QdS, vol. 16, 2014

Testing for (non)linearity in economic time series: a Monte Carlo comparison

Luisa Bisaglia
Dept. of Statistical Sciences, University of Padua, Italy
E-mail: luisa.bisaglia@unipd.it

Margherita Gerolimetto
Dept. of Economics, University Ca' Foscari, Venice, Italy
E-mail:margherita.gerolimetto@unive.it

Abstract In recent years interest has been growing in testing for (non)linearity in economic time series. Several tests are available in literature, some of them are designed to distinguish linearity from a well specified parametric nonlinear model, while others have been developed without a parametric nonlinear alternative in mind. In this paper we review the issue of testing for (non)linearity and examine, via Monte Carlo experiments, the power and size properties of the major linearity tests applied to different nonlinear time series models.

Keywords: Linearity tests; Time series analysis; Nonlinear models for time series

1. Introduction

Linear models have been the focus of theoretical and applied econometrics for most of the 20th Century. Since the contribution by Box and Jenkins (1970), identification and estimation of ARIMA models have become standard statistical tools for economic time series analysis. It was only starting from the 1990s that nonlinear models were greatly developed, also under the stimulus of the economic theory who frequently suggested nonlinear relationships between variables. Consequently, it also emerged the interest in testing whether or not a single economic series or group of series may be generated by a linear model against the alternative that they were nonlinearly related instead.

Linear models have the advantage of being undoubtedly simple and intuitive. However, they also have several limitations, some of which can be overcome via nonlinear modeling: *i)* linear models cannot allow for strong asymmetries in data, *ii)* they are not suitable for data characterized by sudden and irregular jumps, *iii)* they neglect nonlinear dependence, useful for prediction *iv)* they are not suitable for series which are not time reversible. Moreover, a failure to recognize and deal with the presence of nonlinearity in the generating mechanism of a time series can often lead to poorly behaved parameter estimates and to models which miss important serial dependencies altogether. On the other hand, when nonlinear models are used for the analysis of economic time series, the specification of the model is a critical issue. Departures from linearity can be in many directions as nonlinear phenomena are characterized by a huge variety of features and economic theory may be too vague to allow for complete specification.

To assess whether ARIMA models are incapable of fully capturing the dynamics of linear phenomena and possibly to recognize the nonlinear feature, any preliminary analysis should include a linearity check of the Data Generating Process (DGP). Many nonlinearity tests are scattered through the literature[1] but often they do not provide a general answer to the problem since they tend to be designed for detecting special types of nonlinear structures. The purpose of our work is to provide both a review and a comparison of the major tests for detecting nonlinearity in the generating mechanism of an economic time series.[2] In particular, we want to shed some light on how these tests work when applied to a variety of nonlinear models for economic time series via an extensive Monte Carlo simulation experiment, in order to provide a new and fair picture of the performance of the tests, also in comparative terms, while highlighting some particular aspects of the nonlinearity tests.

To our knowledge, there is no recent contribute in literature that compares the tests applied to a variety of parametric models. Of course, there is a number of reviews, among which Davies and Petruccelli (1986), Lee *et al.* (1993), Corduas (1994), Hansen (1999), Teräsvirta (1996), Teräsvirta (2005), Patterson and Ashley (2000) and a very recent one by Giannerini (2012), however they often do not compare via simulations the tests and, in case they do it, the comparison is made only for a very restricted number of tests and a few very specific data generating processes.[3]

A remark is at this stage in order. This survey is restricted to parametric models (for a recent treatment of non parametric models, see Fan and Yao, 2003), and, anyway, to stochastic processes, being chaotic processes beyond the scope of considerations.

The organization of this paper is as follows. Section 2 introduces some nonlinear time series models. Section 3 reviews the most important linearity tests, that will be considered in the Monte Carlo experiment, described in section 4. Section 5 concludes.

[1] The computer codes to implement the tests are often personally written by the researcher and the quality of the codes varies a lot. This is another interesting issue that, however, will not be discussed here, as it is beyond the scope of the paper.

[2] We want to emphasize that in the recent literature there exists a large number of (non)linearity tests, but in this paper we review only those of them that have found application in the analysis of economic time series.

[3] While we are writing this paper some other works on specific nonlinearity topics have appeared in the literature, *e.g.* Giannerini *et al.* (2015) and Chan *et al.* (2015).

2. Some nonlinear time series models

In this Section we briefly review the main nonlinear models that are commonly used in the time series literature.

2.1. Bilinear models

Given a stationary process X_t, a parsimonious representation of X_t as a finite order linear model in the class ARMA(p, q) is:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j a_{t-j} + a_t \tag{1}$$

where $a_t \sim$ WN$(0, \sigma^2)$ and the autoregressive and moving average parts of the model satisfy, respectively, the stationarity and invertibility conditions.

