Journal of Econometrics 56 (1993) 269–290. North-Holland

Testing for neglected nonlinearity in time series models

A comparison of neural network methods and alternative tests

Tae-Hwy Lee, Halbert White, and Clive W.J. Granger *

University of California at San Diego, La Jolla, CA 92093, USA

Received March 1990, final version received November 1991

In this paper a new test, the neural network test for neglected nonlinearity, is compared with the Keenan test, the Tsay test, the White dynamic information matrix test, the McLeod–Li test, the Ramsey RESET test, the Brock–Dechert–Scheinkman test, and the Bispectrum test. The neural network test is based on the approximating ability of neural network modeling techniques recently developed by cognitive scientists. This test is a Lagrange multiplier test that statistically determines whether adding 'hidden units' to the linear network would be advantageous. The performance of the tests is compared using a variety of nonlinear artificial series including bilinear, threshold autoregressive, and nonlinear moving average models, and the tests are applied to actual economic time series. The relative performance of the neural network test is encouraging. Our results suggest that it can play a valuable role in evaluating model adequacy. The neural network test has proper size and good power, and many of the economic series tested exhibit potential nonlinearities.

1. Introduction

Specification and estimation of linear time series models are well-established procedures, based on ARIMA univariate models or VAR or VARMAX multivariate models. However, economic theory frequently suggests nonlinear relationships between variables, and many economists appear to believe that the economic system is nonlinear. It is thus interesting to test whether or not a single

Correspondence to: Halbert White, Department of Economics 0508, University of California at San Diego, La Jolla, CA 92093-0508, USA.

* We would like to acknowledge the helpfulness of Tung Liu who allowed us use of his results on the BDS test, and thank the referees and editors for numerous helpful suggestions. We retain responsibility for any remaining errors. White's research was supported by NSF grant SES-8921382.

economic series or group of series appears to be generated by a linear model against the alternative that they are nonlinearly related. There are many tests presently available to do this. This paper considers a 'neural network' test recently proposed by White (1989b), and compares its performance with several alternative tests using a Monte Carlo study.

It is important to be precise about the meaning of the word 'linearity'. Throughout, we focus on a property best described as 'linearity in conditional mean'. Let $\{Z_t\}$ be a stochastic process, and partition Z_t as $Z_t = (y_t, X_t')'$, where (for simplicity) y_t is a scalar and X_t is a $k \times 1$ vector. X_t may (but need not necessarily) contain a constant and lagged values of y_t. The process $\{y_t\}$ is *linear in mean conditional on* X_t if

$$\mathrm{P}[\mathrm{E}(y_t | X_t) = X_t'\theta^*] = 1 \quad \text{for some} \quad \theta^* \in \mathbb{R}^k.$$

Thus, a process exhibiting autoregressive conditional heteroskedasticity (ARCH) [Engle (1982)] may nevertheless exhibit linearity of this sort because ARCH does not refer to the conditional mean. Our focus is appropriate whenever we are concerned with the adequacy of linear models for forecasting.

The alternative of interest is that y_t is not linear in mean conditional on X_t, so that

$$\mathrm{P}[\mathrm{E}(y_t | X_t) = X_t'\theta] < 1 \quad \text{for all} \quad \theta \in \mathbb{R}^k.$$

When the alternative is true, a linear model is said to suffer from 'neglected nonlinearity'.

Most of the tests treated here have as a first step the extraction of linear structure by the use of an estimated filter. Typically, an AR(p) model is fitted to the series and the test then applied to the estimated residuals. To automate this procedure a particular value of p is used in the simulations, usually $p = 1$ or 2, but when dealing with actual data we shall choose p using the SIC criterion leading to consistent choice of p [Hannan (1980)].

Several tests involve regressing linear model residuals on specific functions of X_t, chosen to capture essential features of possible nonlinearities; the null hypothesis is rejected if these functions of X_t are significantly correlated with the residual. When the null is rejected, the implied alternative model may provide forecasts superior to those from the linear model. These forecasts need not be optimal, merely better. Tests not based on models that imply such forecasts are the McLeod–Li test (based on the autocorrelation of the squared residuals), the Brock–Dechert–Scheinkman test (BDS) (arising from consideration of chaotic processes), and the Bispectrum test. As not all tests are based on alternative forecasting models, we have not considered the relative forecasting abilities of the linear and implied alternative models, although this should be informative and may be considered in further work.

2. The neural network test

Cognitive scientists have recently introduced a class of 'neural network' models inspired by certain features of the way in which information is processed in the brain. An accessible treatment is given by Rumelhart, Hinton, and Williams (1986). A leading model is the 'single hidden layer feedforward network', depicted in fig. 1. In this network, input units ('sensors') send signals x_i, $i = 1, \dots, k$, along links ('connections') that attenuate or amplify the original signals by a factor γ_{ji} ('weights' or 'connection strengths'). The intermediate or 'hidden' processing unit j 'sees' signals $x_i\gamma_{ji}$, $i = 1, \dots, k$, and processes these in some characteristic, typically simple way. Commonly, the hidden units sum the arriving signals [yielding $\tilde{x}'\gamma_j$, where $\tilde{x} = (1, x_1, \dots, x_k)'$, $\gamma_j \equiv (\gamma_{j0}, \gamma_{j1}, \dots, \gamma_{jk})'$] and then produce an output 'activation' $\psi(\tilde{x}'\gamma_j)$, where the 'activation function' ψ is a given nonlinear mapping from $\mathbb{R}$ to $\mathbb{R}$. Often, ψ is a cumulative distribution function (c.d.f.), such as the logistic, $\psi(\lambda) = (1 + e^{-\lambda})^{-1}$, $\lambda \in \mathbb{R}$. Hidden unit signals $\psi(\tilde{x}'\gamma_j)$, $j = 1, \dots, q$, then pass to the output, which sums what it sees to produce an output

$$f(x, \delta) = \beta_0 + \sum_{j=1}^{q} \beta_j \psi(\tilde{x}'\gamma_j), \qquad q \in \mathbb{N}, \tag{2.1}$$

where $\beta_0, \dots, \beta_q$ are hidden to output weights and $\delta = (\beta_0, \dots, \beta_q, \gamma'_1, \dots, \gamma'_q)'$. For convenience and without loss of generality, we suppose that the output unit performs no further transformations.

As discussed by White (1989a, 1990), functions defined by (2.1) belong to a family of flexible functional forms indexed by ψ and q. Hornik, Stinchcombe,

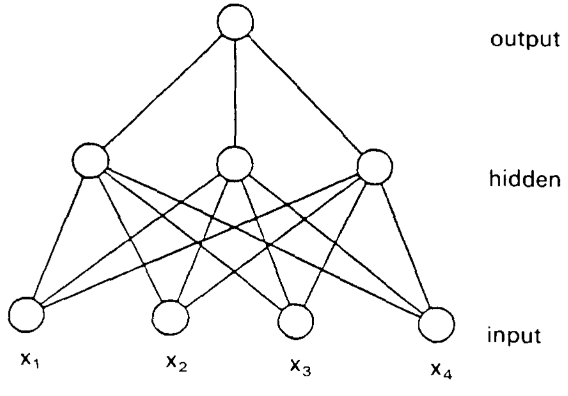

Fig. 1. Single hidden layer feedforward network

and White (1989, 1990) and Stinchcombe and White (1989) – among others – have shown that for wide classes of nonlinear functions ψ, functions of the form (2.1) can provide arbitrarily accurate approximations to arbitrary functions in a variety of normed function spaces (e.g., continuous functions on a compact set with the uniform norm, functions in L_p space, and functions in Sobolev spaces with Sobolev norm), provided that q is sufficiently large. Thus, functions of the form (2.1) are capable of approximating an arbitrary nonlinear mapping. Considerable practical experience shows that when the mapping is fairly smooth, tractable values for q can provide quite good approximations. For example, Lapedes and Farber (1987) well approximated the deterministic chaos of the logistic map using five hidden units, while Gallant and White (1992) well approximated the Mackey–Glass chaos with five hidden units.

