Sequentially Estimating Approximate Conditional Mean Using the Extreme Learning Machine

LIJUAN HUO
School of Humanities and Social Sciences, Beijing Institute of Technology, Haidian, Beijing, 100081, China
Email: ljhuo@bit.edu.cn

JIN SEO CHO
School of Humanities and Social Sciences, Beijing Institute of Technology, Haidian, Beijing, 100081, China
School of Economics, Yonsei University, Seodaemun, Seoul, 03722, Korea
Email: jinseocho@yonsei.ac.kr

This version: November, 2020

Abstract

This paper studies the extreme learning machine (ELM) applied to the Wald test statistic for model specification of the conditional mean, which we call the WELM testing procedure. The omnibus test statistics available in the literature weakly converge to a Gaussian stochastic process under the null that the model is correct, and it makes their application inconvenient. In contrast, the WELM testing procedure is straightforwardly applicable when detecting model misspecification. We apply the WELM testing procedure to the sequential testing procedure formed by a set of polynomial models and estimate an approximate conditional expectation by this. We conduct extensive Monte Carlo experiments and evaluate the performance of the sequential WELM testing procedure and verify that it consistently estimates the most parsimonious conditional mean, when the set of the polynomial models contains a correctly specified model. Otherwise, it consistently rejects all the models in the set.

Key Words: conditional mean specification testing; omnibus test; Gaussian process; extreme learning machine; Wald test statistic; functional regression; sequential testing procedure; consistent correct model estimation.

Acknowledgments: The authors are mostly grateful to the two anonymous referees for their helpful comments. Part of thework for this paperwas donewhen the corresponding authorwas visiting the Department of Economics, The Chinese University of Hong Kong, whose kind hospitality is gratefully acknowledged. Cho appreciates research support from the Yonsei University Research Grant of 2020, and Huo's research is supported by the National Natural Science Foundation of China (Grant-71803009).

Supplementary Materials: The program codes to reproduce the simulation outputs are available online at `https://web.yonsei.ac.kr/jinseocho/swelm.htm`.

The authors declare no conflict of interest.

1 Introduction

Conducting data inference using correctly specified models is most desirable for predicting future observations. If models are misspecified, proper data inference cannot be made, and predicting future observations may be involved with an undesired bias.

Due to this, previous literature develops methodologies to test correct model assumption. For example, in a classic study, Ramsey (1969) provides a test statistic for nonlinearity. In another classic study, Bierens (1990) provides an omnibus model specification test statistic that detects arbitrary model misspecification consistently. In addition to these previous works, a number of studies provide correct model specification testing methodologies (e.g., Keenan, 1985; Tsay, 1986; White, 1987; Lee, White, and Granger, 1993; Cho, Ishida, and White, 2011, among others).

Notwithstanding the rapid development of the correct model specification testing, the researcher may not be able to obtain a correctly specified model and may have to predict future observations using misspecified models. Note that if all candidate models are misspecified by the model specification tests, the model with the smallest mean square error is typically chosen to forecast future observations, even if it is known to be misspecified.

The main goal of the current study to provide a robust methodology to search for a correct model in a systematic way. To do so, we develop a sequential testing procedure that combines the model specification test statistic available in previous literature with high-degree polynomial models so that a close approximation for conditional mean equation can be consistently estimated.

In previous literature, the model specification testing using the artificial neural networks (ANNs) are widely applied due to their universal approximation property (e.g., Hornik, Stinchcombe, and White, 1989, 1990; Stinchcombe and White, 1998). Cho, Ishida, and White (2011) propose an ANN-based quasi-likelihood ratio (QLR) statistic for testing neglected nonlinearity that exploits the generically comprehensively revealing feature of the ANN-based test statistic and overcomes the so-called twofold Davies' (1977, 1987) identification problem to obtain its null limit distribution as a functional of a Gaussian stochastic process.

Notwithstanding the theoretical self-efficacy of the QLR test statistic, it may not be convenient for empirical applications. Its null limit distribution is dependent on the model scopes, so that the asymptotic critical values are different from model to model. Cho, Phillips, and Seo (2020) and Cho, Huang, and White (2020) note this inconvenience and define a Wald test statistic by functional regression so that it follows a chi-squared distribution with one degree of freedom under the null hypothesis that the model is correctly specified. Cho and White (2011) further demonstrate that if the ELM proposed by Huang, Zhu, and Siew (2006) is combined with the Wald test statistic, its computation can be efficiently made in addition to being generically comprehensively revealing. They refer to this as the Wald-ELM (WELM) testing procedure.

The polynomial model is also widely applied for empirical applications. Its popularity lies in the fact that the polynomial model has a recursive structure and is able to uniformly approximate any continuous function. This aspect makes it convenient to apply it to a sequential testing procedure. If a lower degree polynomial model is rejected by a proper testing procedure, we can consider its next higher-degree polynomial model as another approximation and test model adequacy.

Previous studies also apply the sequential testing procedures to polynomial models. Cho and Phillips (2018) develop a sequential testing methodology of testing the null of a polynomial function to identify the polynomial degree by extending the testing methodology in Baek, Cho, and Phillips (2015). Specifically, Baek, Cho, and Phillips (2015) note that if the QLR test statistic in Cho, Ishida, and White (2011) is applied to a linear model augmented by a power transformation, the twofold identification problem is transformed to a trifold Davies's (1977, 1987) identification problem. They overcome this and derive the null limit distribution of the QLR test statistic, and recommend using Hansen's (1996) weighted bootstrap for empirical applications. This recommendation is mainly because the null limit distribution is associated with a Gaussian stochastic process as for the twofold identification problem. Specifically, the null limit distribution is represented by the maximum of the squared Gaussian process, so that the asymptotic critical values are different from model to model, making its application inconvenient in obtaining the asymptotic critical values. Cho and Phillips (2018) further extend the QLR test statistic to testing the null of the polynomial function hypothesis and obtain its null limit distribution by overcoming the multifold identification problem that is further developed from the trifold identification problem in Baek, Cho, and Phillips (2015). In addition to this derivation, they apply the null limit distribution to the sequential testing procedure to search for a close an approximate to conditional mean function. For practical applications of the sequential testing procedure, they also recommend applying the weighted bootstrap as in Baek, Cho, and Phillips (2015).

We employ the WELM test statistic for the goal of this study and apply it to the sequential testing procedure. This statistic is convenient for applications and also possesses the generically comprehensively revealing feature, so that it can be employed for the goal of this study. In addition, the WELM test statistic has features not shared by the test statistic in previous literature. First, the null limit distribution is obtained as a chi-squared distribution, so that the traditional theory on the sequential testing procedure can be applied (e.g., Hosoya, 1989). We do not need to apply the approximation theory on the probability of the maximum of a squared Gaussian process as for the QLR test statistic. Furthermore, as we discuss below, the sequential testing procedure is conducted by reducing the level of significance in response to the rise of the sample size so that the degree estimation error reduces to zero asymptotically. As the null limit distribution is chi-squared, we can easily choose the plans for the level of significance without satisfying the additional condition for the application of the QLR test statistic that the level of significance converges to zero very slowly. This condition does not have to be imposed on our sequential testing procedure. Second, the conditioning variable does not have to be positively valued as required by Cho and Phillips (2011). Even when the conditioning variable is negatively valued, the sequential testing procedure using the WELM test statistic is directly applicable.

The rest of this paper is organized as follows. Section 2 focuses on the polynomial model and provides the null limit distribution of the WELM test statistic along with our description of the literature development of this statistic. Section 3 applies the WELM test statistic to the sequential testing procedure and provides the theoretical results desired by the application. Section 4 conducts extensive simulations using the WELM test statistic and the sequential testing procedure. We consider three different data generating processes (DGPs) and examine how the sequential testing procedure responds to various plans for the level of significance. Section 5 provides concluding remarks and summarizes the main findings. All the mathematical proofs are presented in the Appendix.

2 Method 1: Application of the WELM Testing to the Polynomial Model

In this section, we first describe the main motivation of this study in relation to the development of previous literature in terms of model specification testing. To fix our idea, we focus on the WELM test statistic applied to the polynomial model.

Our primary interest is in developing a statistical methodology to estimate the conditional mean equation of time-series observations. We therefore suppose that data are weakly dependent observations as follows:

Assumption 1 (DGP). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, and let $k \in \mathbb{N}$. Let $\{(y_t, x_t, \boldsymbol{d}_t')' : \Omega \mapsto \mathbb{R}^{2+k} : t = 1, 2, \ldots\}$ be a strictly stationary and absolutely regular process with mixing coefficients β_τ such that for some $\rho > 1$, $\sum_{t=1}^{\infty} \tau^{2\rho/(\rho-1)} \beta_\tau < \infty$ and x_t is strictly non-negative with probability 1.* □

Here, y_t and $\boldsymbol{z}_t := (x_t, \boldsymbol{d}_t')'$ are serially dependent target and explanatory variables, respectively, and $\boldsymbol{z}_t$ can contain the lagged target variables, so that dynamic misspecification can be removed from our consideration. Specifically, the researcher is concerned with a possible nonlinearity with respect to x_t when they attempt to approximate the conditional mean equation by the p-th degree polynomial function:

$$\mathbb{E}[y_t|\mathcal{F}_t] \approx \boldsymbol{x}_t(p)'\boldsymbol{\alpha}_*(p) + \boldsymbol{d}_t'\boldsymbol{\eta}_*,$$

where $\boldsymbol{x}_t(p) := (1, x_t, ..., x_t^p)'$, $\boldsymbol{\theta}_*(p) := (\boldsymbol{\alpha}_*(p)', \boldsymbol{\eta}_*)'$ is the linear coefficient of $(\boldsymbol{x}(p)', \boldsymbol{d}_t')'$, and $\mathcal{F}_t$ is the smallest σ-field generated by $(\boldsymbol{z}_t, y_{t-1}, \boldsymbol{z}_{t-1}, y_{t-2}, \ldots)$.

The polynomial functions are uniformly dense, and this motivates us to estimate the conditional mean by the above specification. Note that Stone-Weierstrass theorem implies that continuous functions are uniformly approximated by polynomial functions with high levels of degrees, so that the above polynomial function becomes a successful approximation for the conditional mean if the degree p is sufficiently large.

The current study seeks to provide a statistical method to estimate the degree of the polynomial function in the most parsimonious manner. Note that the non-local behavior of a high-degree polynomial model is understood as one of the drawbacks of estimating the high-degree polynomial model by regression. That is, the outlier of x_t can substantially affect the estimated forecast, and this can reduce the utility of the polynomial model estimation (e.g., Magee, 1998).

We accommodate this aspect by estimating the polynomial using the most parsimonious model. Specifically, we estimate the polynomial degree p as small as possible, and for this purpose, we provide a sequential testing methodology described in the next section. In particular, our testing approach is based upon the generically comprehensively revealing (GCR) property of an ANN model and ELMs.