The simplest class of nonlinear models is the bilinear model, developed by control engineers to describe the input-output relationship of a deterministic nonlinear system. Indeed, bilinear models have the property that, although they involve only a finite number of parameters, they can approximate with arbitrary accuracy any "well-behaved" non linear relationship (Priestley, 1978). Successively, bilinear models have been transformed into stochastic models and studied by Granger and Andersen (1978), Rao (1981), Rao and Gabr (1984).

The most general form of the bilinear model, BL(p, q, r, s), as defined in Granger and Andersen (1978), is

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + a_t + \sum_{j=1}^{q} \theta_j a_{t-j} + \sum_{i=1}^{r} \sum_{j=1}^{s} \beta_{ij} X_{t-i} a_{t-j} \tag{2}$$

where $a_t \sim$ IID$(0, \sigma^2)$. This model may be regarded as a direct non linear extension of an $ARMA(p, q)$ model, derived by adding the extra terms $X_{t-h} a_{t-i}$. However, because of the generality of model (2), it is very complex to analyze and consequently theoretical properties, such as stationarity and invertibility conditions have been derived only for special cases.

Although bilinear models are a natural extension of the ARMA models, in literature there are only a few applications of these models. One of the most cited is Maravall (1983), who analyses a Spanish currency time series using bilinear models. In Maravall's view, bilinear models seem particularly appropriate for series with occasional outbursts, i.e. sequences of outliers that seem to require a different regime. Intuitively, the bilinear part is mostly dormant when the usual regime operates, but it becomes operative in case of atypical behaviours, acting so as to smooth outliers. This could also be useful to model, for example, seismological data. Recently, Ling *et al.* (2015) propose a generalized autoregressive conditional heteroskedasticity-type maximum likelihood estimator for estimating the unknown parameters for a special bilinear model. For some recent developments on bilinear models see, for example, Rao and Terdik (2003).

2.2. Self-Exciting Threshold AutoRegressive models

Assuming that X_t is expressed as a nonlinear function of its past

$$X_t = f(X_{t-1}, X_{t-2}, ..., X_{t-p}) + a_t$$

where $a_t \sim \text{IID}(0, \sigma^2)$, Tong and Lim (1980) and Tong (1983) define the Self-Exciting Threshold AutoRegressive model (SETAR) as a piecewise linear approximation of the general nonlinear autoregression form

$$X_t = \sum_{j=1}^{k} \left\{ \phi_0^{(j)} + \phi_1^{(j)} X_{t-1} + \ldots + \phi_{p_j}^{(j)} X_{t-p_j} + \sigma^{(j)} a_t \right\} I(X_{t-d} \in A_j) \qquad (3)$$

where $a_t \sim \text{IID}(0, 1)$, d, $p_1, \ldots, p_j$ are some unknown positive integers, $\sigma^{(j)} > 0$ and $\phi_l^{(j)}$ are unknown parameters and A_j forms a partition of $\mathcal{R}$ in the sense that $\cup_{j=1}^{k} A_j = \mathcal{R}$ and $A_i \cap A_j = \not{0}$ for all $i \neq j$.

The SETAR model is nonlinear, provided that $k > 1$ and its theoretical properties are hard to obtain (Chan and Tong (1990); Chan (1993); Chan and Tsay (1998)). One of the most interesting features of the SETAR model is that for some parameter values it can generate limit cycles, amplitude dependent frequencies and jump phenomena. Intuitively, SETAR models exhibit two or more regimes that work as local data generating processes while the X_{t-d} variable takes a certain value.

A special case of SETAR is the very popular TAR (Threshold Autoregressive model)

$$X_t = \sum_{j=1}^{k} \left\{ \phi_0^{(j)} + \phi_1^{(j)} X_{t-1} + \ldots + \phi_{p_j}^{(j)} X_{t-p_j} + \sigma^{(j)} a_t \right\} I(Z_t \in A_j) \qquad (4)$$

where the self-exiting threshod variable, X_{t-d}, is substituted by a weakly exogenous threshold variable Z_t.

In spite of its apparent simplicity, this model is general enough to capture features, neglected by linear models, but commonly observed in practice, such as asymmetries in declining and rising patterns of a process, or the presence of jumps.

