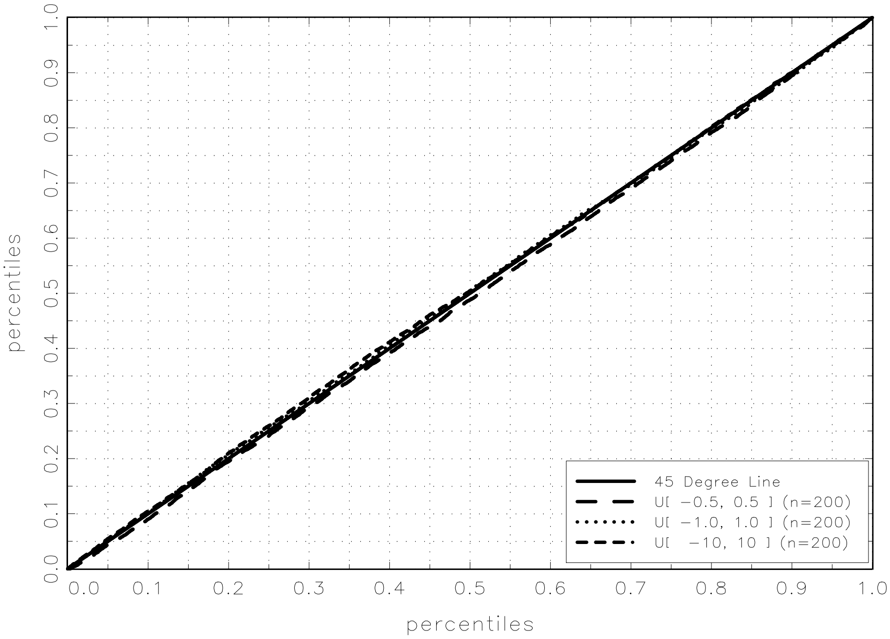


Figure 4: Empirical P-P Plots of $\widehat{T}_n$. Number of Replications: 10,000. DGP: $Y_t = 0.75Y_{t-1} - 0.25Y_{t-2} + \varepsilon_t$ and $\varepsilon_t \sim$ IID $N(0,1)$. Null Model: $Y_t = \alpha_* + \beta_* Y_{t-1} + U_t$. $\delta_j \sim U[-1,1]$. $m = 1$.

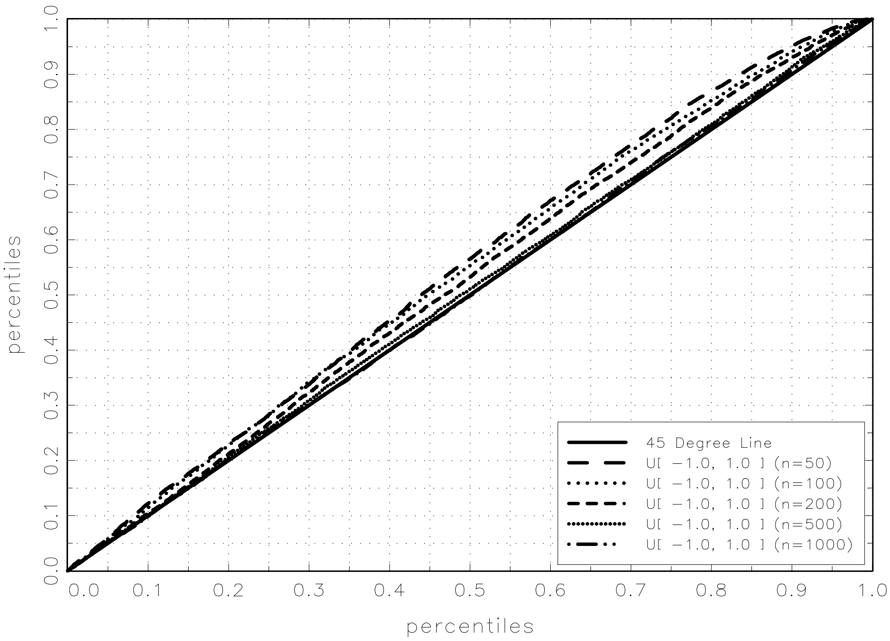


Figure 5: EMPIRICAL P-P PLOTS OF $\widehat{T}_n$. Number of Replications: 10,000. DGP: $Y_t = 0.75Y_{t-1} - 0.25Y_{t-2} + \varepsilon_t$ and $\varepsilon_t \sim$ IID $N(0,1)$. Null Model: $Y_t = \alpha_* + \beta_* Y_{t-1} + U_t$. $\delta_j \sim U[-1,1]$. $m = 2$.