For a precise description of our testing procedure using the extreme learning machine applied to the GCR property, we note that Stinchcombe and White (1998) show that when the regression model is estimated by attaching an analytic function to a linear model, the linear coefficient consistently estimates a non-zero coefficient if and only if the regression model is misspecified for the conditional mean equation; the authors refer to this as the GCR property. Specifically, the following assumption gives the model advocated by Stinchcombe and White (1998):

Assumption 2 (Model)**.** *Let* $\mathcal{M}_p := \{f(\cdot; \boldsymbol{\theta}(p), \lambda, \delta) : (\boldsymbol{\theta}(p), \lambda, \boldsymbol{\delta}) \in \boldsymbol{\Theta}(p) \times \Lambda \times \Delta\}$ *is specified as the alternative model, where*

$$f(\boldsymbol{z}_t; \boldsymbol{\theta}(p), \lambda, \delta) := \boldsymbol{x}_t(p)'\boldsymbol{\alpha}(p) + \boldsymbol{d}_t'\boldsymbol{\eta} + \lambda\Psi(\delta x_t)$$

and $\Psi(\cdot)$*,* δ*,* λ*, are the additional hidden unit constructed by an analytic function, the input-to-hidden weights, and the hidden-to-output weight, respectively.* □

Here, if we let λ_* be the probability limit of the estimated parameter by regression, the GCR property implies that the estimated coefficient for λ_* is consistently different from zero if the p-degree polynomial model is misspecified for the conditional mean. Therefore, we can detect whether the p-th degree polynomial model is correct by testing whether the coefficient of the hidden unit is zero. That is, if the estimated hidden-to-output weight is statistically different from zero, it means that $\mathcal{M}_p$ does not approximate the model sufficiently well. Otherwise, $\mathcal{M}_p$ becomes a successful approximation for the conditional mean, motivating us to rephrase the following hypotheses:

$$\mathcal{H}_0 : \text{For some } \boldsymbol{\theta}_*(p) \in \boldsymbol{\Theta}(p), \quad \mathbb{P}[\mathbb{E}(y_t|\mathcal{F}_t) = \boldsymbol{x}_t(p)'\boldsymbol{\alpha}_*(p) + \boldsymbol{d}_t'\boldsymbol{\eta}_*] = 1 \quad \text{versus}$$

$$\mathcal{H}_1 : \text{For every } \boldsymbol{\theta}(p) \in \boldsymbol{\Theta}(p), \quad \mathbb{P}[\mathbb{E}(y_t|\mathcal{F}_t) = \boldsymbol{x}_t(p)'\boldsymbol{\alpha}(p) + \boldsymbol{d}_t'\boldsymbol{\eta}] < 1$$

into the following equivalent hypotheses:

$$\mathcal{H}_0' : \lambda_* = 0 \quad \text{versus} \quad \mathcal{H}_1' : \lambda_* \neq 0$$

in their framework. This implies that we can let our null model be

$$\mathcal{M}_p^0 := \{f^0(\cdot; \boldsymbol{\theta}(p)) : \boldsymbol{\theta}(p) \in \boldsymbol{\Theta}(p)\}$$

and $f^0(\boldsymbol{z}_t; \boldsymbol{\theta}(p)) := \boldsymbol{x}_t(p)'\boldsymbol{\alpha}(p) + \boldsymbol{d}_t'\boldsymbol{\eta}$ for $p = 1, 2, \ldots$. In what follows, we let $\Psi_t(\delta)$ denote $\Psi(\delta x_t)$ for notational simplicity.

This aspect now implies that the GCR property can be exploited by testing $\mathcal{H}_0'$ against $\mathcal{H}_1'$, and we need to test whether the input-to-output weight is zero.

We now provide the regularity conditions for the regular behavior of the test statistics provided below:

Assumption 3 (Regularity)**.** *(i)* $(\Delta, \mathcal{D}, \mathbb{Q})$ *and* $(\Omega \times \Delta, \mathcal{F} \times \mathcal{D}, \mathbb{P} \cdot \mathbb{Q})$ *are complete probability spaces. (ii) For* $p \in \mathbb{N}$, $\boldsymbol{\Theta}(p)$ *is a non-empty compact and convex set, and* Λ *and* Δ *are non-empty compact and convex subsets such that* 0 *is an interior element of* Λ. *(iii) For* $p \in \mathbb{N}$, $\sum_{t=1}^{n} \boldsymbol{w}_t(p)\boldsymbol{w}_t(p)'$ *is positive definite with probability 1 and* $\mathbb{E}[\boldsymbol{w}_t(p)\boldsymbol{w}_t(p)']$ *is positive definite, where* $\boldsymbol{w}_t(p) := (\boldsymbol{x}_t(p)', \boldsymbol{d}_t')'$. *(iv)* $\Psi : \mathbb{R} \mapsto \mathbb{R}$ *is a non-polynomial analytic function. (v)* $\mathbb{E}[y_t^2] < \infty$, $\mathbb{E}[x_t^{2p}] < \infty$, *and there is a sequence of stationary and ergodic random variables* $\{s_t\}$ *such that (v.a)* $|u_t| \leq s_t$, *(v.b)* $\sup_{\delta \in \Delta} |\Psi(\delta x_t)| \leq s_t$, *(v.c)* $(\int_\Delta \Psi_t(\delta) d\mathbb{Q}(\delta))^2 \leq s_t$, *(v.d)* $\sup_{\delta \in \Delta} |(\partial \Psi(\delta x_t))/(\partial \delta)| \leq s_t$, *and (v.e) for some* $\kappa \geq 4\rho$, $\mathbb{E}[|s_t|^\kappa] < \infty$. $\square$

Assumptions 1, 2, and 3 are obtained by adapting the regularity conditions in Cho and White (2011) to the current polynomial model structure. Their model assumes a nonlinearity with respect to parameters, and we further simplify their assumptions by imposing the polynomial model structure used herein so that the limit results provided below can be obtained as corollaries of their theorems.

Indeed, testing $\mathcal{H}_0' : \lambda_* = 0$ is irregular because it involves Davies' (1977, 1987) identification problem. That is, if $\lambda_* = 0$, δ_* is not identified, δ_* is identified only when $\lambda_* \neq 0$, so that the null limit distribution of the t-test statistic testing $\mathcal{H}_0'$ becomes different from the standard normal distribution. The null limit distribution is found to be characterized by a Gaussian stochastic process indexed by the unidentified parameter δ. That is, if we let t_n be the standard t-test statistic testing $\mathcal{H}_0'$, it follows that

$$t_n \Rightarrow \sup_{\delta \in \Delta} \mathcal{G}(\delta)$$

under $\mathcal{H}_0'$ and Assumptions 1, 2, and 3, where $\mathcal{G}(\cdot)$ is a Gaussian stochastic process such that for every $\delta \in \Delta$, $\mathbb{E}[\mathcal{G}(\delta)] = 0$, and for each (δ, δ'),

$$\mathbb{E}[\mathcal{G}(\delta)\mathcal{G}(\delta')] = \frac{\rho(\delta, \delta')}{\{r(\delta, \delta)\}^{1/2}\{r(\delta', \delta')\}^{1/2}}$$

with

$$\rho(\delta, \delta') = \mathbb{E}[u_t^2 \Psi_t^*(\delta)\Psi_t^*(\delta')] \quad \text{and} \quad r(\delta, \delta') = \mathbb{E}[u_t^2]\mathbb{E}[\Psi_t^*(\delta)\Psi_t^*(\delta')].$$

Here, we let $\Psi_t^*(\delta) := \Psi_t(\delta) - \mathbb{E}[\Psi_t(\delta)\boldsymbol{w}_t(p)']\mathbb{E}[\boldsymbol{w}_t(p)\boldsymbol{w}_t(p)']^{-1}\boldsymbol{w}_t(p)$, and $u_t := y_t - \mathbb{E}[y_t|\mathcal{F}_t]$.

This limit distribution makes it inconvenient to apply the standard t-test statistic when testing $\mathcal{H}_0'$ against $\mathcal{H}_1'$. Note that the limit distribution is affected by too many factors in terms of data and model. If the error u_t is conditionally homoscedastic, the associated Gaussian process is a standard Gaussian process in the sense that for every $\delta \in \Delta$, $\mathcal{G}(\delta) \sim N(0, 1)$. However, this does not hold if u_t is conditionally heteroskedastic. Furthermore, there are many candidate analytic functions for $\Psi(\cdot)$. As Cho and White (2011) highlight, previous literature chooses different functions for $\Psi(\cdot)$, viz., logistic cumulative distribution function in White (1989), exponential function in Bierens (1990), and ridgelet function in Candés (2003), among others. Different covariance kernel structures are obtained for different analytic functions selected for $\Psi(\cdot)$, and this leads to different null limit distributions for the t-test statistic. The empirical researcher applying the standard t-test statistic has to apply different critical values to different models and data, making it more difficult to obtain the asymptotic critical values than the test statistic value itself. This aspect also analogously applies to other standard test statistics such as Wald, Lagrange multiplier, and QLR.

To overcome this, we utilize another testing method that applies the ELM proposed by Huang, Zhu, and Siew (2006). Cho and White (2011) note that the functional ordinary least squares (FOLS) estimator suggested by Cho, Huang, and White (2020) and Cho, Phillips, and Seo (2020) can be exploited to yield a straightforward statistic to test $\mathcal{H}_0'$ against $\mathcal{H}_1'$ by applying the ELM. As we detail below, the FOLS estimator has a limit distribution involved with integration, which lets the estimator follow a normal distribution asymptotically instead of being characterized by the Gaussian process. Using this property, we can convert the FOLS estimator into a Wald test statistic to follow a chi-squared distribution asymptotically under the null hypothesis. Here, the ELM are exploited to compute the involved integrations.

Specifically, we first note that for each δ, $\mathbb{E}[u_t\Psi_t(\delta)] = 0$ under $\mathcal{H}_0$ because $\Psi_t(\delta) := \Psi(\delta x_t)$ is measurable with respect to $\mathcal{F}_t$ and u_t is a martingale difference sequence (MDS) from the fact that $u_t := y_t - \mathbb{E}[y_t|\mathcal{F}_t]$, so that $\mathbb{E}[u_t|\mathcal{F}_t] = 0$. This implies that if $\Psi_t(\delta)$ is regressed against $(1, u_t)$, the estimated coefficient of u_t has to be zero irrespective of δ. Therefore, instead of testing $\mathcal{H}_0'$ vs. $\mathcal{H}_1'$, we opt to test

$$\mathcal{H}_0'' : \beta_*(\cdot) \equiv 0 \quad \text{vs} \quad \mathcal{H}_1'' : \beta_*(\cdot) \neq 0,$$

where for each $\delta \in \Delta$,

$$\begin{bmatrix} \alpha_*(\delta) \\ \beta_*(\delta) \end{bmatrix} := \begin{bmatrix} 1 & \mathbb{E}[u_t] \\ \mathbb{E}[u_t] & \mathbb{E}[u_t^2] \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[\Psi_t(\delta)] \\ \mathbb{E}[u_t\Psi_t(\delta)] \end{bmatrix}.$$

Here, $\mathbb{E}[u_t] = 0$ from that u_t is an MDS. Nevertheless, many of the entities on the right side are unknown to the researcher, necessitating the estimation of each expectation by its sample analog: for each $\delta \in \Delta$,

$$\begin{bmatrix} \widehat{\alpha}_n(\delta) \\ \widehat{\beta}_n(\delta) \end{bmatrix} := \begin{bmatrix} 1 & \sum_{t=1}^n \widehat{u}_t \\ \sum_{t=1}^n \widehat{u}_t & \sum_{t=1}^n \widehat{u}_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^n \Psi_t(\delta) \\ \sum_{t=1}^n \widehat{u}_t\Psi_t(\delta) \end{bmatrix},$$

where $\widehat{u}_t$ is the regression residual obtained from $\mathcal{M}_p^0$, viz.,

$$\widehat{u}_t := y_t - \boldsymbol{w}_t(p)' \left(\sum_{t=1}^n \boldsymbol{w}_t(p)\boldsymbol{w}_t(p)' \right)^{-1} \sum_{t=1}^n \boldsymbol{w}_t(p) y_t,$$

so that it also follows that $\sum_{t=1}^n \widehat{u}_t \equiv 0$ and consistently estimates u_t under $\mathcal{H}_0$. As there are a continuum number of δs in Δ, Cho and White (2011) integrate the above estimators using an adjunct probability measure $\mathbb{Q}(\cdot)$ and obtain the following limit distribution:

$$\begin{bmatrix} \sqrt{n}\int_\Delta(\widehat{\alpha}_n(\delta)-\alpha_*(\delta))d\mathbb{Q}(\delta) \\ \sqrt{n}\int_\Delta \widehat{\beta}_n(\delta)d\mathbb{Q}(\delta) \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 \\ 0 & \mathbb{E}[u_t^2] \end{bmatrix}^{-1} \begin{bmatrix} \int_\Delta \mathcal{G}_1(\delta)d\mathbb{Q}(\delta) \\ \int_\Delta \mathcal{G}_2(\delta)d\mathbb{Q}(\delta) \end{bmatrix} \tag{1}$$

under $\mathcal{H}_0$ and Assumptions 1, 2, and 3, where $\mathbb{Q}(\cdot)$ is an adjunct probability measure defined on Δ (which is selected by the researcher), and $\mathcal{G}_1(\cdot)$ and $\mathcal{G}_2(\cdot)$ are two independent Gaussian processes such that for each $\delta \in \Delta$, $\mathbb{E}[\mathcal{G}_1(\delta)] = 0$ and $\mathbb{E}[\mathcal{G}_2(\delta)] = 0$, and for each δ and $\delta' \in \Delta$,

$$\mathbb{E}[\mathcal{G}_1(\delta)\mathcal{G}_1(\delta')] = \tau(\delta,\delta') := \mathbb{E}[\Psi_t(\delta)\Psi_t(\delta')] - \mathbb{E}[\Psi_t(\delta)]\mathbb{E}[\Psi_t(\delta')], \quad \text{and} \quad \mathbb{E}[\mathcal{G}_2(\delta)\mathcal{G}_2(\delta')] = \rho(\delta,\delta').$$