In SETAR models a regime switch happens when the threshold variable crosses a certain threshold, in other words its conditional mean equation is not continuous with discontinuity points at the thresholds. As a consequence, the parameters change between regimes abruptly and this is quite unrealistic for many real time series. In some cases it is reasonable to assume that the regime switch happens gradually in a smooth fashion. Chan and Tong (1986) were the first to propose and develop these soft thresholding models, called Smooth Transition Autoregressive models (STAR), which allow for "smooth" transitions between regimes (see also Teräsvirta (1994), van Dijk *et al.* (2002)). Two popular choices for the smooth transition function are the logistic function and the exponential function. To better understand similarities and differences between SETAR and STAR model, consider, for simplicity, a SETAR(2) model:

$$\begin{aligned} X_t &= \left\{\phi_0^{(1)} + \phi_1^{(1)} X_{t-1} + \ldots + \phi_{p_1}^{(1)} X_{t-p_1} + \sigma^{(1)} a_t\right\} I(X_{t-d} < r) + \\ &+ \left\{\phi_0^{(2)} + \phi_1^{(2)} X_{t-1} + \ldots + \phi_{p_2}^{(2)} X_{t-p_2} + \sigma^{(2)} a_t\right\} I(X_{t-d} \geq r) \end{aligned} \tag{5}$$

In model (5), the observations, x_t, are generated either from the first regime when x_{t-d} is smaller than the threshold r, or from the second regime when x_{t-d} is greater than the threshold. If the binary indicator function is replaced by a smooth transition function $0 < G(x_{t-d}) < 1$ which depends on a transition variable X_{t-d} the model becomes a smooth transition autoregressive (STAR) model :

$$\begin{aligned} X_t &= \left\{\phi_0^{(1)} + \phi_1^{(1)} X_{t-1} + \ldots + \phi_{p_1}^{(1)} X_{t-p_1} + \sigma^{(1)} a_t\right\} (1 - G(X_{t-d})) + \\ &+ \left\{\phi_0^{(2)} + \phi_1^{(2)} X_{t-1} + \ldots + \phi_{p_2}^{(2)} X_{t-p_2} + \sigma^{(2)} a_t\right\} G(X_{t-d}) \end{aligned} \tag{6}$$

In model (6) the observations x_t switch between two regimes smoothly in the sense that the dynamics of x_t may be determined by both regimes, with one regime having more impacts in some times and the other regime having more impacts in other times. Another interpretation is that STAR models actually allow for a *continuum* of regimes, each associated with a different value of $G(X_{t-d})$. Obviously, X_{t-d} could be substituted by an exogenous variable Z_t as in TAR model. For recent and critic reviews about threshold models in time series analysis, see Tong (2011) and Tong (2015).

2.3. Markov Switching models

Hamilton (1989) introduces Markov Switching model of order p, denoted by MS(p). In case of two regimes, the model can take the following form:

$$X_t = \begin{cases} \alpha_1 + \sum_{i=1}^{p} \phi_{1,i} X_{t-i} + a_{1,t} & \text{if } s_t = 1 \\ \alpha_2 + \sum_{i=1}^{p} \phi_{2,i} X_{t-i} + a_{2,t} & \text{if } s_t = 2 \end{cases} \tag{7}$$

where $a_{i,t} \sim$ IID$(0, \sigma_i^2)$ independent of each other, and s_t assumes values $1, 2$.

The state variable s_t is unobservable and we assume that it is governed by a first order Markov chain with transition probabilities:

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

where $p_{ij} = P(s_t = j|s_{t-1} = i)$ and $p_{11} + p_{12} = p_{21} + p_{22} = 1$.
A small p_{ij} means that the model tends to stay longer in state i. The expected duration of the process to stay in state i is $1/p_{ij}$. The number of regime can be $r \geq 2$.

Although the MS(p) model looks very similar to the SETAR, there is a crucial difference. In particular, in the SETAR model the regimes are defined by the past values of the time series itself and the transition between regimes are governed by a deterministic scheme, once X_{t-d} is observed. In the MS(p) model, instead, regimes are defined by the exogenous state of the Markov chain; the transition scheme is stochastic, hence one is never certain about which state X_t belongs to in a MS model. This difference has important practical implications in forecasting. In a MS(p) model, when the sample size is large, one can use some filtering techniques to draw inferences on the state of X_t, while in a SETAR model, as long as X_{t-d} is observed, the regime of X_t is known. Thus, forecasts of a MS(p) model are always a linear combination of forecasts produced by submodels of individual states. Those of a SETAR model, instead, only come from a single regime provided that X_{t-d} is observed. It is only when the forecast horizon exceeds the delay d also SETAR forecasts become a linear combination of those produced by models of individual regimes.

Moreover, it is much harder to estimate a MS(p) model, because the states are not directly observable. In order to estimate the parameters of a MS model with this uncertainty, one must compute probabilities associated with each possible regime. Such probabilities are estimated using Hamilton's recursive filter (Hamilton, 1994).