Similar approximation-theoretic issues arise in the context of projection pursuit [Friedman and Stuetzle (1981), Diaconis and Shahshahani (1984), Huber (1985), Jones (1991)]. In fact, (2.1) can be viewed as a restricted projection pursuit function in which the functions ψ are given a priori. In standard projection pursuit, ψ may differ for each term (replace ψ with ψ_j) and one must estimate the ψ_j.

The neural network test for neglected nonlinearity uses a single hidden layer network augmented by connections from input to output. Network output o is then

$$o = \tilde{x}'\theta + \sum_{j=1}^{q} \beta_j \psi(\tilde{x}'\gamma_j).$$

When the null hypothesis of linearity is true, i.e.,

$$\mathrm{H}_0: \quad \mathrm{P}[\mathrm{E}(y_t | X_t) = \tilde{X}_t'\theta^*] = 1 \quad \text{for some } \theta^*,$$

then optimal network weights β_j, say β_j^*, are zero, $j = 1, \dots, q$, yielding an 'affine network'. The neural network test for neglected nonlinearity tests the hypothesis $\beta_j^* = 0$, $j = 1, \dots, q$, for particular choice of q and γ_j. The test will have power whenever $\sum_{j=1}^{q} \beta_j \psi(\tilde{x}'\gamma_j)$ is capable of extracting structure from $e_t^* = y_t - \tilde{X}_t'\theta^*$. [Under the alternative, θ^* is the parameter vector of the optimal linear least squares approximation to $\mathrm{E}(y_t | X_t)$.] Recent theoretical work of Stinchcombe and White (1991) suggests that when ψ is the logistic c.d.f., the terms $\psi(\tilde{x}'\gamma_j)$ are generically (in γ_j) able to extract such structure.

Implementing the test as a Lagrange multiplier test leads to testing

$$\mathrm{H}_0': \quad \mathrm{E}(\Psi_t e_t^*) = 0 \quad \text{vs} \quad \mathrm{H}_a': \quad \mathrm{E}(\Psi_t e_t^*) \neq 0,$$

where $\Psi_t \equiv (\psi(\tilde{X}_t'\Gamma_1), \dots, \psi(\tilde{X}_t'\Gamma_q))'$, and $\Gamma = (\Gamma_1, \dots, \Gamma_q)$ is chosen a priori, independently of the random sequence $\{X_t\}$, for given $q \in \mathbb{N}$. We call Ψ_t the 'phantom hidden unit activations'. As in Bierens (1987) and Bierens and Hartog (1988), we shall choose Γ at random. An analysis for Γ chosen to maximize departures of $\mathrm{E}(\Psi_t e_t^*)$ from zero (with $q = 1$, say) can be carried out along the lines of Bierens (1990), but is beyond the scope of the present work.

In constructing the test, we replace e_t^* with estimated residuals $\hat{e}_t = y_t - \tilde{X}_t'\hat{\theta}$, with $\hat{\theta}$ obtained by least squares. This leads to a statistic of the form

$$M_n = \left(n^{-1/2} \sum_{t=1}^{n} \Psi_t \hat{e}_t\right)' \hat{W}_n^{-1} \left(n^{-1/2} \sum_{t=1}^{n} \Psi_t \hat{e}_t\right),$$

where $\hat{W}_n$ is a consistent estimator of $W^* = \mathrm{var}(n^{-1/2} \sum_{t=1}^{n} \Psi_t e_t^*)$. Standard asymptotic arguments lead to the conclusion $M_n \xrightarrow{\mathrm{d}} \chi^2(q)$ as $n \to \infty$ under H_0.

Two practical difficulties may be noted: 1) Elements of Ψ_t tend to be collinear with X_t and with themselves. 2) Computation of $\hat{W}_n$ can be tedious. These can be remedied by 1) conducting the test using $q^* < q$ principal components of Ψ_t not collinear with X_t, denoted Ψ_t^*, and 2) using an equivalent test statistic that avoids explicit computation of $\hat{W}_n$,

$$nR^2 \xrightarrow{\mathrm{d}} \chi^2(q^*),$$

where R^2 is the uncentered squared multiple correlation from a standard linear regression of $\hat{e}_t$ on Ψ_t^*, $\tilde{X}_t$.

3. Alternative tests

In every case an AR(p) model is first fitted to the data and nonlinearity tested for the residuals. In fact any linear model could first be used.

3.1. The Keenan, Tsay, and Ramsey RESET tests

Let y_t be series of interest and let $X_t = (y_{t-1}, \ldots, y_{t-p})'$ be used to explain y_t. (An obvious generalization allows X_t to include other explanatory variables.) In performing the tests, p and any other contents of X_t have to be determined by the user.

The first step of these tests is linear regression of y_t on $\tilde{X}_t$, producing an estimate $\hat{\theta}$, a forecast $f_t = \tilde{X}_t'\hat{\theta}$, and estimated residuals $\hat{e}_t = y_t - \tilde{X}_t'\hat{\theta}$.

Keenan (1985) introduced a test based on the correlation of $\hat{e}_t$ with f_t^2. The Keenan test essentially asks if the squared forecast has any additional forecasting ability, and so tests directly for departures from linearity in mean.

The test of Tsay (1986) has a similar form to the Keenan test, but tests the possibility of improving forecasts by including $p(p+1)/2$ cross-product terms of the components of X_t, of the form $y_{t-j}y_{t-k}$, $k \geq j$, $j, k = 1, \ldots, p$. The test is again directly designed to test for departures from linearity in mean.

The RESET test proposed by Ramsey (1969) generalizes the Keenan test in a different way. Using the polynomials in f_t we can postulate an alternative model of the form

$$y_t = \tilde{X}_t'\theta + a_2 f_t^2 + \cdots + a_k f_t^k + v_t \quad \text{for some} \quad k \geq 2.$$

The null hypothesis is $\mathrm{H_0}$: $a_2 = \cdots = a_k = 0$. Denoting $\hat{e} = (\hat{e}_1, \ldots, \hat{e}_n)'$ and $\hat{v} = (\hat{v}_1, \ldots, \hat{v}_n)'$, the test statistic is $\mathrm{RESET} = [(\hat{e}'\hat{e} - \hat{v}'\hat{v})/(k-1)]/[\hat{v}'\hat{v}/(n-k)]$, which is approximately $F(k-1, n-k)$ when $\mathrm{H_0}$ is true.

As for the neural network tests, collinearity can be avoided by forming the principal components of $(f_t^2, \ldots, f_t^k)$, choosing the $p^* < (k-1)$ largest (except the first principal component so as not to be collinear with $\tilde{X}_t$), and then regressing y_t on these and $\tilde{X}_t$, giving the residual $\hat{u}_t$. The test statistic is $RESET1 = [(\hat{e}'\hat{e} - \hat{u}'\hat{u})/p^*]/[\hat{u}'\hat{u}/(n-k)]$, which is $F(p^*, n-k)$ when $\mathrm{H_0}$ is true.