This null limit distribution is indeed obtained by following the limit distribution theory of the FOLS estimator in Cho, Huang, and White (2020) and Cho, Phillips, and Seo (2020), in which they test the population mean function of functional data by estimating a parametric model by the FOLS estimator. More precisely, the FOLS estimator is obtained by minimizing the following functional mean squared errors:

$$Q_n(\gamma,\xi) := \frac{1}{2n}\sum_{t=1}^n \int_\Delta (\Psi_t(\delta) - \gamma - \xi\widehat{u}_t)^2 d\mathbb{Q}(\delta)$$

with respect to γ and ξ. If we let $(\widehat{\gamma}_n, \widehat{\xi}_n)$ denote the FOLS estimator minimizing $Q_n(\cdot,\cdot)$, it now follows that

$$\begin{bmatrix} \widehat{\gamma}_n \\ \widehat{\xi}_n \end{bmatrix} = \begin{bmatrix} 1 & \sum_{t=1}^n \widehat{u}_t \\ \sum_{t=1}^n \widehat{u}_t & \sum_{t=1}^n \widehat{u}_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^n \int_\Delta \Psi_t(\delta)d\mathbb{Q}(\delta) \\ \sum_{t=1}^n \int_\Delta \widehat{u}_t\Psi_t(\delta)d\mathbb{Q}(\delta) \end{bmatrix}$$
$$\stackrel{\text{a.s.}}{\to} \begin{bmatrix} \gamma_* \\ \xi_* \end{bmatrix} := \begin{bmatrix} 1 & 0 \\ 0 & \mathbb{E}[u_t^2] \end{bmatrix}^{-1} \begin{bmatrix} \int_\Delta \mathbb{E}[\Psi_t(\delta)]d\mathbb{Q}(\delta) \\ \int_\Delta \mathbb{E}[u_t\Psi_t(\delta)]d\mathbb{Q}(\delta) \end{bmatrix}$$

under Assumptions 1, 2, and 3, leading to that

$$\begin{bmatrix} \sqrt{n}(\widehat{\gamma}_n - \gamma_*) \\ \sqrt{n}(\widehat{\xi}_n - \xi_*) \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 \\ 0 & \mathbb{E}[u_t^2] \end{bmatrix}^{-1} \begin{bmatrix} \int_\Delta \mathcal{G}_1(\delta)d\mathbb{Q}(\delta) \\ \int_\Delta \mathcal{G}_2(\delta)d\mathbb{Q}(\delta) \end{bmatrix}.$$

Note that this limit distribution is now identical to that in (1), and $\xi_* = 0$ under $\mathcal{H}_0$.

Using the FOLS estimator, Cho and White (2011) test the null hypothesis using the Wald test statistic.

Note that integrating Gaussian processes produces a normally distributed random variable, implying that

$$\left[\begin{array}{c} \int_\Delta \mathcal{G}_1(\delta) d\mathbb{Q}(\delta) \\ \int_\Delta \mathcal{G}_2(\delta) d\mathbb{Q}(\delta) \end{array}\right] \sim N\left[\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} \sigma_\gamma^2 & 0 \\ 0 & \sigma_\xi^2 \end{array}\right)\right],$$

where

$$\sigma_\gamma^2 := \int_\Delta \int_\Delta \tau(\delta, \delta') d\mathbb{Q}(\delta) d\mathbb{Q}(\delta') \quad \text{and} \quad \sigma_\xi^2 := \int_\Delta \int_\Delta \rho(\delta, \delta') d\mathbb{Q}(\delta) d\mathbb{Q}(\delta').$$

This further implies that

$$\left[\begin{array}{c} \sqrt{n} \int_\Delta (\widehat{\alpha}_n(\delta) - \alpha_*(\delta)) d\mathbb{Q}(\delta) \\ \sqrt{n} \int_\Delta \widehat{\beta}_n(\delta) d\mathbb{Q}(\delta) \end{array}\right] \overset{\text{A}}{\sim} N\left[\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} \sigma_\gamma^2 & 0 \\ 0 & \sigma_\xi^2/\sigma_u^2 \end{array}\right)\right] \tag{2}$$

under $\mathcal{H}_0$, where $\sigma_u^2 := \mathbb{E}[u_t^2]$, and the null limit distribution of the FOLS estimator now motivates us to construct the following Wald test statistic:

$$\mathcal{W}_n := n\left(\frac{\widehat{\sigma}_{u,n}^2}{\widehat{\sigma}_{\xi,n}^2}\right)\left(\int_\Delta \widehat{\beta}_n(\delta) d\mathbb{Q}(\delta)\right)^2,$$

which follows a chi-squared distribution under $\mathcal{H}_0''$ and Assumptions 1, 2 and 3, where $\widehat{\sigma}_{\xi,n}^2$ and $\widehat{\sigma}_{u,n}^2$ are consistent estimators for σ_ξ^2 and σ_u^2, respectively. Under $\mathcal{H}_1''$, $\int_\Delta \beta_*(\delta) d\mathbb{Q}(\delta)$ is not necessarily equal to zero, and we can expect power for this test statistic from this aspect. Note that the test statistic is defined by following Wald's (1943) test principle. Due to its trivial null limit behavior, its empirical applicability is more straightforward relative to other test statistics requiring extra efforts to obtain the asymptotic critical values from the researcher, viz., the QLR test statistic in Baek, Cho, and Phillips (2015) and Cho and Phillips (2018).

Nevertheless, the burden of computing the Wald test statistic can be immense due to the involved integrations. To compute the statistic, it is necessary to compute the integration of $\Psi_t(\cdot)$ for each t, and if n is large, the involved computation burden can be immense.

Cho and White (2011) recommend resolving this issue by applying the ELM proposed by Huang, Zhu, and Siew (2006). That is, if we let $\{\delta_i : i = 1, 2, \ldots, m\}$ be a set of identically and independently distributed (IID) random variables following $\mathbb{Q}$ distribution, it follows that

$$\bar{\Psi}_{t,m} := \frac{1}{m} \sum_{i=1}^{m} \Psi_t(\delta_i) \overset{\text{a.s.}}{\to} \int_\Delta \Psi_t(\delta) d\mathbb{Q}(\delta)$$

by the law of large numbers, so that the FOLS estimator can be well approximated by

$$\begin{bmatrix} \widehat{\gamma}_{m,n} \\ \widehat{\xi}_{m,n} \end{bmatrix} := \begin{bmatrix} 1 & \sum_{t=1}^{n} \widehat{u}_t \\ \sum_{t=1}^{n} \widehat{u}_t & \sum_{t=1}^{n} \widehat{u}_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^{n} \bar{\Psi}_{t,m} \\ \sum_{t=1}^{n} \widehat{u}_t \bar{\Psi}_{t,m} \end{bmatrix}$$

if m is sufficiently large. To implement this plan, we formally assume the following condition:

Assumption 4 (ELM)**.** $\{\delta_j\}$ *is a sequence of identically and IID random variables defined on* $(\Delta, \mathcal{D}, \mathbb{Q})$. $\square$

Then, we can expect that

$$\mathcal{W}_{m,n} := n\widehat{\sigma}_n^2 \left(\frac{\widehat{\xi}_{m,n}^2}{\widehat{\sigma}_{\xi,m,n}^2} \right) \overset{\text{A}}{\sim} \mathcal{X}_1^2$$

under $\mathcal{H}_0'''$ and Assumptions 1, 2, 3, and 4, where

$$\widehat{\sigma}_{\xi,m,n}^2 := \frac{1}{n} \sum_{t=1}^{n} \widehat{u}_t^2 \bar{\Psi}_{t,m,n}^{*2} \quad \text{and}$$

$$\bar{\Psi}_{t,m,n}^{*} := \bar{\Psi}_{t,m} - \left(\frac{1}{n} \sum_{t=1}^{n} \bar{\Psi}_{t,m} \boldsymbol{w}_t(p) \right) \left(\frac{1}{n} \sum_{t=1}^{n} \boldsymbol{w}_t(p) \boldsymbol{w}_t(p)' \right)^{-1} \boldsymbol{w}_t(p).$$

Note that the only difference between $\mathcal{W}_n$ and $\mathcal{W}_{m,n}$ is in the fact that $\widehat{\xi}_{m,n}$ is used to estimate $\int_{\Delta} \widehat{\beta}_n(\delta) d\mathbb{Q}(\delta)$. Following Cho and White (2011), we also refer to it as Wald-ELM (WELM) test statistic.

Cho and White (2011) show by simulation that the null distribution of the WELM test statistic is well approximated by the chi-squared distribution by letting n and m be sufficiently large when their null model is the first-order autoregressive model. In addition, they also verify that the WELM test statistic displays a respectful power.

Before moving to the next section, we collect the main claims in this section into the following lemma:

Lemma 1. *Given Assumptions 1, 2, 3, and 4, (i)* $\mathcal{W}_n \overset{\text{A}}{\sim} \mathcal{X}_1^2$ *under* $\mathcal{H}_0''$ *; and for any positive sequence* $\{c_n\}$ *such that* $c_n = o(n)$*, if* $\int_{\Delta} \beta_*(\delta) d\mathbb{Q}(\delta) \neq 0$, $\mathbb{P}(\mathcal{W}_n > c_n) \to 1$ *under* $\mathcal{H}_1''$*; and (ii)* $\mathcal{W}_{m,n} \overset{\text{A}}{\sim} \mathcal{X}_1^2$ *under* $\mathcal{H}_0''$ *as* m *and* $n \to \infty$ *; and for any positive sequence* $\{c_n\}$ *such that* $c_n = o(n)$, $\mathbb{P}(\mathcal{W}_{m,n} > c_n) \to 1$ *under* $\mathcal{H}_1''$ *as* m *and* $n \to \infty$. $\square$

Lemma 1(*i* and *ii*) are made by Cho, Huang, and White (2020) and Cho and White (2011), respectively in a general context, but we provide their proofs in the Appendix to fit the current context.

3 Method 2: Sequential WELM Testing Procedure

In this section, we examine the sequential testing procedure combined with the WELM test statistic. The WELM test statistic developed by Cho and White (2011) focuses on specification testing. We further develop a testing methodology to estimate the most parsimonious polynomial model by combining the WELM test statistic with a sequential testing procedure.

To fix our idea on the sequential testing procedure, we first provide our model. Note that the model in Assumption 2 assumes a p-th degree polynomial model, and we now suppose that there are $\bar{p}$ polynomial models in total:

$$\mathcal{M}(\bar{p}) := \{\mathcal{M}_p : p = 1, 2, \ldots, \bar{p}\} \quad \text{and} \quad \mathcal{M}^0(\bar{p}) := \{\mathcal{M}_p^0 : p = 1, 2, \ldots, \bar{p}\},$$

so that $\mathcal{M}(\bar{p})$ and $\mathcal{M}^0(\bar{p})$ are the sets of the alternative and null models, respectively. Note that these model sets encompass the models in Assumption 2 as special cases. That is, $\mathcal{M}_p$ and $\mathcal{M}_p^0$ in Assumption 2 are elements of $\mathcal{M}(\bar{p})$ and $\mathcal{M}^0(\bar{p})$, respectively.