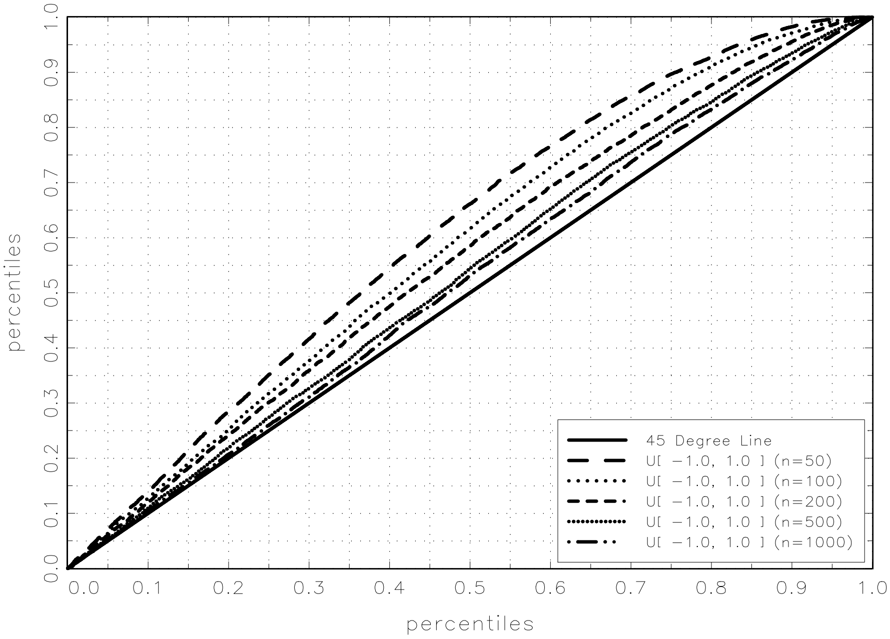
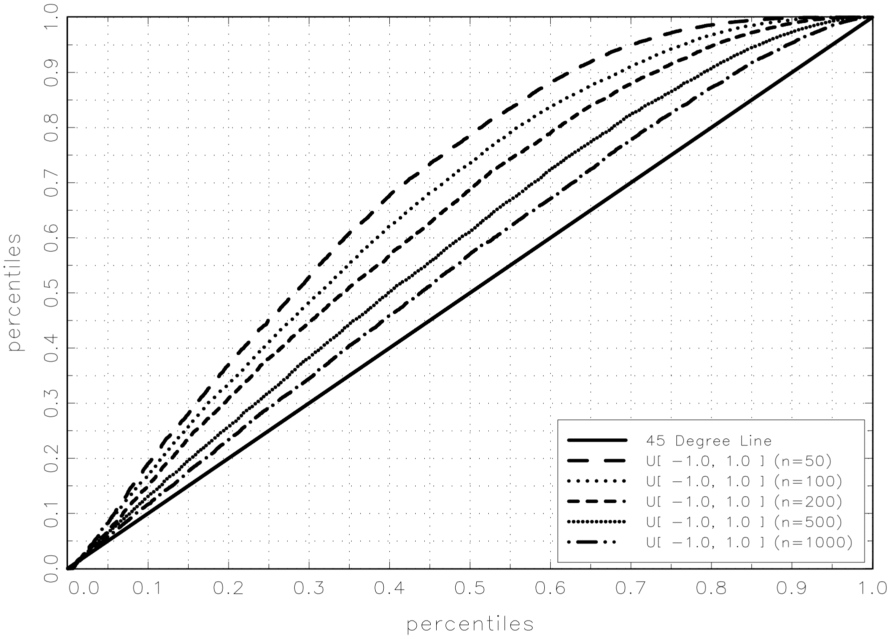


Figure 6: EMPIRICAL P-P PLOTS OF $\widehat{T}_n$. Number of Replications: 10,000. DGP: $Y_t = 0.75Y_{t-1} - 0.25Y_{t-2} + \varepsilon_t$ and $\varepsilon_t \sim$ IID $N(0,1)$. Null Model: $Y_t = \alpha_* + \beta_* Y_{t-1} + U_t$. $\delta_j \sim U[-1,1]$. $m = 3$.


percentiles
45 Degree Line
U[−1.0, 1.0] (n=50)
U[−1.0, 1.0] (n=100)
U[−1.0, 1.0] (n=200)
U[−1.0, 1.0] (n=500)
U[−1.0, 1.0] (n=1000)
percentiles