A Lagrange multiplier version of the test is obtained by regressing $\hat{e}_t$ on $\tilde{X}_t$ and $f_t^2, \ldots, f_t^k$ to get an R^2 statistic. Under regularity conditions nR^2 is asymptotically distributed as $\chi^2(k-1)$ under the null. Again, forming the principal components of $(f_t^2, \ldots, f_t^k)$, choosing the p^* largest, and then regressing $\hat{e}_t$ on these and $\tilde{X}_t$, also gives an R^2 statistic. Under regularity conditions, the statistic $RESET2 = nR^2$ is distributed as $\chi^2(p^*)$ for n large, under $\mathrm{H_0}$. For this test both k and p^* have to be selected by the user. The RESET tests are sensitive primarily to departures from linearity in mean.

3.2. *The White dynamic information matrix test*

White (1987, 1992) proposed a specification test for dynamic (time series) models, based on the covariance of conditional score functions. For the normal linear model

$$y_t = \tilde{X}_t'\theta + e_t, \quad e_t \sim \mathrm{N}(0, \sigma^2),$$

the log-likelihood is

$$\log f_t(x_t, \theta, \sigma) = \text{constant} - \log\sigma - (y_t - \tilde{X}_t'\theta)^2/2\sigma^2.$$

With $u_t = (y_t - \tilde{X}_t'\theta)/\sigma$, the conditional score function is

$$s_t(X_t, \theta, \sigma) \equiv \nabla \log f_t(X_t, \theta, \sigma) = \sigma^{-1}(u_t, u_t X_t', u_t^2 - 1)',$$

where ∇ is the gradient operator with respect to θ and σ. Denoting $s_t^* = s_t(X_t, \theta^*, \sigma^*)$, correct specification implies $\mathrm{E}(s_t^*) = 0$ and $\mathrm{E}(s_t^* s_{t-\tau}^{*\prime}) = 0$, $t = 1, 2, \ldots, \tau = 1, \ldots, t$. The dynamic information matrix test can be based on the indicator $m_t = S \text{ vec } s_t s_{t-1}'$, where S is a nonstochastic selection matrix focusing attention on particular forms of possible misspecification.

Denoting $\hat{s}_t = s_t(X_t, \hat{\theta}, \hat{\sigma})$ and $\hat{m}_t = S \text{ vec } \hat{s}_t \hat{s}'_{t-1}$, where $\hat{\theta}$ and $\hat{\sigma}$ are the quasi-maximum likelihood estimators (QMLEs), the following versions of the test statistic can be formed: 1) $WHITE1 = n\hat{\mu}'_n \hat{J}_n^{-1} \hat{\mu}_n$ where $\hat{\mu}_n = n^{-1}\sum_{t=1}^n \hat{m}_t$, $\hat{J}_n = n^{-1}\sum \hat{m}_t \hat{m}'_t - (n^{-1}\sum \hat{m}_t \hat{s}'_t)[n^{-1}\sum \hat{s}_t \hat{s}'_t]^{-1}(n^{-1}\sum \hat{s}_t \hat{m}'_t)$; 2) $WHITE2 = nR^2$, where R^2 is the (constant unadjusted) squared multiple correlation coefficient of the regression of the constant unity on the explanatory variables $\hat{m}_t$, $\hat{s}_t$; 3) $WHITE3 = nR^2$ where R^2 is the (constant unadjusted) squared multiple correlation coefficient from the regression of $\hat{u}_t = (y_t - \tilde{X}'_t \hat{\theta})/\hat{\sigma}$ on $\tilde{X}_t$ and $\hat{k}_t$, with $\hat{k}_t$ being defined from $\hat{m}_t = \hat{k}_t \hat{u}'_t$. Under H_0, *WHITE1*, *WHITE2*, and *WHITE3* all have the $\chi^2(q)$ distribution asymptotically, where q is the dimension of m_t. These tests will be sensitive to departures from linearity in mean to the extent that these departures induce autocorrelation in s_t^*. Other misspecifications resulting in such autocorrelations will also be detected.

3.3. The McLeod and Li test

It was noted in Granger and Andersen (1978) that for a linear stationary process

$$\text{corr}(y_t^2, y_{t-k}^2) = [\text{corr}(y_t, y_{t-k})]^2 \quad \text{for all} \quad k,$$

and so departures from this would indicate nonlinearity. McLeod and Li (1983) use the squared residuals from a linear model and apply a standard Box–Ljung Portmanteau test for serial correlation. This test is sensitive to departures from linearity in mean that result in apparent ARCH structure; ARCH itself will also be detected.

3.4. The BDS test

Whilst conducting research on tests for chaos, Brock, Dechert, and Scheinkman (1986) derived a test appropriate for detecting general stochastic nonlinearity. For a series y_t, define

$$C_m(\varepsilon) = n^{-2}[\text{number of pairs } (i,j) \text{ such that}$$
$$|y_i - y_j| < \varepsilon, |y_{i+1} - y_{j+1}| < \varepsilon, \ldots, |y_{i+m-1} - y_{j+m-1}| < \varepsilon],$$

so that $y_i, \ldots, y_{i+m-1}$ and $y_j, \ldots, y_{j+m-1}$ are two segments of the series of length m, such that all corresponding pairs of points differ from each other by size ε. The test statistic is $BDS = n^{1/2}[C_m(\varepsilon) - C_1(\varepsilon)^m]$. Under the null hypothesis that the series is independently and identically distributed, *BDS* is asymptotically normally distributed with zero mean and a known complicated variance. The test is interesting because it arises from very different considerations than the others. For our implementation, y_t is replaced by the linear model estimated residuals. The BDS test will be sensitive to departures from linearity in mean, but may also have power against series linear in mean with ARCH.

3.5. The Bispectrum test

Following earlier work by Subba Rao and Gabr (1980, 1984), a test based on the bispectrum was suggested by Hinich (1982) and Ashley, Patterson, and Hinich (1986). If $\{y_t\}$ is a zero mean stationary series, it can be expressed as $y_t = \sum_{i=0}^{\infty} a_i \varepsilon_{t-i}$, where $\{\varepsilon_t\}$ is purely random and the weights $\{a_i\}$ are fixed. Define the second- and third-order autocovariances by $c(h) = \mathrm{E}[y_{t+h} y_t]$ and $c(h, k) = \mathrm{E}[y_{t+h} y_{t+k} y_t]$, and write their corresponding Fourier transforms (the power spectrum and power bispectrum) as $S(\omega) = \sum_h c(h)\exp(-2\pi i \omega h)$ and $B(\omega_1, \omega_2) = \sum_{h,k} c(h, k)\exp[-2\pi i(\omega_1 h + \omega_2 k)]$. It can be shown that

$$\frac{|B(\omega_1, \omega_2)|^2}{S(\omega_1)S(\omega_2)S(\omega_1 + \omega_2)} = \frac{\mu_3^2}{\sigma_\varepsilon^6} \quad \text{for all} \quad (\omega_1, \omega_2),$$

where $\sigma_\varepsilon^2 = \mathrm{E}\varepsilon_t^2$ and $\mu_3 = \mathrm{E}\varepsilon_t^3$. The square root of this is called the skewness of $\{y_t\}$. The fact that the skewness of such a time series is independent of (ω_1, ω_2) is used to test nonlinearity. The test statistic of Hinich (1982) is based on the interquartile range of the estimated ratio of the skewness over a set of frequency pairs, (ω_1, ω_2). This proved to be too expensive to use in the Monte Carlo simulation but was used in section 6. We thank the authors of this test for providing the software to perform the test. In our implementation y_t is again replaced by the linear model estimated residual. The Bispectrum test is sensitive to departures from linearity in mean, but will also detect ARCH.