The most parsimonious model, which we seek to estimate by a sequential testing procedure, is obtained by testing smaller models against larger models sequentially. Specifically, the following procedure is proposed as our sequential testing procedure:

Step 1: We test $\mathcal{M}_1^0$ against $\mathcal{M}_1$ using the WELM test statistic. If $\mathcal{M}_1^0$ cannot be rejected at the level of significance $\boldsymbol{\alpha}$, we stop the sequential testing procedure and conclude that the conditional mean is linear with respect to x_t. Otherwise, we move to the next step. Note that the regression residual is computed by regressing y_t on $(1, x_t)$ when computing the WELM test.

Step 2: We test $\mathcal{M}_2^0$ against $\mathcal{M}_2$ using the WELM test statistic. If $\mathcal{M}_2^0$ cannot be rejected at the level of significance $\boldsymbol{\alpha}$, we stop the sequential testing procedure; otherwise, we move to the next step. In this way, we continue our testing procedure until we reach $p = \bar{p}$. As in the first step, the regression residual is computed by regressing y_t on $(1, x_t, \ldots, x_t^p)$ to compute the WELM statistic, which tests $\mathcal{M}_p^0$ against $\mathcal{M}_p$ for $p = 2, 3, \ldots, \bar{p} - 1$.

Step 3: We test $\mathcal{M}_{\bar{p}}^0$ against $\mathcal{M}_{\bar{p}}$ using the WELM test statistic. If $\mathcal{M}_{\bar{p}}^0$ cannot be rejected, we stop the sequential testing procedure to conclude that $\mathbb{E}[y_t|\mathcal{F}_t]$ is sufficiently well approximated by $\mathcal{M}_{\bar{p}}^0$; otherwise, we conclude that $\mathcal{M}^0(\bar{p})$ is entirely misspecified for $\mathbb{E}[y_t|\mathcal{F}_t]$.

Using this procedure, the most parsimonious and correct model is consistently detected. For a specific discussion, for some $\boldsymbol{\alpha}_*(p)$ and $\boldsymbol{\eta}_*$, we let p_* be defined as

$$p_* := \min\{p \in \mathbb{N} : \mathbb{E}[y_t|\mathcal{F}_t] = \boldsymbol{x}_t(p)'\boldsymbol{\alpha}_*(p) + \boldsymbol{d}_t'\boldsymbol{\eta}_*\}.$$

Note that p_* is the smallest polynomial degree such that the conditional mean is equal to the conditional mean. If $p > p_*$, the coefficients of degrees greater than p_* must be zero. Therefore, if $\mathcal{M}_{p_*}^0$ can be estimated, the most parsimonious polynomial model can be estimated, and the sequential testing procedure described above is designed to estimate p_*. Note that the WELM testing procedure has the GCR property by Stinchcombe and White (1998) and the sequential testing procedure starts model-testing from the smallest model to larger ones. Therefore, if the lower degree polynomial model is misspecified for the conditional mean, it will be consistently rejected by the WELM test statistic, so that we can expect to estimate the most parsimonious correct model by the sequential testing procedure. From this result, we obtain the following corollary:

Corollary 1. *Given Assumption 1, if Assumptions 2, 3, and 4 hold for each* $p \in \mathbf{P} := \{1, 2, \ldots, \bar{p}\}$ *and* $p_* \in \mathbf{P}$*, for any* $\epsilon > 0$*,* $\lim_{n\to\infty} \mathbb{P}(|\widehat{p}_n(\boldsymbol{\alpha}) - p_*| > \epsilon) = \boldsymbol{\alpha}$*, where* $\widehat{p}_n(\boldsymbol{\alpha})$ *is the polynomial degree estimator obtained by applying the WELM test statistics to the sequential testing procedure with the level of significance* $\boldsymbol{\alpha}$*.* □

As Corollary 1 is obvious from its structure, we do not prove it separately.

Corollary 1 implies that the degree estimator $\widehat{p}_n(\boldsymbol{\alpha})$ has a consistent estimation error that is equal to the level of significance $\boldsymbol{\alpha}$, so if this estimation error is not removed from the above procedure, the degree estimator is not consistent for p_*.

We further note that the significance level $\boldsymbol{\alpha}$ is selected by the researcher. We can let $\boldsymbol{\alpha}$ be dependent on the sample size n, so that if $\boldsymbol{\alpha}_n \to 0$ as $n \to \infty$, the degree estimation error can be allowed to converge to zero, leading to a consistent estimator. We contain this result in the following theorem:

Theorem 1. *Given Assumption 1, if Assumptions 2 and 3 hold for each* $p \in \mathbf{P} := \{1, 2, \ldots, \bar{p}\}$*,* $p_* \in \mathbf{P}$*, and* $\boldsymbol{\alpha}_n = 1 - C(c_n)$ *such that for some* $\delta \in (0, 1)$*,* $c_n = O(n^\delta)$*, then for any* $\epsilon > 0$*,* $\lim_{n\to\infty} \mathbb{P}(|\widehat{p}_n(\boldsymbol{\alpha}_n) - p_*| > \epsilon) = 0$*, where* $C(\cdot)$ *is the chi-squared distribution function with one degree of freedom.* □

The results in Corollary 1 and Theorem 1 correspond to the results using the sequential testing procedure in previous literature. Hosoya (1989) examines the sequential testing procedure for a set of models nested by larger models using the likelihood-ratio test statistic, so that the likelihood-ratio test statistics can be sequentially applied using the chi-squared null limit distributions. Nevertheless, the models assumed by

Hosoya (1989) do not have the identification problem as we examine herein. Theorem 2 of Cho and Phillips (2018) also provides a result analogous to Theorem 1 of the current study, but their conditions are more relaxed herein in the following senses. First, they apply the QLR test statistic for their sequential testing problem, which compares the mean square errors obtained from the null and alternative models such that the alternative model is constructed by letting $\Psi(\delta x_t) = x_t^{\delta}$. They show that a multifold identification problem exists under the null that the conditional mean is correctly specified by the polynomial model, and due to this, their QLR test statistic weakly converges to a functional of a Gaussian stochastic process. Consequently, the null limit distribution of their test statistic does not follow a chi-squared distribution. The null limit distribution is obtained using the weighted bootstrapping proposed by Hansen (1996), making its application inconvenient. Second, the particular form of the power transformation for $\Psi_t(\cdot)$ restricts their applications. If x_t is negatively valued, it may not be properly defined. Note that $x_t^{\delta} = \exp(\delta \log(x_t))$, which is defined only when $x_t > 0$, so that the application of their methodology is restrictive if x_t can be negatively valued. Finally, the level of significance $\boldsymbol{\alpha}_n$ is assumed to slowly converge to zero relative to the convergence rate herein. They require that $\log(\boldsymbol{\alpha}_n)/n \to 0$ in addition to $\boldsymbol{\alpha}_n \to 0$, whereas the latter is only assumed in Theorem 1. This requirement is imposed mainly because the null limit distribution of the QLR test statistic is characterized by the maximum of the squared Gaussian process. The tail distribution of the maximum is approximated by associating it with that from the squared fractional Brownian motion via Slepian inequality. $\log(\boldsymbol{\alpha}_n)/n \to 0$ is required to yield a sequence of critical values uniformly dominated by those from the squared fractional Brownian motion. On the contrary, our sequential testing procedure does not need to satisfy this additional condition.

4 Results: Monte Carlo Simulations

In this section, we illustrate the sequential WELM testing procedure via conducting Monte Carlo simulations using stationary time-series observations.

4.1 Linear Function and Sequential Testing Procedure

Without loss of generality, we first suppose the following dynamic and stationary time-series DGP:

$$y_t = \alpha_{0*} + \alpha_{1*} x_t + \eta_* y_{t-1} + \epsilon_t$$

where $x_t = \phi_* x_{t-1} + u_t$, $(\epsilon_t, u_t) \sim$ IID $N(0, \sigma_*^2 I_2)$, $y_0 \sim N(0, \sigma_y^2)$, $x_0 \sim$ IID $N(0, \sigma_x^2)$, and $t = 1, 2, ..., n$ such that $|\phi_*| < 1$ and $\eta_*| < 1$. Note that the last two inequality conditions are imposed for the stationarity of data.

Given this DGP condition, we let our model be constructed by polynomial models. We first consider a linear model as the first-degree polynomial model. For this purpose, we let the explanatory variable vector $\boldsymbol{x}_t(p)$ be simply x_t, so that $p = 1$, and we also let d_t be the lagged dependent variable y_{t-1}. Therefore, if we let $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \eta)'$, the null model $\mathcal{M}_1^0$ becomes $\{\Phi(\cdot,) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, where $\Phi(\mathbf{X}_t, \boldsymbol{\theta}) := \alpha_0 + \alpha_1 x_t + \eta y_{t-1}$. For our alternative model, we let the exponential function be $\Psi(\cdot)$ so that

$$\mathcal{M}_1 := \{f(\cdot; \boldsymbol{\theta}, \lambda, \delta) : (\boldsymbol{\theta}, \lambda, \delta) \in \boldsymbol{\Theta} \times \Lambda \times \Delta\} \quad \text{and}$$

$$f(\mathbf{X}_t; \boldsymbol{\theta}, \lambda, \delta) := \alpha_0 + \alpha_1 x_t + \eta y_{t-1} + \lambda \exp(\delta x_t)$$

such that $\Lambda := [-\bar{\lambda}, \bar{\lambda}]$ and $\Delta := [\underline{\delta}, \bar{\delta}]$. Next, we compute the WELM test statistic $\widehat{\mathcal{W}}_{n,m}$ by first approximating $\bar{\Psi}_{m,t} := \int_\Delta \Psi(\mathbf{X}_t' \boldsymbol{\delta}) d\mathbb{Q}(\boldsymbol{\delta})$ via $\bar{\Psi}_{m,t} := m^{-1} \sum_{i=1}^m \exp(\delta x_t)$ and next by letting $\boldsymbol{w}_t(1) := [1, x_t, y_{t-1}]'$, where we suppose that $\mathbb{Q}$ is a probability measure uniformly distributed on $\Delta = \left[\underline{\delta}, \bar{\delta}\right]$. Note that this linear model is correctly specified for the DGP. Therefore, we should expect that the WELM test statistic rejects this model $\boldsymbol{\alpha} \times 100\%$ asymptotically when the level of significance is $\boldsymbol{\alpha}$.

Next, we extend the model scope to higher-degree polynomial models. For this purpose, we further let $\boldsymbol{x}_t(p) := (1, x_t, x_t^2, ..., x_t^p)$ to specify the following null and alternative models:

$$\mathcal{M}_p^0 = \{\alpha_0 + \alpha_1 x_t + \alpha_2 x_t^2 + \cdots + \alpha_p x_t^p + \eta y_{t-1} : \boldsymbol{\theta} := (\alpha_0, \ldots, \alpha_p, \eta) \in \boldsymbol{\Theta}(p)\}, \quad \text{and}$$

$$\mathcal{M}_p = \{\alpha_0 + \alpha_1 x_t + \alpha_2 x_t^2 + \cdots + \alpha_p x_t^p + \eta y_{t-1} + \lambda \exp(\delta x_t) : \boldsymbol{\theta} := (\alpha_0, \ldots, \alpha_p, \eta) \in \boldsymbol{\Theta}(p), \lambda \in \Lambda, \delta \in \Delta\}.$$

Given this, we further let $\boldsymbol{w}_t(p) := [1, x_t, x_t^2, \ldots, x_t^p, y_{t-1}]'$ to compute the WELM test statistic to test the p-th degree polynomial model. Note that if $p = 1$, the WELM test statistic is exactly the same as that obtained using the linear model. For $p = 2, 3$ and $\bar{p} = 4$, the null models $\mathcal{M}_p^0$ are correctly specified, so that the WELM test statistics are also expected to reject the null model $\boldsymbol{\alpha} \times 100\%$ asymptotically. By this, we construct the following sets of alternative and null models:

$$\mathcal{M}(4) := \{\mathcal{M}_p : p = 1, 2, \ldots, 4\} \quad \text{and} \quad \mathcal{M}^0(4) := \{\mathcal{M}_p^0 : p = 1, 2, \ldots, 4\}$$

to apply the sequential testing procedure.