Following McCulloch and Tsay (1993) it is possible to generalize the MS model by considering the transition probabilities as logistic or probit functions of some explanatory variable available at time $t - 1$.

2.4. Long-memory models

It is generally accepted that many time series of practical interest exhibit strong dependence, i.e., long memory. For such series, the sample autocorrelations decay slowly and the spectral density exhibits a pole at the origin. To describe these features, a particular class of models is required, one such is the class of the autoregressive fractionally integrated moving average (ARFIMA) models. Although ARFIMA are linear models, they are often considered nonlinear, because their features change dramatically the statistical behaviour of estimates and predictions. As a consequence, many of the theoretical results and methodologies used for analyzing short memory linear time series (as for example ARMA processes) are no longer appropriate for long memory models. For these reasons we also consider the class of ARFIMA models as nonlinear.

There exist different definitions of long memory processes. In the time domain, a stationary discrete time series is said to be long memory if its autocorrelation function decays to zero like a power function. This definition implies that the dependence between successive observations decays slowly as the number of lags tends to infinity. On the other hand, in the frequency domain, a stationary discrete time series is said to be long memory if its spectral density is unbounded at the zero frequency. Other definitions are equivalent and can be found in Beran (1994). More recently Boutahar *et al.* (2007) provides an updated review on the topic.

In this paper we consider one of the most popular long memory processes that takes into account this particular behaviour of the autocorrelation and of the spectral density function, i.e. the ARFIMA(p, d, q), independently introduced by Granger and Joyeux (1980) and Hosking (1981). This process simply generalizes the usual ARIMA(p, d, q) process by allowing d to assume any real value.

Let $a_t \sim$ WN$(0, \sigma^2.)$ The process $\{X_t,\ t \in \mathbf{Z}\}$ is said to be an ARFIMA(p, d, q) process with $d \in (-0.5, 0.5)$, if it is stationary and satisfies the difference equation

$$\Phi(B)\,\Delta(B)\,(X_t - \mu) = \Theta(B)a_t, \tag{8}$$

where $\Phi(z)$ and $\Theta(z)$ are polynomials of degree p and q, respectively, satisfying $\Phi(z) \neq 0$ and $\Theta(z) \neq 0$ for all z such that $|z| \leq 1$, B is the backward shift operator, $\Delta(B) = (1 - B)^d = \sum_{j=0}^{\infty} \pi_j B^j$ with $\pi_j = \Gamma(j - d)/[\Gamma(j + 1)\Gamma(-d)]$, and $\Gamma(\cdot)$ is the gamma function.

The estimation of the long memory parameter d has been of interest for many authors (see Palma (2007) for a good review). In the following we will concentrate on ARFIMA processes with $d \in (0, 0.5)$: for this range of values the process is stationary, invertible and possesses long range dependence.

2.5. ARCH class models

Data in which the variances of the error terms change with the time t, suffer from heteroskedasticity. The standard warning is that in the presence of heteroskedasticity, the regression coefficients for an ordinary least squares regression are still unbiased, but the standard errors and confidence intervals estimated by conventional procedures will be too narrow, giving a false sense of precision. Instead of considering this as a problem to be corrected, ARCH and GARCH models treat heteroskedasticity as a feature to be modeled. As a result, not only are the deficiencies of least squares corrected, but a prediction is computed for the variance of each error term.

The ARCH and GARCH models (AutoRegressive Conditional Heteroskedasticity and Generalized AutoRegressive Conditional Heteroskedasticity) are designed to deal with these issues. They have become widespread tools for dealing with time series heteroskedastic models. The goal of such models is to provide a volatility measure that can be used in financial decisions concerning risk analysis, portfolio selection and derivative pricing.

The first model that provides a systematic framework for volatility modeling is the ARCH model of Engle (1982), used to parametrize conditional heteroskedasticity in a wage-price equation for the United Kingdom.

Formally, let ϵ_t be a random variable that has a mean and a variance conditionally on the information set F_{t-1} (the σ-field generated by $\epsilon_{t-j},\ j \geq 1$), an ARCH$(p)$ model assumes that:

$$\epsilon_t = \sigma_t a_t, \ \ \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2$$

where $a_t \sim$ IID$(0, 1)$, $\alpha_0 > 0$ and $\alpha_i \geq 0,\ i = 1, 2, \ldots, p$.
The parameter restrictions form a necessary and sufficient condition for positivity of the conditional variance. In practice a_t is often assumed to follow the $N(0, 1)$ or a standardized Student t-distribution. It is possible to prove that: (*i*) the unconditional variance of ϵ_t is constant, that is, unconditionally the process is homoskedastic; (*ii*) ϵ_t have zero-autocovariances; (*iii*) ϵ_t has heavier tails than the Normal distribution (heavy tails are a common feature of financial data, for this reason ARCH models are very popular in this field). Besides that, other reasons for choosing ARCH models are that they are simple and easy to handle, they take care of clustered errors, nonlinearities and changes in the econometricians ability to forecast.