4. The simulation design

Two blocks of univariate series were generated from models chosen to represent a variety of stable nonlinear situations. Throughout $\varepsilon_t \sim \mathrm{N}(0, 1)$ is a white noise series.

4.1. Block1

(i) Autoregressive (AR)

$$y_t = 0.6 y_{t-1} + \varepsilon_t,$$

a representative linear model.

(ii) Bilinear (BL)

$$y_t = 0.7 y_{t-1} \varepsilon_{t-2} + \varepsilon_t,$$

a bilinear model having the same covariance properties as a white noise; see Granger and Andersen (1978).

(iii) Threshold autoregressive (TAR)

$$\begin{aligned} y_t &= \quad 0.9 y_{t-1} + \varepsilon_t \quad \text{for} \quad |y_{t-1}| \le 1, \\ &= -0.3 y_{t-1} + \varepsilon_t \quad \text{for} \quad |y_{t-1}| > 1, \end{aligned}$$

an example considered by Tong (1983).

(iv) Sign autoregressive (SGN)

$$y_t = \operatorname{sgn}(y_{t-1}) + \varepsilon_t,$$

where $\operatorname{sgn}(x) = 1$ if $x > 0$, $= 0$ if $x = 0$, $= -1$ if $x < 0$. This is a particular form of nonlinear autoregression (NLAR).

(v) Nonlinear autoregressive (NAR)

$$y_t = (0.7|y_{t-1}|)/(|y_{t-1}| + 2) + \varepsilon_t,$$

another NLAR model, closely related to a class of models known as rational NLAR.

The models of the second block have been used in previous papers on nonlinear testing, particularly by Keenan (1985), Tsay (1986), Ashley, Patterson and Hinich (1986), and Chan and Tong (1986), and so are included to allow comparison with these studies.

4.2. Block2

(1) MA(2) (Model1)

$$y_t = \varepsilon_t - 0.4\varepsilon_{t-1} + 0.3\varepsilon_{t-2}.$$

(2) Heteroskedastic MA(2) (Model2)

$$y_t = \varepsilon_t - 0.4\varepsilon_{t-1} + 0.3\varepsilon_{t-2} + 0.5\varepsilon_t\varepsilon_{t-2},$$

where the forecastable part of y_t is linear and the final product term introduces heteroskedasticity.

(3) Nonlinear MA (Model3)

$$y_t = \varepsilon_t - 0.3\varepsilon_{t-1} + 0.2\varepsilon_{t-2} + 0.4\varepsilon_{t-1}\varepsilon_{t-2} - 0.25\varepsilon_{t-2}^2,$$

where the final two terms give a nonlinear MA model that is typically noninvertible.

(4) AR(2) (Model4)

$$y_t = 0.4y_{t-1} - 0.3y_{t-2} + \varepsilon_t,$$

a stationary linear AR model.

(5) Bilinear AR (Model5)

$$y_t = 0.4y_{t-1} - 0.3y_{t-2} + 0.5y_{t-1}\varepsilon_{t-1} + \varepsilon_t,$$

a model containing both linear and bilinear terms.

(6) Bilinear ARMA (Model6)

$$y_t = 0.4y_{t-1} - 0.3y_{t-2} + 0.5y_{t-1}\varepsilon_{t-1} + 0.8\varepsilon_{t-1} + \varepsilon_t.$$

It is seen that three linear models are included for purposes of comparison. Data were generated with sample sizes 50, 100, and 200. 1000 replications were used to obtain estimates of power. In a few sample cases the plot of y_t against y_{t-1} was constructed. It was generally true that there was no obvious nonlinearity visible, except possibly for the SGN autoregressive series.

All of the tests can be generalized in fairly obvious ways to consider nonlinear relationships between pairs of series. As part of an initial exploration of this situation, two nonlinearly related pairs of series were also generated.

4.3. Bivariate models

SQ

$$y_t = x_t^2 + \varepsilon_t,$$

EXP

$$y_t = \exp(x_t) + \varepsilon_t,$$

where $x_t = 0.6x_{t-1} + e_t$, ε_t and e_t are independent, $\varepsilon_t \sim \mathrm{N}(0, v)$ and $e_t \sim \mathrm{N}(0, 1)$ white noises. Three different values for the variance of ε_t were used, $v = 1, 25, 400$. Plots of y_t against x_t with $v = 1$ clearly showed a nonlinear relationship, but this was no longer clear when $v = 400$.

For all the simulations except for TSAY2, the information set is $X_t = y_{t-1}$ for Block1, $X_t = (y_{t-1}, y_{t-2})'$ for Block2, and $X_t = x_t$ for the bivariate models. For the TSAY2 test, $X_t = (y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5})'$ for Block1 and Block2 and $X_t = (x_t, x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4})'$ for bivariate models.

In performing neural network tests the logistic c.d.f. $\psi(\lambda) = [1 + \exp(-\lambda)]^{-1}$ is used. The input to hidden unit weights Γ_{ji} were randomly generated from the uniform distribution on $[-2, 2]$. The variables y_t, X_t have been rescaled onto $[0, 1]$. We choose $q = 10$ for NEURAL1 and $q = 20$ for NEURAL2. We use $q^* = 2$ largest principal components (excluding the first principal component) for the Block1 and bivariate models and $q^* = 3$ for the Block2 models.

For the White dynamic information matrix tests, appropriate choice of S gives

$$m_t' = \sigma^{-2}(u_t u_{t-1}, X_t u_t u_{t-1}, X_{t-1} u_t u_{t-1}, X_t X_{t-1} u_t u_{t-1}),$$

where $X_t = y_{t-1}$ for Block1, $X_t = (y_{t-1}, y_{t-2})'$ for Block2, and $X_t = x_t$ for the bivariate models, so that $q = 4$ for the Block1 and bivariate models and $q = 8$ for the Block2 models.

We choose 20 degrees of freedom for the McLeod and Li tests, and $k = 5$, $p^* = 1$ in RESET1 and RESET2 for every model.

5. Results of the simulation

Critical values for the various tests were constructed using both the asymptotic theory and by simulation using a linear model. For Block1 the AR(1) (Model (i)) was used and for Block2 the AR(2) (Model4) was used, for sample sizes 50, 100, and 200, and with 6000 replications. The bivariate models used the AR(1) values. 10%, 5%, 2.5%, and 1% critical values were constructed for each test, but to conserve space only 5% values are reported.

Tables 1 and 2 show the 5% critical values simulated for sample sizes 50, 100, and 200, and the asymptotic values.