For this DGP and the models, we conduct simulations and report the simulation results in Table 1, which are obtained by applying the sequential testing procedure. We generate data by letting $(\alpha_{1*}, \eta_*, \phi_*, \sigma_*^2) = (0.5, 0.5, 0.5, 1.0)$ and also let the levels of significance $\boldsymbol{\alpha}$ be 10%, 5%, and 1%. Given these simulation environments, we examine the empirical rejection rates of the WELM test statistic for $n = 50$, 100, 200, 500, 1000, 2000, and 5000. We also let $m = 5,000$, and the total number of experiments is 5,000.

The simulation results can be summarized as follows. First, the sequential testing procedure stops mostly at the first step, which implies that the sequential WELM test identifies the correct degree of the unknown polynomial function correctly. More specifically, as the sample size n increases, the WELM test statistic detects the linear model as the correct model approximately $(1 - \boldsymbol{\alpha}) \times 100\%$. This aspect is observed irrespective of the sample size, so that we can expect that the WELM test statistic controls the type-I error precisely; thereby, the most parsimonious correct model can be efficiently estimated. Second, even when the sequential testing procedure estimates the models whose polynomial degrees are greater than unity, most of the selected models are quadratic models. This implies that the sequential testing procedure has a strong tendency to select the next most parsimonious model for the conditional mean function. As a result, mostly selected models are linear or quadratic functions. Third, as the level of significance $\boldsymbol{\alpha}$ decreases, more precise estimation results are delivered from the experiments. However, this result in another way implies that the estimation error cannot be eliminated altogether as long as the level of significance is fixed.

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

We therefore conduct another simulation by letting the level of significance be dependent upon the sample size. Specifically, we let the level of significance $\boldsymbol{\alpha}_n$ be $n^{-1/2}$, n^{-1}, $n^{-3/2}$, and n^{-2}. Note that for these levels of significance, $\boldsymbol{\alpha}_n$ reduces to zero n increases, so that the sequential testing procedure is expected to eliminate the estimation error asymptotically. Among the levels of significance, n^{-2} approaches zero more quickly than the other levels of significance. Table 2 reports the simulation results obtained by 5,000 experiments. Figures in the first panel denote $\widehat{P}_n(\boldsymbol{\alpha}_n) := r^{-1}\sum_{i=1}^{r}\mathbb{I}(\hat{p}_{n,i} = 1)$, where r denotes the total number of experiments set to be $5,000$, and $\hat{p}_{n,i}$ denotes the degree estimated by the sequential testing procedure from the i-th experiment when the level of significance is $\boldsymbol{\alpha}_n$. Here, $\mathbb{I}(\cdot)$ denotes the indicator function. Note that for each plan for the level of significance $\boldsymbol{\alpha}_n$, $\widehat{P}_n(\boldsymbol{\alpha}_n)$ estimates the empirical probability for the estimated degree by the sequential testing procedure to be equal to 100%. The other figures in parentheses denote the hypothetical proportion measured by $(1 - \boldsymbol{\alpha}_n) \times 100\%$. Note that as $\boldsymbol{\alpha}_n$ reduces to zero more quickly, the hypothetical proportion more quickly arrives at 100%. In addition to the sequential testing procedure, we also compare these estimation results with standard information criterion-based estimations using Akaike's information criterion (AIC), Bayesian information criterion (BIC), and small-sample corrected AIC (AICc). These information criteria are applied to the null models $\mathcal{M}_p^0$ with $p = 1, 2, 3, 4$ and compute the proportions measured by $\widetilde{P}_n := r^{-1} \sum_{i=1}^{r} \mathbb{I}(\widetilde{p}_{n,i} = 1)$, where $\widetilde{p}_{n,i}$ denotes the degree selected by the information criterion. Figures in the second panel report the proportions estimated by the information criteria. Finally, we apply the same information criteria to the alternative models $\mathcal{M}_p$ with $p = 1, 2, 3, 4$ and report the estimated proportions in the third panel that are obtained by the same methodology. We distinguish them from the earlier information criteria by attaching "′" to AIC, BIC, and AICc, so that AIC′, BIC′, and AICc′ denote the information criteria applied to the alternative models.

The simulation results reported in Table 2 can be summarized as follows. First, for every significance level $\boldsymbol{\alpha}_n$, the distance between $\widehat{P}_n(\boldsymbol{\alpha}_n)$ and $(1 - \boldsymbol{\alpha}_n)$ gets close to zero as the sample size n increases. This suggests that the first-degree polynomial model is successfully estimated by the sequential estimation procedure. Second, the distance between $\widehat{P}_n(\boldsymbol{\alpha}_n)$ and $(1 - \boldsymbol{\alpha}_n)$ is closest to zero when the plan for the level of significance is set to be $\boldsymbol{\alpha}_n = n^{-2}$. This implies that the sequential testing procedure can estimate the first-degree polynomial model more precisely by letting the level of significance converge to zero more quickly. Third, as the second panel shows, BIC converges to 100% with the increase of sample size, but AIC and AICc are not so fast as BIC. Fourth, as the third panel shows, BIC′ performs similarly to BIC, whereas AIC′ and AICc′ perform a little worse than AIC and AICc, respectively. Finally, when comparing the BIC (or BIC′) with the sequential WELM testing procedure, the performance of the information criteria is inferior to those of the sequential testing procedure when $\boldsymbol{\alpha}_n$ converges to zero quickly. Specifically, if we let $\boldsymbol{\alpha}_n$ be $n^{-3/2}$ or n^{-2}, the performance of the sequential testing procedure is better than that obtained by BIC uniformly in the sample size. In contrast, if we let $\boldsymbol{\alpha}_n$ be $n^{-1/2}$, the performance of BIC is superior to the sequential testing procedure uniformly in the sample size. In the middle, if $\boldsymbol{\alpha}_n$ reduces to zero at a moderate rate, viz., n^{-1}, the performance of the sequential testing procedure is dependent upon the sample size n. That is, if n is relatively small, the sequential testing procedure performs better than BIC, but if n is relatively large, BIC performs better than the sequential testing procedure. This aspect implies that letting the level of significance converge to zero as quickly as possible can produce the best estimation result if the first-degree polynomial model is a correct model.

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

4.2 Quadratic Function and Sequential Testing Procedure

We extend the earlier simulation by conducting another simulation. We examine a different DGP. Specifically, we suppose that data are generated by $y_t = \alpha_{1*}x_t + \alpha_{2*}x_t^2 + \eta_* y_{t-1} + \epsilon_t$, where $x_t = \phi_* x_{t-1} + u_t$, $(x_0, y_0) \sim$ IID $N(0, I_2)$, $(\epsilon_t, u_t) \sim$ IID $N(0, \sigma_*^2 I_2)$, and $(\alpha_{1*}, \alpha_{2*}, \eta_*, \phi_*, \sigma_*^2) = (0.5, 0.5, 0.5, 0.5, 1.0)$. Therefore, the first-degree polynomial model is now incorrectly specified, but the second-, third-, and fourth-degree polynomial models are correctly specified. Hence, the desired sequential testing procedure should estimate the second-degree polynomial model as the most parsimonious and correctly specified model. We attempt this second simulation to verify whether the lessons we could have obtained from the simulations in section 4.1 are still valid for other DGPs.

As before, we first conduct simulations by fixing the levels of significance and next by letting them depend on the sample size. The simulation results for the first and second cases are reported in Tables 3 and 4, respectively, and they are obtained under the same simulation environments as for Tables 1 and 2, respectively.

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

We can summarize the simulation results as follows. First, Table 3 shows that the portion of the linear model selected by the sequential testing procedure decreases to zero as the sample size n increases. For each level of significance, 10%, 5%, and 1%, the first-degree polynomial model is selected less and less as n increases. This aspect implies that the WELM test statistic has a consistent power to reject the misspecified model. Second, according to Table 3, the second-degree polynomial model is asymptotically selected $(1 - \boldsymbol{\alpha}) \times 100\%$, and this implies that the WELM test statistic controls the type-I error efficiently. Hence, the most parsimonious and correctly specified model can be consistently selected by the sequential testing procedure. Third, as before, the estimation error incurred by the sequential testing procedure cannot be removed altogether as long as the level of significance is fixed irrespective of the sample size. Fourth, Table 4 reports the portions of the polynomial degrees estimated by the sequential WELM testing procedure with the significance levels dependent on the sample size, and the information criteria. As we can see, for $\boldsymbol{\alpha}_n = n^{-1}$, $\boldsymbol{\alpha}_n = n^{-6/4}$, and $\boldsymbol{\alpha}_n = n^{-2}$, the distance between $\widehat{P}_n(\boldsymbol{\alpha}_n)$ and $(1 - \boldsymbol{\alpha}_n)$ gets close to zero with the increase in sample size. Fifth, if the sample size is relatively small, slowly converging levels of significance better estimate the correct degree than the quickly converging levels. For example, if $n = 50$, letting $\boldsymbol{\alpha}_n = n^{-1/2}$ produces higher portions than that obtained by letting $\boldsymbol{\alpha}_n$ be n^{-2}. Nevertheless, as the sample size increases, they show different estimation patterns. For $\boldsymbol{\alpha}_n = n^{-1/2}$, the proportion converges to 100% very slowly, whereas for $\boldsymbol{\alpha}_n = n^{-2}$, the proportion converges to 100% very quickly, implying that the plans for the level of significance have to be carefully chosen to apply them to the sequential testing procedure. If a relatively large sample data set is examined, the correct degree of polynomial model can be better estimated by letting the level of significance converge to zero quickly. On the contrary, if the sample size is small, a level of significance converging to zero relatively slowly should be chosen. Sixth, we also compare the performances of the information criteria and observe that BIC overall performs better than AIC and AICc, and the same thing holds among AIC′, BIC′, and AICc′. We also note that BIC always better estimates than BIC′. Finally, we compare the simulation results using the sequential testing procedure with BIC. If the sample size is small, BIC always dominates all estimation results made by the sequential testing procedures, but if the sample size is large enough, say more than 2,000, the sequential testing procedure with the level of significance converging to zero quickly better estimates than BIC. This simulation result is somewhat different from what we could observe in section 4.1. The sequential testing procedure does not alway performing better than BIC. If the polynomial function has a lower degree in the DGP, the sequential testing procedure may perform better than the information criterion. In particular, if the sample size is sufficiently large, use of the sequential testing procedure appears more amenable.

$<<<<<$ Insert Table 4 around here. $>>>>>$

4.3 Misspecified Models and Sequential Testing Procedure

As our final simulation, we now suppose that none of the models are correctly specified by supposing that $y_t = \pi_* \cos(y_{t-1}) + \epsilon_t$, where $y_0 \sim N(0, \sigma^2_{y_0})$ and $u_t \sim$ IID $N(0, \sigma^2_u)$. Here, we let $(\pi_*, \sigma^2_{y_0}, \sigma^2_u) = (1.0, 1.0, 1.0)$. We apply the same models as before and select the best model using the sequential testing procedure. Note that $\cos(\cdot)$ function is expressed as an infinite-degree polynomial function by Taylor's expansion, so that the fourth-degree polynomial model cannot be correctly specified for this DGP. This implies that the sequential testing procedure is expected to estimate a degree greater than 4. Our primary interest in this simulation is in investigating how the earlier finite sample properties of the sequential testing procedure are modified by this new DGP condition. As the model conditions and simulation environments are the same as before, we do not iterate.