In spite of their simplicity, ARCH models often require many parameters to adequately describe the volatility process of an asset return, thus Bollerslev (1986) proposes a useful extension known as the Generalized ARCH (GARCH) model.

Formally a GARCH(p, q) model assumes that:

$$\epsilon_t = \sigma_t a_t, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \tag{9}$$

where $a_t \sim$ IID$(0, 1)$, $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$ and $\sum_{i=1}^{max(p,q)} (\alpha_i + \beta_i) < 1$.

The latter constraint on $\alpha_i + \beta_i$ implies that the unconditional variance of ϵ_t is finite, whereas its conditional variance σ_t^2 evolves over time, a_t is often assumed to be a standard normal or standardized Student-t distribution.

A possible limitation of ARCH and GARCH models is that they assume that positive and negative shocks have the same effects on volatility as the latter depends on the square of the previous shocks. Actually, many financial series respond differently to positive and negative shocks and ARCH models do not provide any new insight for understanding the source of variations of this type of time series. To overcome this many others ARCH-type models (IGARCH, EGARCH, GARCH-M, CHARMA, APARCH, FIGARCH, ...) have been developed in literature (see, for example, Tsay, 2010). Finally, for nonlinear GARCH models see also Teräsvirta (2006) .

3. Testing linearity

From a general and intuitive point of view[4], to test for (non)linearity, the system of hypotheses is:

$$\begin{cases} H_0 : linearity \\ H_1 : nonlinearity \end{cases}$$

Sometimes, the DGP under H_1 is specifically prechosen and in this case testing for nonlinearity is in fact testing for a specific nonlinear structure. In some other cases, the DGP under H_1 is still relatively general and the problem of hypothesis testing is then also generic.

3.1. Linearity against non specific nonlinear alternatives

McLeod and Li (1983) test

A portmanteau-test type statistic, based on the autocorrelation function of squared residuals obtained from an ARMA model fit, has been proposed by McLeod and Li (1983). The idea is to apply the Ljung-Box statistics to the squared residuals of an ARMA(p, q) model to check for model inadequacy. Consequently, the null hypothesis is H_0 : ARMA(p, q) and the test statistic is:

$$Q(m) = n(n+2)\sum_{i=1}^{m} \frac{\hat{\rho}_i^2(a_t^2)}{n-i}$$

where n is the sample size, m is a properly chosen number of autocorrelations used in the test, a_t denotes the residual series, and $\hat{\rho}_i(a_t^2)$ is the lag-i ACF of a_t^2. Under the null hypothesis

$$Q(m) \rightarrow \chi^2_{m-p-q}$$

where the number of degrees of freedom is $m-p-q$ since the statistic is computed from the observed residuals and, typically, m is taken around 20. The motivation for using squared data values to detect nonlinearity is provided by a result inherent in the work of Granger and Newbold (1976). They showed that for a series X_t which is normal (and therefore linear)

$$\rho_k(X_t^2) = (\rho_k(X_t))^2$$

Consequently, any departure from this result presumably would indicate a degree of nonlinearity, as pointed out by Granger and Andersen (1978).

The Q-statistic is also useful in detecting conditional heteroskedasticity of a (returns) series ϵ_t and is asymptotically equivalent to the Lagrange multiplier test statistic of Engle (1982) for ARCH models illustrated in the next pages.

[4] In the following subsections for each test the specific H_0 *vs* H_1 settings will be clarified in detail.

Under this circumstance, the null hypothesis of the statistic is

$$H_0 : \beta_1 = \cdots = \beta_m = 0$$

under the alternative that at least one of the $\beta_i, \quad i = 1, \ldots, m$, is significantly different from zero, where β_i is the coefficient of ϵ^2_{t-i} in the linear regression

$$\epsilon_t^2 = \beta_0 + \beta_1 \epsilon_{t-1}^2 + \cdots + \beta_m \epsilon_{t-m}^2 + a_t, \quad t = m+1, \ldots, n$$

As shown by Davies and Petruccelli (1986) via simulations, Q has higher power when the time series is really generated by an ARCH model, whereas it may result quite ineffective with respect to other structures.