Table 3 illustrates the stability of the 5% critical values using 6000 simulations. 5% critical values were obtained from each of the AR(1) models $x_t = \phi x_{t-1} + \varepsilon_t$, with ϕ taking the values -0.9, -0.6, -0.3, 0.0, 0.3, 0.6, and 0.9, for sample size $n = 100$. Again, with this sample size, 1000 replications of the AR(1) model with $\phi = 0.6$ were constructed, and the tests applied with the critical values obtained from the previous simulation. The 'power' or percent

Table 1

Critical values (5%) for Block1 models.[a]

Test	$n = 50$	$n = 100$	$n = 200$
NEURAL1	5.52 (5.99)	5.40 (5.99)	5.58 (5.99)
NEURAL2	5.52 (5.99)	5.45 (5.99)	5.65 (5.99)
KEENAN	3.09 (4.05)	3.28 (3.94)	3.67 (3.84)
TSAY1	3.16 (4.05)	3.32 (3.94)	3.69 (3.84)
TSAY2	2.50 (2.03)	2.02 (1.80)	1.88 (1.67)
WHITE1	9.95 (9.49)	9.15 (9.49)	8.92 (9.49)
WHITE2	11.84 (9.49)	10.82 (9.49)	10.02 (9.49)
WHITE3	9.01 (9.49)	9.24 (9.49)	9.34 (9.49)
MCLEOD	31.14 (31.41)	31.42 (31.41)	31.67 (31.41)
RESET1	3.01 (4.06)	3.19 (3.94)	3.59 (3.84)
RESET2	3.05 (3.84)	3.11 (3.84)	3.39 (3.84)

[a] The first number in each cell is the simulated critical value from AR(1), $y_t = 0.6y_{t-1} + \varepsilon_t$, with 6000 replications, and the second number in parentheses is the asymptotic critical value.

rejection in each case is shown in table 3. Thus the figure 4.2 at the top left means that data from an AR(1) with $\phi = 0.6$ led to rejection of the linearity null hypothesis on 4.2% of occasions by the NEURAL1 test using the critical value obtained from the AR(1) model with $\phi = -0.9$. With 1000 replications, the 95% confidence intervals of these powers (sizes) around the hoped for value of 5%, is 3.6 to 6.4%. The final column shows the rejection percentage of H_0 using the asymptotic 5% critical values. Virtually all of the neural network test results lie in the 95% confidence interval for size, as do all of the values in the column when $\phi = 0.6$. However, some tests do not perform satisfactorily, particularly the KEENAN, TSAY1, TSAY2, and WHITE1 tests, suggesting either that the critical values are not stable across simulation models or that the asymptotic theory is not appropriate with a sample size of 100. Because of its complexity, the BDS test was not calculated in this particular exercise.

Table 2
Critical values (5%) for Block 2 models.[a]

Test	$n = 50$	$n = 100$	$n = 200$
NEURAL1	7.72 (7.81)	7.47 (7.81)	7.98 (7.81)
NEURAL2	7.74 (7.81)	7.52 (7.81)	7.85 (7.81)
KEENAN	3.70 (4.06)	3.84 (3.94)	4.00 (3.84)
TSAY1	2.84 (2.82)	2.63 (2.70)	2.69 (2.60)
TSAY2	2.23 (2.03)	1.86 (1.80)	1.74 (1.67)
WHITE1	23.45 (15.51)	18.58 (15.51)	16.90 (15.51)
WHITE2	19.36 (15.51)	17.42 (15.51)	16.74 (15.51)
WHITE3	14.98 (15.51)	14.86 (15.51)	15.49 (15.51)
MCLEOD	31.41 (31.41)	31.29 (31.41)	31.19 (31.41)
RESET1	3.49 (4.06)	3.77 (3.94)	3.92 (3.84)
RESET2	3.57 (3.84)	3.71 (3.84)	3.82 (3.84)

[a] The first number in each cell is the simulated critical value from Model4, $y_t = 0.4y_{t-1} - 0.3y_{t-2} + \varepsilon_t$, with 6000 replications, and the second number in parentheses is the asymptotic critical value.

Table 3
Size of tests and similarity.[a]

Test	-0.9	-0.6	-0.3	0.0	0.3	0.6	0.9	Asymp.
NEURAL1	4.2	4.2	4.0	4.9	4.8	5.6	4.0	4.2
NEURAL2	4.2	4.1	4.1	4.8	4.7	5.6	3.5	4.0
KEENAN	3.0	3.5	3.5	3.5	4.0	5.1	8.9	3.1
TSAY1	3.0	3.5	3.5	3.5	4.0	5.1	8.9	3.2
TSAY2	8.0	7.4	7.6	9.6	7.8	6.1	5.1	10.1
WHITE1	2.5	3.0	5.3	4.5	5.4	4.1	2.4	3.6
WHITE2	4.5	4.1	4.9	4.5	4.2	4.1	3.6	7.4
WHITE3	4.0	4.0	4.7	4.4	5.1	4.1	3.9	3.8
MCLEOD	4.4	4.4	4.4	4.4	4.4	4.4	4.7	4.4
RESET1	3.7	3.7	3.7	3.7	4.1	5.0	4.9	3.5
RESET2	3.9	3.7	3.6	3.6	3.7	5.2	5.9	3.6

[a] (1) Each column shows the power (%) for AR(1) $y_t = 0.6y_{t-1} + \varepsilon_t$, using the 5% critical values simulated with $y_t = \phi y_{t-1} + \varepsilon_t$, $\phi = -0.9, -0.6, -0.3, 0.0, 0.3, 0.6, 0.9$. The last column shows the power for the AR(1) using 5% asymptotic critical values. (2) 95% confidence interval of the observed size is (3.6, 6.4), since if the true size is s the observed size follows the (asymptotic) normal distribution with mean s and variance $s(1 - s)/1000$. (3) Sample size = 100, replications = 1000.

Table 4 shows the power of the tests using the Block1 models plus the bivariate models with $v = 1$ and with sample size $n = 200$. The first number is the power using the 5% critical value from the simulation of the AR(1) model with $\phi = 0.6$ and below it, in parentheses, is the power using the 5% theoretical critical value. A great deal of variation in the power of the tests is observed. Most tests have reasonable power for the bilinear model, but the White dynamic information matrix test and McLeod–Li test are particularly successful. For the threshold autoregressive data, only the neural network test has any success and similarly for the SGN autoregressive model, where the neural network test is very powerful, the RESET test has some power, and the other tests have very

Table 4
Power for Block1 and bivariate models.[a]

Test	AR	BL	TAR	SGN	NAR	SQ	EXP
NEURAL1	5.2	58.0	79.8	98.9	20.4	100.0	100.0
	(4.5)	(55.6)	(77.2)	(98.6)	(18.3)	(100.0)	(100.0)
NEURAL2	5.4	57.1	80.7	99.3	19.9	100.0	100.0
	(4.4)	(55.5)	(78.3)	(98.7)	(17.9)	(100.0)	(100.0)
KEENAN	4.7	39.3	4.2	15.0	23.8	100.0	100.0
	(4.2)	(38.1)	(3.9)	(13.8)	(21.8)	(100.0)	(100.0)
TSAY1	4.7	39.3	4.2	15.0	23.8	100.0	100.0
	(4.3)	(38.2)	(3.9)	(13.9)	(21.8)	(100.0)	(100.0)
TSAY2	4.8	47.7	2.7	5.2	2.8	66.9	36.2
	(8.4)	(55.5)	(5.2)	(8.4)	(5.4)	(72.8)	(42.8)
WHITE1	5.2	97.5	4.1	7.7	4.5	75.5	71.1
	(3.2)	(96.8)	(3.0)	(5.9)	(3.2)	(71.5)	(66.4)
WHITE2	5.8	95.9	5.9	12.8	5.7	67.0	62.7
	(7.0)	(96.8)	(7.4)	(14.3)	(7.4)	(70.5)	(66.2)
WHITE3	5.5	99.5	4.2	12.1	5.5	79.5	60.0
	(5.2)	(99.4)	(3.9)	(11.7)	(5.1)	(78.9)	(59.6)
MCLEOD	4.6	90.2	5.0	3.8	4.8	23.0	15.0
	(4.9)	(90.3)	(5.2)	(3.9)	(5.0)	(23.2)	(15.0)
RESET1	5.0	38.5	6.1	32.3	23.5	96.0	61.2
	(3.9)	(36.0)	(5.1)	(30.3)	(21.3)	(95.9)	(58.7)
RESET2	5.2	40.2	7.0	32.7	26.2	95.4	59.5
	(3.4)	(36.2)	(5.3)	(29.5)	(22.1)	(95.2)	(56.0)
BDS	(6.1)	(98.8)	(14.5)	(12.6)	(5.9)	(70.6)	(49.6)

[a] Power using 5% critical value simulated with AR(1) model is shown (except for BDS test), and power using 5% asymptotic critical value is shown in parentheses. Sample size = 200, replications = 1000. The results for BDS test in Liu (1990) with embedding dimension $m = 2$ and $\varepsilon = (0.8)^6$ are reported here. The series were divided by the range of the data before applying the BDS tests.

little power. The particular nonlinear AR model used (NAR) is difficult to detect, as the power is low for all the tests. The bivariate cases are discussed later.