$<<<<<$ Insert Table 5 around here. $>>>>>$

We report the simulation results in Tables 5 and 6. Table 5 is obtained by fixing the levels of significance, and Table 6 is obtained by letting the levels of significance depend on the sample size. The simulation results are summarized as follows. First, as the sample size n increases, the empirical rejection rates also increase for each degree $p = 1, 2$, and 3, and the sequential testing procedure concludes that the polynomial degree is greater than or equal to 4 for most experiments. For example, if $n = 2,000$, the sum of the portions of $p = 1, 2, 3$ are only 0.58%, 4.82%, and 11.55% for significance levels 10%, 5%, and 1%, respectively, and they further decrease as n increases to $5,000$. This result indicates that the power of the sequential WELM testing procedure performs well if the sample size is sufficiently large. Second, when an incorrect model is selected, the quadratic model is overall selected more often than the linear or cubic models. That is, the second-degree polynomial model is more preferred to the first- and third-degree polynomial models. This is mainly because the cosine function is an even function around zero, so that the quadratic function may better approximate the cosine function when the sample size is not sufficiently large. Third, we now let the levels of significance depend on the sample size and examine the simulation results reported in Table 6. As we can see, the distance between $\widehat{P}_n(\boldsymbol{\alpha}_n)$ and $(1-\boldsymbol{\alpha}_n)$ reduces with the rise of sample size n. Although the distance is not so close to zero as in Tables 2 and 4, the distance reduces. Fourth, if n is small, slowly converging levels of significance better estimate than the quickly converging plans. Nevertheless, as n increases, the portions converge to 100% more quickly when we let $\boldsymbol{\alpha}_n$ be n^{-2} than when we let $\boldsymbol{\alpha}_n$ be $n^{-1/2}$. Hence, if the data set has a large sample size, the level of significance converging to zero relatively quickly should be chosen. This is the same observation in section 4.2. Fifth, we now compare the performances of the information criteria and observe that AIC overall performs better than BIC and AICc, and the same thing holds among AIC′, BIC′, and AICc′. We also note that AIC always better estimates than AIC′. Finally, we compare the simulation results using the sequential testing procedure with AIC. AIC always dominates all estimations made by the sequential testing procedures. This simulation result implies that BIC is not always the best performing information criterion, and the sequential testing procedure can dominate BIC even when the sample size is small. Furthermore, if all models of consideration are misspecified, it is not easy to draw regular patterns among the sequential testing procedure and the information criteria.

<<<<< Insert Table 6 around here. >>>>>

5 Conclusion

We apply the Wald test statistic assisted by the ELM to test correct model assumption and estimate a close approximate to the conditional mean. When testing for model misspecification of conditional mean, the omnibus test statistics typically weakly converge to a Gaussian stochastic process under the null hypothesis that the model is correctly specified. This aspect makes their applications inconvenient. We define the Wald test statistic using the functional regression and apply the ELM to compute the test statistic efficiently; we refer to it as the Wald-ELM (WELM) test statistic following Cho and White (2011). The WELM test statistic is GCR and follows a chi-squared distribution under the null. We further apply the WELM test statistic to a sequential testing procedure to search for an approximate conditional expectation and conduct extensive Monte Carlo experiments to evaluate its performance. Via simulation, we verify that if the candidate polynomial models are correctly specified, the sequential WELM testing procedure estimates the most parsimonious and correct model consistently. Further, it consistently rejects all candidate models if none of the polynomial models are correctly specified. We further compare the performance of the standard information criteria such as Bayesian and Akaike information criteria, and its small-sample adjusted version. From this comparison, we find that the model estimation via the sequential testing procedure has a competitive power in estimating the most parsimonious and correct model.

Appendix

Proof of Lemma 1: (*i*) We first note that theorem 2 of Cho, Huang, and White (2020) implies that

$$\begin{bmatrix} \sqrt{n}\int_\Delta(\widehat{\alpha}_n(\delta)-\alpha_*(\delta))d\mathbb{Q}(\delta) \\ \sqrt{n}\int_\Delta(\widehat{\beta}_n(\delta)-\beta_*(\delta))d\mathbb{Q}(\delta) \end{bmatrix} \overset{\text{A}}{\sim} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\gamma^2 & 0 \\ 0 & \sigma_\xi^2/\sigma_u^2 \end{pmatrix}\right], \tag{3}$$

so that

$$\sqrt{n}\int_\Delta \widehat{\beta}_n(\delta)d\mathbb{Q}(\delta) \overset{\text{A}}{\sim} N(0,\sigma_\xi^2/\sigma_u^2)$$

under $\mathcal{H}_0''$. Therefore,

$$\mathcal{W}_n := n\left(\frac{\widehat{\sigma}_{u,n}^2}{\widehat{\sigma}_{\xi,n}^2}\right)\left(\int_\Delta \widehat{\beta}_n(\delta)d\mathbb{Q}(\delta)\right)^2 \overset{\text{A}}{\sim} \mathcal{X}_1^2$$

under $\mathcal{H}_0''$. In contrast, under $\mathcal{H}_1''$, $a_* := \int_\Delta \beta_*(\delta)d\mathbb{Q}(\delta) \neq 0$, so that (3) implies that

$$\left(\frac{\widehat{\sigma}_{u,n}^2}{\widehat{\sigma}_{\xi,n}^2}\right)\left(\sqrt{n}\int_\Delta \widehat{\beta}_n(\delta)d\mathbb{Q}(\delta) - \sqrt{n}a_*\right)^2 \overset{\text{A}}{\sim} \mathcal{X}_1^2,$$

where the left side is asymptotically identical to

$$\mathcal{W}_n - 2n\left(\frac{\widehat{\sigma}_{u,n}^2}{\widehat{\sigma}_{\xi,n}^2}\right)a_*\left(\int_\Delta \widehat{\beta}_n(\delta)d\mathbb{Q}(\delta) - a_*\right) - n\left(\frac{\widehat{\sigma}_{u,n}^2}{\widehat{\sigma}_{\xi,n}^2}\right)a_*^2,$$

so that $\mathcal{W}_n = n(\widehat{\sigma}_{u,n}^2/\widehat{\sigma}_{\xi,n}^2)a_*^2 + O_{\mathbb{P}}(\sqrt{n})$, and this implies that $\mathcal{W}_n/n = (\sigma_u^2/\sigma_\xi^2)a_*^2 + o_{\mathbb{P}}(1)$ under $\mathcal{H}_1''$ from the fact that both $\widehat{\sigma}_{u,n}^2$ and $\widehat{\sigma}_{\xi,n}^2$ are consistent for $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_\xi^2$, respectively. Therefore, for any $c_n = o(n)$,

$\lim_{n\to\infty} \mathbb{P}(\mathcal{W}_n > c_n) = 1$ as desired.

(*ii*) Note that

$$\mathcal{W}_{m,n} := n\widehat{\sigma}_n^2 \left(\frac{\widehat{\xi}_{m,n}^2}{\widehat{\sigma}_{\xi,m,n}^2} \right),$$

where

$$\begin{bmatrix} \widehat{\gamma}_{m,n} \\ \widehat{\xi}_{m,n} \end{bmatrix} := \begin{bmatrix} 1 & \sum_{t=1}^n \widehat{u}_t \\ \sum_{t=1}^n \widehat{u}_t & \sum_{t=1}^n \widehat{u}_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^n \bar{\Psi}_{t,m} \\ \sum_{t=1}^n \widehat{u}_t \bar{\Psi}_{t,m} \end{bmatrix}, \quad \widehat{\sigma}_{\xi,m,n}^2 := \frac{1}{n}\sum_{t=1}^n \widehat{u}_t^2 \bar{\Psi}_{t,m,n}^{*2},$$

$$\bar{\Psi}_{t,m,n}^* := \bar{\Psi}_{t,m} - \left(\frac{1}{n}\sum_{t=1}^n \bar{\Psi}_{t,m}\boldsymbol{w}_t(p)\right)\left(\frac{1}{n}\sum_{t=1}^n \boldsymbol{w}_t(p)\boldsymbol{w}_t(p)'\right)^{-1} \boldsymbol{w}_t(p), \quad \text{and}$$

$$\bar{\Psi}_{t,m} := \frac{1}{m}\sum_{i=1}^m \Psi_t(\delta_i).$$

Therefore, if for each t, $\bar{\Psi}_{t,m} \overset{\text{a.s.}}{\to} \int_\Delta \Psi_t(\delta)\mathbb{Q}(\delta)$, it follows that as $m \to \infty$, $\widehat{\xi}_{m,n} \overset{\text{a.s.}}{\to} \widehat{\xi}_n$ and $\widehat{\sigma}_{\xi,m,n}^2 \overset{\text{a.s.}}{\to} \widehat{\sigma}_{\xi,n}^2$, so that $\mathcal{W}_{m,n} \to \mathcal{W}_n$ as $m \to \infty$, and the desired result follows from Lemma 1(*i*).

We now note that the law of large numbers can apply to $\bar{\Psi}_{t,m}$, so that as $m \to \infty$, $\bar{\Psi}_{t,m} \overset{\text{a.s.}}{\to} \mathbb{E}_{\mathbb{Q}}[\Psi_t(\delta_i)]$ from the fact that δ_i is drawn from $\mathbb{Q}(\cdot)$ and independent of the data observations. Furthermore, $\mathbb{E}_{\mathbb{Q}}[\Psi_t(\delta_i)]$ $= \int_\Delta \Psi_t(\delta)\mathbb{Q}(\delta)$ by the definition of the expectation. This completes the proof. ■

Proof of Corollary 1: For notational simplicity, we let $\mathcal{W}_{m,n}(p)$ denote the WELM test statistic testing $\mathcal{M}_p^0$ against $\mathcal{M}_p$.

We now note that from the definition of $\widehat{p}_n(\boldsymbol{\alpha})$, viz.,

$$\widehat{p}_n(\boldsymbol{\alpha}) := \arg\min\{\mathbf{P} : \mathcal{W}_{m,n}(p) \leq cv(\boldsymbol{\alpha})\},$$

where $cv(\boldsymbol{\alpha})$ is the critical value obtained from the chi-squared distribution with one degree of freedom and level of significance $\boldsymbol{\alpha}$, if $p < p_*$,

$$\lim_{n\to\infty} \mathbb{P}(\widehat{p}_n(\boldsymbol{\alpha}) = p) = 0, \tag{4}$$

because if $p < p_* \in \mathbf{P}$, the model is misspecified, so that for any positive sequence $\{c_n\}$ such that $c_n = o(n)$,

$$\mathbb{P}(\mathcal{W}_{m,n}(p) > c_n) \to 1 \tag{5}$$

as $n \to \infty$. Therefore, it follows that

$$\lim_{n\to\infty} \mathbb{P}(\mathbb{I}(p = p_*)|\mathcal{W}_{m,n}(p)) = 0,$$

where $\mathbb{P}(\mathbb{I}(p = p_*)|\mathcal{W}_{m,n}(p))$ denotes the conditional probability for the hypothesized polynomial degree p to be equal to p_* conditional on that the hypothesis $\mathcal{M}_p^0$ is tested by $\mathcal{W}_{m,n}(p)$, implying (4). Therefore,

$$\lim_{n\to\infty} \mathbb{P}(\widehat{p}_n(\boldsymbol{\alpha}) \geq p_*) = 1. \tag{6}$$

In contrast, if $p \geq p_*$, the model is correctly specified and $\mathcal{W}_{m,n}(p) \overset{\text{A}}{\sim} \mathcal{X}_1^2$ from the structure of the WELM test statistic, so that

$$\mathbb{P}(\mathcal{W}_{m,n}(p) > cv(\boldsymbol{\alpha})) \to \boldsymbol{\alpha}$$

as $n \to \infty$. That is, it follows that for each $p \geq p_*$,

$$\lim_{n\to\infty} \mathbb{P}\left(\mathbb{I}(p = p_*)|\mathcal{W}_{m,n}(p)\right) = 1 - \boldsymbol{\alpha}. \tag{7}$$

Therefore, the definition of $\widehat{p}_n(\boldsymbol{\alpha})$, (6), and (7) imply that $\widehat{p}_n(\boldsymbol{\alpha})$ consistently estimates the minimum value of $\{p \in \mathbf{P} : p \geq p_*\}$ with probability $1 - \boldsymbol{\alpha}$, which is p_*. This implies the desired result. ■

Proof of Theorem 1: Let cv_n be the critical value corresponding to $\boldsymbol{\alpha}_n$, viz.,

$$cv_n = C^{-1}(1 - \boldsymbol{\alpha}_n),$$

which is $O(n^\delta)$ and also $o(n)$ because $\delta \in (0, 1)$ by the given condition. Therefore, for each p, (5) implies that

$$\mathbb{P}(\mathcal{W}_{m,n}(p) > cv_n) \to 1,$$

implying that

$$\lim_{n\to\infty} \mathbb{P}(\widehat{p}_n(\boldsymbol{\alpha}_n) \geq p_*) = 1. \tag{8}$$

Contrary to this, if $p \geq p_*$, $\mathcal{W}_{m,b}(p) \overset{\text{A}}{\sim} \mathcal{X}_1^2$, so that

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{W}_{m,n}(p) > cv_n) - \boldsymbol{\alpha}_n = 0,$$

and $\boldsymbol{\alpha}_n = o(1)$ because $cv_n = O(n^{\delta})$ for some $\delta > 0$. Therefore, for each $p \geq p_*$, it follows that

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{W}_{m,n}(p) > cv_n) = 0, \tag{9}$$

and this and (8) imply that

$$\lim_{n\to\infty} \mathbb{P}(\widehat{p}_n(\boldsymbol{\alpha}_n) = p_*) - (1 - \boldsymbol{\alpha}_n) = 0,$$

because $\widehat{p}_n(\boldsymbol{\alpha}_n)$ is defined to be the smallest degree among the degrees satisfying (9). The final equation now implies that $\mathbb{P}(\widehat{p}_n(\boldsymbol{\alpha}_n) = p_*) = 1 + o(1)$. This completes the proof. ■

References

BAEK, Y.I., CHO, J.S., AND PHILLIPS, P.C. (2015): "Testing Linearity Using Power Transforms of Regressors," *Journal of Econometrics*, 187, 376–384.