BDS test

The BDS test (Brock *et al.*, 1987), developed within chaos theory, is one of the most popular tests for nonlinearity. It is a nonparametric test, originally designed to test for independence and identical distribution (*iid*), but shown to have also power against a large gamma of linear and nonlinear alternatives, Brock *et al.* (1991). Moreover it can be used as a portmanteau test or miss-specification test when applied to the residuals from a fitted model.

The BDS statistics is based on the correlation integral, a measure of the number of times with which temporal pattern are repeated in the data. Given a time series X_t, $t = 1, 2, ..., n$ and define its m-history as $X_t^m = (x_t, x_{t-1}, ..., x_{t-m+1})$, the correlation integral at the embedding dimension m is

$$C_{m,T}(\epsilon) = \sum_{t<s} I_\epsilon \left(X_t^m, X_s^m\right) \left\{ \frac{2}{T_m(T_m - 1)} \right\}$$

where $T_m = T - (m-1)$ and $I_{X_t^m, X_s^m}$ is an indicator function which equals 1 if the sup norm $\|X_t^m - X_s^m\| < \epsilon$ and equals 0 otherwise. Basically, $C_{m,T}(\epsilon)$ counts up the number of m-histories that lie within a hypercube of size ϵ of each other. Put it differently, the correlation integral estimates the probability that any two m-dimensional points are within a distance of ϵ of each other

$$P(|X_t - X_s| < \epsilon, |X_{t-1} - X_{s-1}| < \epsilon, \ldots, |X_{t-m+1} - X_{s-m+1}| < \epsilon)$$

If the X_t are *iid*, this probability should be equal to the following in the limiting case

$$C_{1,T}(\epsilon)^m = P(|X_t - X_s| < \epsilon)^m$$

Brock *et al.* (1996) define the BDS statistics to test the null hypothesis of linearity against the alternative of nonlinearity, as follows

$$V_{m\epsilon} = \sqrt{T} \frac{C_{m,T}(\epsilon) - C_{1,T}(\epsilon)^m}{s_{m,T}}$$

where $s_{m,T}$ is the standard deviation and can be estimated consistently as documented by Brock *et al.* (1987). Under fairly moderate regularity conditions, the BDS statistic converges in distribution to $N(0,1)$

White (1989) and Terasvirta et al (1993) Neural Network tests

The Neural Network test (White, 1989) for neglected nonlinearity, NN test herafter, is built on neural network models. One of the most common is the single hidden layer feedforward network where unit inputs send a vector X of signals X_i, $i = 1, \dots, k$ along links (connections) that attenuate or amplify the original signals by a factor γ_{ij} (weights). The intermediate or hidden processing unit j receives the signals $X_i\gamma_{ij}$, $i = 1, \dots, k$ and processes them. In general, incoming signals are summed by the hidden units so that an output is produced by means of an activation function $\Phi(\tilde{X}', \gamma_j)$, where Φ is typically the logistic function[5] and $\tilde{X} = (1, X_1, \dots, X_k)$, passed to the output layer

$$f(X,\delta) = \beta_0 + \sum_{j=1}^{q} \beta_j \Phi(\tilde{X}'\gamma_j), \quad q \in N \tag{10}$$

where $\beta_0, \dots, \beta_q$ are hidden to output weights and $\delta = (\beta_0, \dots, \beta_q, \gamma_1', \dots, \gamma_q')'$.

The NN test in particular employs a single hidden layer network, augmented by connections from input to output. The output o of the network is

$$o = \tilde{X}'\theta + \sum_{j=1}^{q} \beta_j \Phi(\tilde{X}'\gamma_j)$$

and the null hypothesis of linearity is equivalent to the optimal weights of the network being equal to zero, that is the null hypothesis of the NN test is $\beta_j^* = 0$ for $j = 1, 2, \dots, q$ for given q and γ_j.

Operatively, the NN test can be implemented as a Lagrange multiplier test:

$$\begin{cases} H_0 : E(\Phi_t e_t^*) = 0 \\ H_1 : E(\Phi_t e_t^*) \neq 0 \end{cases}$$

where the elements $\Phi_t \equiv (\Phi(\tilde{X}_t'\Gamma_1), \dots, \Phi(\tilde{X}_t'\Gamma_q))$ and $\Gamma \equiv (\Gamma_1, \dots, \Gamma_q)$ are chosen a priori, independently of X_t and for given q. To practically carry out the test, the element e_t are replaced by the OLS residuals $e_t = y_t - \tilde{X}'\hat{\theta}$, to obtain the test statistic