Table 5 shows similar power results (5% critical value, $n = 200$) for the Block2 series. To provide the simulation critical values, 6000 simulations of Model4, an AR(2) model, were generated. 1000 replications for each of the six models were generated, and the simulated and asymptotic 5% critical values were used to obtain estimated powers. Model1 is linear, being an MA(2), and most tests reject the null near the correct 5% rate, but the WHITE tests reject frequently, indicating dynamic misspecification, as an MA(2) is only poorly approximated by an AR(2) model. All of the tests have good power against the bilinear models

Table 5
Power for Block2 models.[a]

Test	Model1	Model2	Model3	Model4	Model5	Model6
NEURAL1	5.0	17.2	98.1	5.8	94.2	85.8
	(5.3)	(18.1)	(98.4)	(6.2)	(94.5)	(86.4)
NEURAL2	4.8	16.8	99.2	5.5	91.7	82.7
	(4.8)	(17.0)	(99.2)	(5.5)	(92.0)	(82.9)
KEENAN	4.3	15.8	87.9	4.5	86.5	83.0
	(4.3)	(16.5)	(88.3)	(4.9)	(86.8)	(83.7)
TSAY1	5.0	18.8	99.2	6.0	98.8	89.8
	(5.0)	(19.7)	(99.3)	(6.0)	(98.8)	(90.0)
TSAY2	3.6	18.4	97.1	5.6	98.8	92.4
	(3.8)	(18.9)	(97.2)	(6.1)	(98.9)	(92.8)
WHITE1	37.2	42.3	84.4	6.3	100.0	97.2
	(42.7)	(50.0)	(88.6)	(9.3)	(100.0)	(98.4)
WHITE2	22.7	34.3	81.9	6.1	100.0	97.6
	(28.0)	(41.9)	(86.2)	(8.6)	(100.0)	(98.7)
WHITE3	21.7	28.8	88.9	5.0	100.0	99.4
	(21.6)	(28.7)	(88.9)	(5.0)	(100.0)	(99.4)
MCLEOD	6.2	8.9	37.1	4.6	79.4	69.0
	(5.9)	(8.8)	(36.1)	(4.2)	(79.3)	(68.9)
RESET1	4.0	16.9	86.2	4.1	78.2	66.5
	(4.1)	(16.9)	(86.6)	(4.2)	(78.2)	(67.0)
RESET2	4.0	16.1	86.7	4.1	77.3	62.9
	(3.8)	(16.1)	(86.7)	(4.0)	(77.2)	(62.7)
BDS	(7.3)	(8.1)	(32.8)	(6.1)	(99.4)	(97.1)

[a] Power using 5% critical value simulated with Model4 (AR(2)) is shown (except for BDS test), and power using 5% asymptotic critical value is shown in parentheses. Sample size = 200, replications = 1000. The results for BDS test in Liu (1990) with embedding dimension $m = 2$ and $\varepsilon = (0.8)^6$ are reported here. The series were divided by the range of the data before applying the BDS tests.

Table 6
Power vs. sample size for Block1 and bivariate models.[a]

	AR	BL	TAR	SGN	NAR	SQ	EXP
			NEURAL1				
$n = 50$	4.8	27.7	32.2	49.0	9.0	100.0	98.7
$n = 100$	5.6	43.0	54.3	84.1	13.8	100.0	100.0
$n = 200$	5.2	58.0	79.8	98.9	20.4	100.0	100.0
			TSAY1				
$n = 50$	4.6	25.3	9.6	17.8	12.2	100.0	99.3
$n = 100$	5.1	33.0	7.5	16.1	16.8	100.0	100.0
$n = 200$	4.7	39.3	4.2	15.0	23.8	100.0	100.0
			WHITE3				
$n = 50$	5.0	81.0	4.8	6.0	4.3	20.1	17.8
$n = 100$	4.1	98.0	4.0	8.8	5.2	47.1	32.6
$n = 200$	5.5	99.5	4.2	12.1	5.5	79.5	60.0
			RESET2				
$n = 50$	4.6	24.9	10.6	21.5	12.5	69.6	47.1
$n = 100$	5.2	33.9	8.6	26.5	18.2	84.9	57.0
$n = 200$	5.2	40.2	7.0	32.7	26.2	95.4	59.5

[a] Power using 5% simulated critical values is shown. Replications = 1000, sample size $n = 50, 100, 200$.

Table 7
Power vs. sample size for Block2 models.[a]

	Model1	Model2	Model3	Model4	Model5	Model6
			NEURAL1			
$n = 50$	5.7	10.7	49.3	4.9	64.3	52.0
$n = 100$	5.2	13.2	84.5	5.5	86.2	73.1
$n = 200$	5.0	17.2	98.1	5.8	94.2	85.8
			TSAY1			
$n = 50$	6.3	10.6	51.5	4.4	74.7	56.8
$n = 100$	4.6	12.4	85.6	4.5	94.5	78.1
$n = 200$	5.0	18.8	99.2	6.0	98.8	89.8
			WHITE3			
$n = 50$	6.3	13.0	30.2	4.6	79.2	71.4
$n = 100$	10.8	18.6	58.3	5.8	99.6	95.6
$n = 200$	21.7	28.8	88.9	5.0	100.0	99.4
			RESET2			
$n = 50$	5.9	6.9	34.5	6.1	36.0	35.4
$n = 100$	3.3	11.5	60.5	6.2	56.7	49.8
$n = 200$	4.0	16.1	86.7	4.1	77.3	62.9

[a] Power using 5% simulated critical values is shown. Replications = 1000, sample size $n = 50, 100, 200$.

and the nonlinear moving average model (Model3), but little power against the heteroskedastic MA model (Model2).

Similar tables were constructed for sample sizes 50 and 100 but are not shown in full detail. Tables 6 and 7 show power against sample size for four of the tests, NEURAL1, TSAY1, WHITE3, and RESET2. The results are generally as expected, with power increasing as sample size increases for most of the nonlinear models.

From these results and others not shown some general comments can be made:

(i) NEURAL1 and NEURAL2 are virtually the same; thus the extra work for NEURAL2 is probably not worthwhile.
(ii) TSAY1 and TSAY2 both have nuisance parameter problems; TSAY1 seems to be better.
(iii) The WHITE tests all have nuisance parameter problems; WHITE3 is generally better.
(iv) RESET1 and RESET2 are virtually identical, RESET2 being marginally better.
(v) The McLeod–Li test is generally weak compared to the alternatives.
(vi) The TSAY1 test is virtually always more powerful than the KEENAN test.
(vii) The BDS test is good with bilinear data and has average power in other cases.
(viii) No single test was uniformly superior to the others.