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

CANDÈS, E.J. (2003): "Ridgelets: Estimating with Ridge Functions," *Annals of Statistics*, 31, 1561–1599.

CHO, J.S., HUANG, M., AND WHITE, H. (2020): "Testing for a Constant Mean Function Using Functional Regression," Discussion Paper, School of Economics, Yonsei University.

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

CHO, J.S. AND PHILLIPS, P.C.B. (2018): "Sequentially Testing Polynomial Model Hypotheses Using Power Transforms of Regressors," *Journal of Applied Econometrics*, 33, 141–159.

CHO, J.S., PHILLIPS, P.C.B., AND SEO, J. (2020): "Parametric Conditional Mean Inference with Functional Data Applied to Lifetime Income Curves," Discussion Paper, School of Economics, Yonsei University.

CHO, J.S. AND WHITE, H. (2011): "Testing Correct Model Specification Using Extreme Learning 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.

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

HOSOYA, Y. (1989): "Hierarchical Statistical Models and a Generalized Likelihood Ratio Test," *Journal of Royal Statistical Society, Series B*, 51, 435–447.

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

KEENAN, D.M. (1985): "A Tukey Nonadditivity-type Test for Time Series Nonlinearity," *Biometrika*, 72, 39–44.

LEE, T.-H., WHITE, H., AND GRANGER, C.W. (1993): "Testing for Neglected Nonlinearity in Time Series Models: A Comparison of Neural Network Methods and Alternative Tests," *Journal of Econometrics*, 56, 269–290.

MAGEE, L. (1998): "Nonlocal Behavior in Polynomial Regressions," *The American Statistician*, 52, 20–22.

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

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

TSAY, R.S. (1986): "Nonlinearity Tests for Time Series," *Biometrika*, 73(2):461–466.

WALD, A. (1943): "Tests if Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large," *Transactions of the American Mathematical Society*, 54, 426–486.

WHITE, H. (1987): "Specification Testing in Dynamic Models," *Advances in Econometrics, Fifth World Congress*. Vol. 1. New York, NY: Cambridge University Press, pp. 1–58.

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

Nominal level (%)	$p\backslash n$	50	100	200	500	1,000	2,000	5,000
10%	1*	89.88	89.82	89.48	90.36	89.94	89.92	89.70
	2	9.02	9.04	8.78	8.10	8.12	7.72	7.90
	3	0.94	1.02	1.66	1.46	1.66	2.10	2.00
	≥ 4	0.16	0.12	0.08	0.08	0.28	0.26	0.40
5%	1*	96.12	95.92	96.44	95.5	94.4	95.24	95.18
	2	3.6	3.84	3.38	4.1	4.98	4.18	4.16
	3	0.28	0.22	0.18	0.34	0.56	0.5	0.54
	≥ 4	0	0.02	0	0.06	0.06	0.08	0.12
1%	1*	99.69	99.5	99.44	99.41	99.3	99.25	99.29
	2	0.31	0.5	0.56	0.59	0.7	0.75	0.71
	3	0	0	0	0	0	0	0
	≥ 4	0	0	0	0	0	0	0

Table 1: ESTIMATED POLYNOMIAL DEGREES BY THE SEQUENTIAL WELM TESTING PROCEDURE (IN PERCENT). Number of replications: 5,000. This table reports the portion of estimated polynomial degrees via the sequential WELM testing procedure. DGP: $y_t = \alpha_{1*}x_t + \eta_* y_{t-1} + \epsilon_t$, where $x_t = \phi_* x_{t-1} + u_t$, $(x_0, y_0) \sim$ IID $N(0, I_2)$, $(\epsilon_t, u_t) \sim$ IID $N(0, \sigma_*^2 I_2)$, $\delta_i \sim$ IID $U(0,1)$, and $(\alpha_{1*}, \eta_*, \phi_*, \sigma_*^2) = (0.5, 0.5, 0.5, 1.0)$. Here, the given hypotheses are provided as follows: $\mathcal{H}_0^{(1)}$: $\mathbb{E}[y_t|x_t, y_{t-1}] = \theta_{0*} + \theta_{1*}x_t + \theta_{y*}y_{t-1}$; $\mathcal{H}_0^{(2)}$: $\mathbb{E}[y_t|x_t, y_{t-1}] = \theta_{0*} + \theta_{1*}x_t + \theta_{2*}x_t^2 + \theta_{y*}y_{t-1}$; $\mathcal{H}_0^{(3)}$: $\mathbb{E}[y_t|y_{t-1}] = \theta_{0*} + \theta_{1*}x_t + \theta_{2*}x_t^2 + \theta_{3*}x_t^3 + \theta_{y*}y_{t-1}$; and $\mathcal{H}_0^{(4)}$: $\mathbb{E}[y_t|x_t] = \theta_{0*} + \theta_{1*}x_t + \theta_{2*}x_t^2 + \theta_{3*}x_t^3 + \theta_{4*}x_t^4 + \theta_{y*}y_{t-1}$. We further let $\Psi(x_t\delta) = \exp(x_t\delta)$ to compute the WELM test statistic.

Methods$\backslash n$	50	100	200	500	1,000	2,000	5,000
Seqn. Estmtn.	84.06	85.48	85.28	84.84	85.76	86.18	86.82
with $\boldsymbol{\alpha}_n = n^{-1/2}$	(85.86)	(90.00)	(92.93)	(95.53)	(96.83)	(97.76)	(98.59)
Seqn. Estmtn.	98.90	98.80	98.40	98.30	98.48	98.06	98.22
with $\boldsymbol{\alpha}_n = n^{-1}$	(98.00)	(99.00)	(99.50)	(99.80)	(99.90)	(99.95)	(99.98)
Seqn. Estmtn.	100.0	99.90	99.88	99.90	98.78	99.84	99.72
with $\boldsymbol{\alpha}_n = n^{-3/2}$	(99.71)	(99.90)	(99.96)	(99.99)	(100.0)	(100.0)	(100.0)
Seqn. Estmtn.	100.0	100.0	99.98	99.96	100.0	99.98	100.0
with $\boldsymbol{\alpha}_n = n^{-2}$	(99.96)	(99.99)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)
AIC	75.40	77.74	77.64	76.74	78.04	78.8	78.16
BIC	92.86	95.48	97.24	98.34	98.92	99.34	99.68
AICc	81.34	80.56	78.68	77.38	78.30	78.86	78.24
AIC$'$	70.64	73.94	73.94	75.76	75.54	76.14	76.22
BIC$'$	92.12	95.80	97.74	98.56	99.02	99.38	99.68
AICc$'$	78.84	77.50	76.30	76.34	75.94	76.34	76.32

Table 2: PORTION OF SEQUENTIALLY ESTIMATED POLYNOMIAL DEGREES BY THE SEQUENTIAL WELM TESTING PROCEDURE (IN PERCENT). Number of replications: 5,000. This table reports the percentages of the correctly estimated polynomial degree via the sequential WELM testing procedure and the information criteria. Figures in the first panel denote $\widehat{P}_n(\boldsymbol{\alpha}_n) \times 100$, and those in the second and third panels are $\tilde{P}_n \times 100$. In addition, figures in parentheses denote $(1 - \boldsymbol{\alpha}_n) \times 100$, where we let $\widehat{P}_n(\boldsymbol{\alpha}_n) := r^{-1} \sum_{i=1}^{r} \mathbb{I}(\hat{p}_{n,i} = p_*)$. r is the number of iterations, $\hat{p}_{n,i}$ denotes the degree estimator obtained by the sequential testing procedure for the i-th simulation, and $\mathbb{I}(\cdot)$ is the indicator function. Similarly, $\widetilde{P}_n := r^{-1} \sum_{i=1}^{r} \mathbb{I}(\tilde{p}_{n,i} = p_*)$, where $\tilde{p}_{n,i}$ is the degree estimator obtained by the information criteria. MODEL: $\mathcal{M}_p := \{\boldsymbol{x}_t(p)'\boldsymbol{\alpha}(p) + \eta y_{t-1} + \Psi(\delta x_t)\}$, where $p = 1, 2, 3, 4$. AIC, BIC, and AICc are the information criteria applied to $\mathcal{M}_p^0 := \{\boldsymbol{x}_t(p)'\boldsymbol{\alpha}(p) + \eta y_{t-1}\}$, and AIC$'$, BIC$'$, and AICc$'$ are those applied to $\mathcal{M}_p$, where $p = 1, 2, 3, 4$. DGP: $y_t = \alpha_{1*} x_t + \eta_* y_{t-1} + \epsilon_t$, where $x_t = \phi_* x_{t-1} + u_t$, $(x_0, y_0) \sim$ IID $N(0, I_2)$, $(\epsilon_t, u_t) \sim$ IID $N(0, \sigma_*^2 I_2)$, $\delta_i \sim$ IID $U(0,1)$, and $(\alpha_{1*}, \eta_*, \phi_*, \sigma_*^2) = (0.5, 0.5, 0.5, 1.0)$.

Nominal level (%)	$p\backslash n$	50	100	200	500	1,000	2,000	5,000
10%	1	7.56	2.18	0.78	0.16	0.06	0.00	0.00
	2*	81.30	87.14	89.10	89.40	90.48	89.98	90.20
	3	9.56	9.22	8.98	8.94	8.18	8.56	8.08
	≥ 4	1.58	1.46	1.14	1.50	1.28	1.46	1.72
5%	1	21.36	7.38	3.32	0.56	0.16	0.00	0.00
	2*	74.58	88.28	91.72	95.02	96.02	95.92	95.30
	3	3.80	4.12	4.66	4.10	3.40	3.64	4.28
	≥ 4	0.26	0.22	0.30	0.32	0.42	0.44	0.42
1%	1	61.34	28.62	12.46	2.86	0.96	0.14	0.02
	2*	38.32	70.94	86.90	96.74	98.36	98.96	99.34
	3	0.34	0.44	0.60	0.40	0.66	0.88	0.62
	≥ 4	0.00	0.00	0.04	0.00	0.02	0.02	0.02

Table 3: ESTIMATED POLYNOMIAL DEGREES BY THE SEQUENTIAL WELM TESTING PROCEDURE (IN PERCENT). Number of replications: 5,000. This table reports the portion of estimated polynomial degrees via the sequential WELM testing procedure. DGP: $y_t = \alpha_{1*}x_t + \alpha_{2*}x_t^2 + \eta_* y_{t-1} + \epsilon_t$, where $x_t = \phi_* x_{t-1} + u_t$, $(x_0, y_0) \sim$ IID $N(0, I_2)$, $(\epsilon_t, u_t) \sim$ IID $N(0, \sigma_*^2 I_2)$, $\delta_i \sim$ IID $U(0,1)$, and $(\alpha_{1*}, \alpha_{2*}, \eta_*, \phi_*, \sigma_*^2) = (0.5, 0.5, 0.5, 0.5, 1.0)$. Here, the hypotheses are provided as follows: $\mathcal{H}_0^{(1)} : \mathbb{E}[y_t|x_t, y_{t-1}] = \theta_{0*} + \theta_{1*}x_t + \theta_{y*}y_{t-1}$; $\mathcal{H}_0^{(2)} : \mathbb{E}[y_t|x_t, y_{t-1}] = \theta_{0*} + \theta_{1*}x_t + \theta_{2*}x_t^2 + \theta_{y*}y_{t-1}$; $\mathcal{H}_0^{(3)} : \mathbb{E}[y_t|y_{t-1}] = \theta_{0*} + \theta_{1*}x_t + \theta_{2*}x_t^2 + \theta_{3*}x_t^3 + \theta_{y*}y_{t-1}$; and $\mathcal{H}_0^{(4)} : \mathbb{E}[y_t|x_t] = \theta_{0*} + \theta_{1*}x_t + \theta_{2*}x_t^2 + \theta_{3*}x_t^3 + \theta_{4*}x_t^4 + \theta_{y*}y_{t-1}$. We further let $\Psi(x_t\delta) = \exp(x_t\delta)$ to compute the WELM test statistic.