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

[5] By definition, Φ belongs to a class of flexible functional forms. White (1989) showed that for wide class of nonlinear functions Φ, the neural network can provide arbitrarily accurate approximations to arbitrary functions in various normed function spaces if q is large enough.

where $\hat{W}$ is a consistent estimator of $W^* = var(n^{-1/2}\sum_{t=1}^{n}\Phi_t e_t^*)$ and under H_0 $M_n \xrightarrow{d} \chi^2(q)$. To circumvent multicollinearity of Φ_t with themselves and X_t as well as computational issues when obtaining $\hat{W}_n$, two practical solutions are adopted. First, the test is conducted for $q* < q$ principal components of Φ_t, $\Phi_t e_t^*$. Second, the following equivalent test statistic is used to avoid calculation of $\hat{W}_n$,

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

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

Teräsvirta *et al.* (1993) proved that the result of this test is affected by the presence of the intercept in the power of the logistic function chosen as activation function. Moreover, he documented a loss of power due to the random choice of the γ parameters. Building on this, Teräsvirta *et al.* (1993) replaced the expression $\sum_{j=1}^{q}\beta_j\Phi(\tilde{X}'\gamma_j)$ in (10) with an approximation based on the Taylor expansion and derived an alternative LM test has been shown to have better power properties.

Ramsey (1969) RESET test

Ramsey (1969) proposes a specification test for linear least squares regression analysis, whose argument is that nonlinearity will be reflected in the diagnostics of a fitted linear model if the residuals of the linear model are correlated with terms to a certain power. In other words, this test, referred to as a RESET test, focuses on specification errors in the linear regression, including those coming from unmodeled non-linearity and is readily applicable to linear AR models.

Consider the linear AR(p) model:

$$X_t = \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + a_t.$$

The first step of the RESET test is to obtain the least squares estimate $\hat{\phi}$, compute the residuals $\hat{a}_t = X_t - \hat{X}_t$, and the sum of squared residuals:

$$SSR_0 = \sum_{i=p+1}^{n} \hat{a}_t^2$$

where n is the sample size.

In the second step, consider the linear regression

$$\hat{a}_t = \mathbf{X_{t-1}'}\mathbf{a} + \mathbf{M_{t-1}'}\mathbf{b} + v_t$$

where $\mathbf{X_{t-1}} = (1, X_{t-1}, \ldots, X_{t-p})$ and $\mathbf{M_{t-1}} = (\hat{X}_t^2, \ldots, \hat{X}_t^{s+1})$ for some $s > 1$, and compute the least squares residuals

$$\hat{v}_t = \hat{a}_t - \mathbf{X_{t-1}'}\hat{\mathbf{a}} - \mathbf{M_{t-1}'}\hat{\mathbf{b}}$$

In the third step sum of squared residuals is computed

$$SSR_1 = \sum_{i=p+1}^{n} \hat{v}_t^2$$

If the linear AR(p) model is adequate, then $\mathbf{a}$ and $\mathbf{b}$ should be zero, so the null hypothesis of linearity can be tested in the fourth step by the usual F statistic given by:

$$F = \frac{(SSR_0 - SSR_1)/g}{SSR_1/(n-p-g)} \;\; with \;\; g = s + p + 1$$

which under linearity and normality, has an $F_{g,n-p-g}$.

Keenan's (1985) test and Tsay's (1986) test

Keenan (1985) proposes a nonlinearity test for time series that uses $\hat{X}_t^2$ only and modifies the second step of the RESET test to avoid multicollinearity between $\hat{X}_t^2$ and $\mathbf{X_{t-1}}$. In particular, Keenan assumes that the series can be approximated (Volterra expansion) as follows:

$$X_t = \mu + \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} \theta_u a_{t-u} + \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} \theta_{uv} a_{t-u} a_{t-v}$$

Clearly, if $\sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} \theta_{uv} a_{t-u} a_{t-v}$ is zero, the approximation is linear, so Keenan's idea shares the principle of an F test. The procedure is in the same steps as Ramsey's test. Firstly, select (with a selection criterion, e.g. AIC) the value p of the number of lags involved in the regression, then fit X_t on $(1, X_{t-1}, \ldots, X_{t-p}$ to obtain the fitted values $(\hat{X}_t)$, the residuals set $(\hat{a}_t)$ and the residual sum of squares SSR. Then regress $\hat{X}_t^2$ on $(1, X_{t-1}, \ldots, X_{t-p})$ to obtain the residuals set $(\hat{\zeta}_t)$. Finally calculate

$$\hat{\eta}_t = \frac{\sum_{t=p+1}^{n} \hat{a}_t \hat{\zeta}_t}{\sum_{t=p+1}^{n} \hat{\zeta}_t^{\,2}}$$

and the test statistic equals

$$\hat{F} = \frac{(n-2p-2)\hat{\eta^2}}{(SSR - \hat{\eta^2})}$$

Under the null hypothesis of linearity, i.e.

$$H_0 : \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} \theta_{uv} a_{t-u} a_{t-v} = 0$$

and the assumption that (a_t) are IID Gaussian, asymptotically $\hat{F} \sim F_{1,n-2p-2}$.