Table 8 shows the power of four tests in the bivariate case $y_t = x_t^2 + \varepsilon_t$, $\varepsilon_t \sim \mathrm{N}(0, \sigma^2)$, where $x_t = 0.6x_{t-1} + e_t$, $e_t \sim \mathrm{N}(0, 1)$. The three values of σ chosen were 1, 5, and 20, corresponding to approximate signal-to-noise ratios:

σ	1	5	20
$\mathrm{var}(x^2)/\mathrm{var}(\varepsilon)$	7.0	0.28	0.019

Not all situations were simulated. The NEURAL1 and TSAY1 tests are quite powerful in most situations, and have some power even with signal-to-noise ratios around 2% (corresponding to $\sigma = 20$).

Table 9 shows the same information for the bivariate model $y_t = \exp(x_t) + \varepsilon_t$, $\varepsilon_t \sim \mathrm{N}(0, \sigma^2)$, where x_t is as before and the same set of σ values are used. The signal-to-noise ratios are

σ	1	5	20
$\mathrm{var}(\exp(x))/\mathrm{var}(\varepsilon)$	2.16	0.086	0.005

Table 8
Power vs. sample size and noise for bivariate model (SQ).[a]

	$\sigma = 1$	$\sigma = 5$	$\sigma = 20$
	NEURAL1		
$n = 50$	100.0		
$n = 100$	100.0		
$n = 200$	100.0	100	30
	TSAY1		
$n = 50$	100.0		
$n = 100$	100.0		
$n = 200$	100.0	100	42
	WHITE3		
$n = 50$	20.1		
$n = 100$	47.1		
$n = 200$	79.5	5	9
	RESET2		
$n = 50$	69.6		
$n = 100$	84.9		
$n = 200$	95.4	66	17

[a] Power using 5% simulated critical values is shown. 1000 replications for $\sigma = 1$, 100 replications for $\sigma = 5$, 20. Not all situations were simulated. Sample size $n = 50,100,200$.

Table 9
Power vs. sample size and noise for bivariate model (EXP).[a]

	$\sigma = 1$	$\sigma = 5$	$\sigma = 20$
	NEURAL1		
$n = 50$	98.7		
$n = 100$	100.0		
$n = 200$	100.0	99	25
	TSAY1		
$n = 50$	99.3		
$n = 100$	100.0		
$n = 200$	100.0	97	24
	WHITE3		
$n = 50$	17.8		
$n = 100$	32.6		
$n = 200$	60.0	15	4
	RESET2		
$n = 50$	47.1		
$n = 100$	57.0		
$n = 200$	59.5	24	10

[a] Power using 5% simulated critical values is shown. 1000 replications for $\sigma = 1$, 100 replications for $\sigma = 5$, 20. Not all situations were stimulated. Sample size $n = 50,100,200$.

The results are similar to the previous case. It is encouraging that the NEURAL and TSAY tests have such good power in many cases.

6. Tests on actual economic time series

To illustrate the behavior of the various tests for actual economic time series, five economic series were analyzed. The series were first transformed to produce stationary sequences. If z_t is the original series, $y_t = \Delta z_t$ or $y_t = \Delta \log z_t$ is fitted by AR(p), where p is determined by the SIC criterion [Sawa (1978) Hannan (1980)]. Thus $X_t = (y_{t-1}, \dots, y_{t-p})'$. For the moment, we assume the absence of ARCH effects. We discuss the consequences of ARCH below.

Table 10 shows asymptotic p-values for the various tests; a low p-value suggests rejection of the null. As the neural network test involves randomly selecting the Γ parameters, it can be repeated several times with different draws. We obtain p-values for several draws of the neural network test, but these are not independent. Despite dependence, the Bonferroni inequality provides an upper bound on the p-value. Let $P_1, \dots, P_m$ be p-values corresponding to m test

Table 10
Tests on actual economic time series.[a]

Test	(1)	(2)	(3)	(4)	(5)
NEURAL	0.288	0.001	0.024	0.000	0.070
	0.277	0.001	0.024	0.000	0.975
	0.289	0.243	0.024	0.000	0.623
	0.283	0.003	0.024	0.000	0.749
	0.283	0.053	0.024	0.000	0.451
Simple Bonferroni	1.387	0.003	0.118	0.000	0.349
Hochberg Bonferroni	0.289	0.003	0.024	0.000	0.349
KEENAN	0.533	0.000	0.727	0.001	0.888
TSAY	0.532	0.000	0.726	0.001	0.066
WHITE1	0.131	0.127	0.423	0.014	0.059
WHITE2	0.116	0.134	0.423	0.015	0.061
WHITE3	0.254	0.000	0.579	0.001	0.012
MCLEOD	0.743	0.000	0.960	0.000	0.025
RESET1	0.122	0.004	0.040	0.000	0.682
RESET2	0.118	0.004	0.040	0.000	0.675
BISPEC	0.001	0.000	0.000	0.014	0.403
BDS	0.724	0.000	0.921	0.000	0.060

[a](1) US/Japan exchange rate, $y_t = \Delta \log z_t$, AR(1), 1974:1–1990:7, monthly, 199 observations; (2) US three-month T-bill interest rate, $y_t = \Delta z_t$, AR(6); (3) US M2 money stock, $y_t = \Delta \log z_t$, AR(1); (4) US personal income, $y_t = \Delta \log z_t$, AR(1); (5) US unemployment rate, $y_t = \Delta z_t$, AR(4). Series (2), (3), (4), and (5) are monthly for 1959:1–1990:7 with 379 observations. Series (1) and (2) are not seasonally adjusted while the others are. BDS test: embedding dimension $m = 2$, $\varepsilon = (0.8)^6$.

statistics, and $P_{(1)}, \ldots, P_{(m)}$ the ordered p-values. The Bonferroni inequality leads to rejection of H_0 at the α level if $P_{(1)} \leq \alpha/m$, so $\alpha = mP_{(1)}$ is the Bonferroni bound. A disadvantage of this simple procedure is that it is based only on the smallest p-value, and so may be too conservative. Holm (1979), Simes (1986), Hommel (1988, 1989), and Hochberg (1988) discuss improved Bonferroni procedures. Hochberg's modification is used here, defined by the rule 'reject H_0 at the α level if there exists a j such that $P_{(j)} \leq \alpha/(m - j + 1)$, $j = 1, \ldots, m$'. The improved Bonferroni bound is $\alpha = \min_{j=1,\ldots,m}(m - j + 1)P_{(j)}$.

In table 10, $m = 5$ neural network tests are conducted for each series. Both simple and Hochberg Bonferroni bounds are reported. These results illustrate a cautionary feature of the neural network test, as quite different p-values are found for different draws. Thus, if one had relied on a single use of the test, quite different conclusions would be possible, just as would be true if one relied on single but different standard tests for nonlinearity. Use of Bonferroni procedures with multiple draws of the neural network test appears to provide useful insurance against using a single test that by chance looks in the wrong direction.

Although the results illustrate application of several tests for neglected nonlinearity to various economic time series, we must strongly emphasize that they do not by themselves provide definitive evidence of neglected nonlinearity in mean. The reason is that, unlike the situation found in our simulations, we cannot be sure that there are not other features of the series studied that lead to the observed results; in particular, the possible presence of ARCH effects cannot be ruled out.