Methods$\backslash n$	50	100	200	500	1,000	2,000	5,000
Seqn. Estmtn.	80.18	83.12	83.34	85.00	86.28	85.80	85.62
with $\boldsymbol{\alpha}_n = n^{-1/2}$	(85.86)	(90.00)	(92.93)	(95.53)	(96.84)	(97.76)	(98.59)
Seqn. Estmtn.	53.98	81.48	90.78	96.52	98.00	98.10	98.44
with$\boldsymbol{\alpha}_n = n^{-1}$	(94.68)	(96.84)	(98.12)	(99.05)	(99.44)	(99.67)	(99.83)
Seqn. Estmtn.	15.70	49.46	76.74	93.28	97.82	99.48	99.84
with $\boldsymbol{\alpha}_n = n^{-3/2}$	(99.71)	(99.90)	(99.96)	(99.99)	(100.0)	(100.0)	(100.0)
Seqn. Estmtn.	2.48	20.66	56.92	86.98	95.88	98.98	99.90
with $\boldsymbol{\alpha}_n = n^{-2}$	(99.96)	(99.99)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)
AIC	81.08	83.92	83.26	84.66	84.08	83.94	83.86
BIC	92.22	96.50	97.44	98.66	98.92	99.50	99.64
AICc	85.42	85.68	84.22	85.16	84.28	84.08	83.90
AIC$'$	64.90	76.28	78.16	78.80	78.74	78.64	77.78
BIC$'$	71.44	93.54	97.56	98.70	99.12	99.22	99.74
AICc$'$	69.60	80.10	79.80	79.48	79.04	78.86	77.86

Table 4: PORTION OF SEQUENTIALLY ESTIMATED POLYNOMIAL DEGREES BY THE SEQUENTIAL WELM TESTING PROCEDURE (IN PERCENT). Number of replications: 5,000. This table reports the percentages of the correctly estimated polynomial degree via the sequential WELM testing procedure and the information criteria. Figures in the first panel denote $\widehat{P}_n(\boldsymbol{\alpha}_n)\times 100$, and those in the second and third panels are $\tilde{P}_n \times 100$. In addition, figures in parentheses denote $(1-\boldsymbol{\alpha}_n)\times 100$, where we let $\widehat{P}_n(\boldsymbol{\alpha}_n) := r^{-1}\sum_{i=1}^{r}\mathbb{I}(\hat{p}_{n,i} = p_*)$; r is the number of iterations. $\hat{p}_{n,i}$ denotes the degree estimator obtained by the sequential testing procedure for the i-th simulation and $\mathbb{I}(\cdot)$ is the indicator function. Similarly, $\widetilde{P}_n := r^{-1}\sum_{i=1}^{r}\mathbb{I}(\tilde{p}_{n,i} = p_*)$, where $\tilde{p}_{n,i}$ is the degree estimator obtained by the information criteria. MODEL: $\mathcal{M}_p := \{\boldsymbol{x}_t(p)'\boldsymbol{\alpha}(p) + \eta y_{t-1} + \Psi(\delta x_t)\}$, where $p = 1, 2, 3, 4$. AIC, BIC, and AICc are the information criteria applied to $\mathcal{M}_p^0 := \{\boldsymbol{x}_t(p)'\boldsymbol{\alpha}(p) + \eta y_{t-1}\}$, and AIC$'$, BIC$'$, and AICc$'$ are those applied to $\mathcal{M}_p$, where $p = 1, 2, 3, 4$. DGP: $y_t = \alpha_{1*}x_t + \alpha_{2*}x_t^2 + \eta_* y_{t-1} + \epsilon_t$, where $x_t = \phi_* x_{t-1} + u_t$, $(x_0, y_0) \sim$ IID $N(0, I_2)$, $(\epsilon_t, u_t) \sim$ IID $N(0, \sigma_*^2 I_2)$, $\delta_i \sim$ IID $U(0,1)$, and $(\alpha_{1*}, \alpha_{2*}, \eta_*, \phi_*, \sigma_*^2) = (0.5, 0.5, 0.5, 0.5, 1.0)$.

Nominal level (%)	$p\backslash n$	50	100	200	500	1,000	2,000	5,000
10%	1	20.50	5.10	1.10	0.08	0.04	0.00	0.00
	2	56.46	51.04	30.90	6.78	1.94	0.58	0.12
	3	11.24	17.90	18.88	6.08	0.36	0.00	0.00
	≥ 4	11.8	25.96	49.12	87.06	97.66	99.42	99.88
5%	1	39.82	13.50	3.76	0.86	0.26	0.00	0.00
	2	55.16	71.94	66.90	42.70	19.30	4.82	0.30
	3	4.22	11.30	18.22	11.90	1.76	0.00	0.00
	≥ 4	0.80	3.17	11.12	44.54	78.68	95.18	99.70
1%	1	71.18	30.31	5.80	0.50	0.04	0.00	0.00
	2	27.98	65.51	78.66	48.81	22.47	9.36	1.91
	3	0.42	2.34	8.76	16.90	9.06	2.19	0.67
	≥ 4	0.42	1.84	6.78	33.79	68.43	88.45	97.42

Table 5: ESTIMATED POLYNOMIAL DEGREES BY THE SEQUENTIAL WELM TESTING PROCEDURE (IN PERCENT). Number of replications: 5,000. This table reports the portion of estimated polynomial degrees via the sequential WELM testing procedure. DGP: $y_t = \pi_* \cos(y_{t-1}) + \epsilon_t$, where $y_0 \sim N(0, \sigma_{y_0}^2)$, $\delta_i \sim$ IID $U(-1, 1)$, and $u_t \sim$ IID $N(0, \sigma_u^2)$. Here, we let $(\pi_*, \sigma_{y_0}^2, \sigma_u^2) = (1.0, 1.0, 1.0)$. The hypotheses are provided as follows: $\mathcal{H}_0^{(1)} : \mathbb{E}[y_t|x_t, y_{t-1}] = \theta_{0*} + \theta_{1*}x_t + \theta_{y*}y_{t-1}$; $\mathcal{H}_0^{(2)} : \mathbb{E}[y_t|x_t, y_{t-1}] = \theta_{0*} + \theta_{1*}x_t + \theta_{2*}x_t^2 + \theta_{y*}y_{t-1}$; $\mathcal{H}_0^{(3)} : \mathbb{E}[y_t|y_{t-1}] = \theta_{0*} + \theta_{1*}x_t + \theta_{2*}x_t^2 + \theta_{3*}x_t^3 + \theta_{y*}y_{t-1}$; and $\mathcal{H}_0^{(4)} : \mathbb{E}[y_t|x_t] = \theta_{0*} + \theta_{1*}x_t + \theta_{2*}x_t^2 + \theta_{3*}x_t^3 + \theta_{4*}x_t^4 + \theta_{y*}y_{t-1}$. All these null hypotheses are misspecified for the DGP. We further let $\Psi(x_t\delta) = \exp(x_t\delta)$ to compute the WELM test statistic.

Methods$\backslash n$	50	100	200	500	1,000	2,000	5,000
Seqn. Estmtn.	14.38	27.76	47.30	67.90	61.68	41.80	11.80
with $\boldsymbol{\alpha}_n = n^{-1/2}$	(85.86)	(90.00)	(92.93)	(95.53)	(96.84)	(97.76)	(98.59)
Seqn. Estmtn.	4.32	11.36	29.04	63.24	75.84	63.72	31.08
with$\boldsymbol{\alpha}_n = n^{-3/4}$	(94.68)	(96.84)	(98.12)	(99.05)	(99.44)	(99.67)	(99.83)
Seqn. Estmtn.	1.04	4.42	14.36	47.30	74.64	78.58	54.68
with $\boldsymbol{\alpha}_n = n^{-1}$	(98.00)	(99.00)	(99.50)	(99.80)	(99.90)	(99.95)	(99.98)
Seqn. Estmtn.	0.00	0.00	0.16	2.90	19.50	53.60	81.64
with $\boldsymbol{\alpha}_n = n^{-2}$	(99.96)	(99.99)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)
AIC	19.74	32.94	59.70	94.26	99.74	100.0	100.0
BIC	5.38	9.26	22.08	65.66	96.16	99.98	100.0
AICc	15.12	29.90	58.22	94.14	99.74	100.0	100.0
AIC$'$	8.82	9.64	13.06	26.14	45.72	76.26	98.14
BIC$'$	0.90	0.74	0.88	2.00	5.72	19.92	67.76
AICc$'$	5.34	7.84	11.84	25.50	45.40	76.12	98.12

Table 6: PORTION OF SEQUENTIALLY ESTIMATED POLYNOMIAL DEGREES BY THE SEQUENTIAL WELM TESTING PROCEDURE (IN PERCENT). Number of replications: 5,000. This table reports the percentages of the correctly estimated polynomial degree via the sequential WELM testing procedure and the information criteria. Figures in the first panel denote $\widehat{P}_n(\boldsymbol{\alpha}_n) \times 100$, and those in the second and third panels are $\widetilde{P}_n \times 100$. In addition, figure in parentheses denote $(1 - \boldsymbol{\alpha}_n) \times 100$, where we let $\widehat{P}_n(\boldsymbol{\alpha}_n) := r^{-1} \sum_{i=1}^r \mathbb{I}(\hat{p}_{n,i} = p_*)$. r is the number of iterations, $\hat{p}_{n,i}$ denotes the degree estimator obtained by the sequential testing procedure for the i-th simulation, and $\mathbb{I}(\cdot)$ is the indicator function. Similarly, $\widetilde{P}_n := r^{-1} \sum_{i=1}^r \mathbb{I}(\tilde{p}_{n,i} = p_*)$, where $\tilde{p}_{n,i}$ is the degree estimator obtained by the information criteria. MODEL: $\mathcal{M}_p := \{\boldsymbol{x}_t(p)'\boldsymbol{\alpha}(p) + \eta y_{t-1} + \Psi(\delta x_t)\}$, where $p = 1, 2, 3, 4$. AIC, BIC, and AICc are the information criteria applied to $\mathcal{M}_p^0 := \{\boldsymbol{x}_t(p)'\boldsymbol{\alpha}(p) + \eta y_{t-1}\}$, and AIC$'$, BIC$'$, and AICc$'$ are those applied to $\mathcal{M}_p$, where $p = 1, 2, 3, 4$. DGP: $y_t = \pi_* \cos(y_{t-1}) + \epsilon_t$, where $y_0 \sim N(0, \sigma_{y_0}^2)$, $\delta_i \sim$ IID $U(-1, 1)$, $u_t \sim$ IID $N(0, \sigma_u^2)$, and $(\pi_*, \sigma_{y_0}^2, \sigma_u^2) = (1.0, 1.0, 1.0)$.