Generally, ARCH will have one of two effects: either it will cause the size of the test to be incorrect while still resulting in a test statistic bounded in probability under the null (as for the neural network, KEENAN, TSAY, WHITE, and RESET tests), or it will directly lead (asymptotically) to rejection despite linearity in mean (as with the McLeod and Li, BDS, and Bispectrum tests). Two remedies suggest themselves: one may either (1) remove the effect of ARCH, or (2) remove ARCH. The first is relevant to tests with adversely affected size. The effect of ARCH can be removed using a heteroskedasticity-consistent covariance matrix estimator in computing the various test statistics. The approach of Wooldridge (1990) may prove especially useful. The second approach is possible whenever one is confidently able to specify the form of the ARCH effect. However, use of a misspecified ARCH model in the procedure will again adversely affect the size of the test. Furthermore, if the alternative is true, the fitted ARCH model can be expected to absorb some or perhaps even much of the neglected nonlinearity. Conceivably, this could have adverse impact on the power of the procedure. Consideration of either of these remedies raises issues that take us well beyond the scope of the present study; their investigation is left to other work.

Thus, we can take the empirical results of this section as indicating that either neglected nonlinearity or ARCH may be present in series (1)–(5), but that further investigation is needed. The results of this paper are only a first step on the way to analyzing methods capable of unambiguous detection of neglected nonlinearity in real-world settings.

7. Conclusions

As with any Monte Carlo study, results obtained can be considerably influenced by the design of the experiment, in this case the choice of nonlinear models and choices made in constructing the test studied. Assuming that the models chosen are of general interest, we have found that several of the tests studied here have good power against a variety of alternatives, but no one of these tests dominates all others. The new neural network test considered in this paper appears to perform as well as or better than standard tests in certain contexts. Application of the tests studied here to actual economic time series suggests the possible presence of neglected nonlinearity, but further investigation is needed to separate out possible effects of ARCH.

References

Ashley, R., D. Patterson, and M. Hinich, 1986, A diagnostic test for nonlinear serial dependence in time series fitting errors, Journal of Time Series Analysis 7, 165–178.

Bierens, H.J., 1987, ARMAX model specification testing, with an application to unemployment in the Netherlands, Journal of Econometrics 35, 161–190.

Bierens, H.J., 1990, A consistent conditional moment test of functional form, Econometrica 58, 1443–1458.

Bierens, H.J. and J. Hartog, 1988, Nonlinear regression with discrete explanatory variables, with an application to the earnings function, Journal of Econometrics 38, 269–299.

Brock, W.A., W.D. Dechert, and J.A. Scheinkman, 1986, A test for independence based on the correlation dimension, Discussion paper (Department of Economics, University of Wisconsin, Madison, WI).

Chan, W.S. and H. Tong, 1986, On tests for nonlinearity in time series analysis, Journal of Forecasting 5, 217–228.

Diaconis, P. and M. Shahshahani, 1984, On nonlinear functions of linear combinations, SIAM Journal on Scientific and Statistical Computing 5, 175–191.

Engle, R.F., 1982, Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflations, Econometrica 50, 987–1007.

Friedman, J.H. and W. Stuetzle, 1981, Projection pursuit regression, Journal of the American Statistical Association 76, 817–823.

Gallant, A.R. and H. White, 1992, On learning the derivatives of an unknown mapping with multilayer feedforward networks, Neural Networks 5, 129–138.

Granger, C.W.J. and A.P. Anderson, 1978, An introduction to bilinear time series models (Vandenhoech und Ruprecht, Göttingen).

Hannan, E.J., 1980, The estimation of the order of an ARMA process, Annals of Statistics 8, 1071–1081.

Hinich, M.J., 1982, Testing for gaussianity and linearity of a stationary time series, Journal of Time Series Analysis 3, 169–176.

Hochberg, Y., 1988, A sharper Bonferroni procedure for multiple tests of significance, Biometrika 75, 800–802.
Holm, S., 1979, A simple sequentially rejective multiple test procedure, Scandinavian Journal of Statistics 6, 65–70.
Hommel, G., 1988, A stagewise rejective multiple test procedure based on a modified Bonferroni test, Biometrika 75, 383–386.
Hommel, G., 1989, A comparison of two modified Bonferroni procedures, Biometrika 76, 624–625.
Hornik, K., M. Stinchcombe, and H. White, 1989, Multi-layer feedforward networks are universal approximators, Neural Networks 2, 359–366.
Hornik, K., M. Stinchcombe, and H. White, 1990, Universal approximation of an unknown mapping and its derivatives using multi-layer feedforward networks, Neural Networks 3, 551–560.
Huber, P., 1985, Projection pursuit, Annals of Statistics 13, 435–475.
Jones, L.K., 1991, A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, Annals of Statistics, forthcoming.
Keenan, D.M., 1985, A Tukey nonadditivity type test for time series nonlinearity, Biometrika 72, 39–44.
Lapedes, A. and R. Farber, 1987, Nonlinear signal processing using neural networks, Paper presented at the IEEE conference on neural information processing system – Natural and synthetic.
Liu, Tung, 1990, How to decide if a series is chaotic: The estimation of correlation exponent, Ph.D. thesis (University of California, San Diego, CA).
McLeod, A.I. and W.K. Li, 1983, Diagnostic checking ARMA time series models using squared residual autocorrelations, Journal of Time Series Analysis 4, 169–176.
Ramsey, J.B., 1969, Tests for specification errors in classical linear least squares regression analysis, Journal of the Royal Statistical Society B 31, 350–371.
Rumelhart, D.E., G.E. Hinton, and R.J. Williams, 1986, Learning internal representations by error propagation, in: D.E. Rumelhart and J.L. McClelland, eds., Parallel distributed processing: Explorations in the microstructures of cognition – 1 (MIT Press, Cambridge, MA) 318–362.
Sawa, T., 1978, Information criteria for discriminating among alternative regression models, Econometrica 46, 1273–1291.
Simes, R.J., 1986, An improved Bonferroni procedure for multiple tests of significance, Biometrika 73, 751–754.
Stinchcombe, M. and H. White, 1989, Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions, in: Proceedings of the international joint conference on neural networks, Washington, DC (IEEE Press, New York, NY) I: 613–618.
Stinchcombe, M. and H. White, 1991, Using feedforward network to distinguish multivariate populations, Discussion paper (University of California, San Diego, CA).
Subba Rao, T. and M. Gabr, 1980, A test for linearity of stationary time series, Journal of Time Series Analysis 1, 145–158.
Subba Rao, T. and M. Gabr, 1984, An introduction to bispectral analysis and bilinear time series models, Lecture notes in statistics 24 (Springer-Verlag, New York, NY).
Tong, H., 1983, Threshold models in nonlinear times series analysis (Springer-Verlag, New York, NY).
Tsay, R.S., 1986, Nonlinearity tests for times series, Biometrika 73, 461–466.
White, Halbert, 1987, Specification testing in dynamic models, in: T.F. Bewley, ed., Advances in econometrics, Fifth world congress – 1 (Cambridge University Press, Cambridge) 1–58.
White, Halbert, 1989a, Some asymptotic results for learning in single hidden layer feedforward network models, Journal of the American Statistical Association 84, 1003–1013.
White, Halbert, 1989b, An additional hidden unit test for neglected nonlinearity in multilayer feedforward networks, in: Proceedings of the international joint conference on neural networks, Washington, DC (IEEE Press, New York, NY) II: 451–455.
White, Halbert, 1990, Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings, Neural Networks 3, 535–550.
White, Halbert, 1992, Estimation, inference and specification analysis (Cambridge University Press, New York, NY).
Wooldridge, J.M., 1990, A unified approach to robust, regression-based specification tests, Econometric Theory 6, 17–43.