Tsay (1986) improved on the power of the Keenan (1985) test by allowing for disaggregated nonlinear variables (all cross products $X_{t-i}X_{t-j}$, $i, j = 1, \ldots, p$) thus generalizing Keenan test by explicitly looking for quadratic serial dependence in the data. While the first step of Keenan test is unchanged, in the second step of Tsay test, instead of $(\hat{X}_t)^2$, the products $X_{t-i}X_{t-j}$, $i, j = 1, \ldots, p$ are regressed on $(1, X_{t-1}, \ldots, X_{t-p}$. Hence, the corresponding test statistic $\tilde{F}$ is asymptotically distributed as $F_{m,n-m-p-1}$, where $m = p(p-1)/2$.

3.2. Linearity against specific nonlinear alternatives

TAR-LR test

Chan and Tong (1986) propose a likelihood ratio (LR) test for discriminating a particular subset of the self-exciting TAR models, i.e. TAR$(2, p, p)$, from linear AR models when p, R and d are known (or assumed). Using the same notation as in the previous section, $H_0 : X_t \sim$ AR(p), is tested against H_1:

$$X_t = \begin{cases} \phi_{1,0} + \sum_{i=1}^{p} \phi_{1,i} X_{t-i} + a_{1,t} & \text{if } X_{t-d} < r \\ \phi_{2,0} + \sum_{i=1}^{p} \phi_{2,i} X_{t-i} + a_{2,t} & \text{if } X_{t-d} \geq r \end{cases}$$

where r is the threshold. Assuming that a_t is IID independent of X_s, $s < t$, the Chan and Tong LR test is given by:

$$LR_1 = \left\{\sigma^2(NL, r)/\sigma^2\right\}^{\frac{n-p+1}{2}}$$

where $\sigma^2(NL, r)$ and σ^2 are the respective estimators of the error variance from TAR$(2; p, p)$ and AR(p) models. Under the null hypothesis of linearity, the AR coefficients in the TAR regimes will be not significantly different, i.e. $H_0 : \phi_i^1 = \phi_i^2$ $(i = 0, 1, \ldots, p)$, and $-2log(LR_1)$ is asymptotically distributed as χ^2_{p+1}. It is well-established that this kind of test suffers from the Davies (1987) problem, since threshold parameter, r, is not identified under the null hypothesis of linearity. The parameter r is referred to as a nuisance parameter under the null hypothesis. Consequently, the asymptotic distribution of the likelihood ratio is very different from that of the conventional likelihood ratio statistics. Chan (1991), and Andrews and Ploberger (1994) provide further discussion on hypothesis testing involving nuisance parameters under the null hypothesis. In practice, r is generally unknown and needs to be estimated.[6] The LR test then turns into:

$$LR_2 = \left\{\sigma^2(NL)/\sigma^2\right\}^{\frac{n-p+1}{2}}$$

As a consequence, the likelihood function is irregular and the asymptotic distribution of the statistics is no longer χ^2. However, Chan and Tong (1986) propose a numerical evaluation of the likelihood function and a likelihood ratio test based on that numerical approximation. For the restricted case indicated above, theoretical results allow tabulation of the asymptotic null distribution of LR_2 (see Moeanaddin and Tong (1988), Chan and Tong (1990), for details).

Engle (1982) LM test

The Lagrange multipliers (LM) test by Engle (1982) has been introduced to test for ARCH effects mainly for its computational simplicity, as the LM test only demands estimation of the linear model. It is equivalent to the F statistic to test for the null hypothesis of coefficients not significantly different from zero in the regression of the squared residuals from the fit of a linear model on the lagged (up to m) values of the same squared residuals.

$$\hat{a}_t^2 = \alpha_0 + \alpha_1 \hat{a}_{t-1}^2 + \cdots + \alpha_m \hat{a}_{t-1}^2 + \epsilon_t, \quad t = m+1, \ldots, n$$

Once the quantities $SSR_0 = \sum_{t=m+1}^{n}(a_t^2 - \bar{a})^2$ and $SSR_1 = \sum_{t=m+1}^{n}\hat{\epsilon}^2$ are computed, the F statistic is easily obtained:

$$F = \frac{(SSR_0 - SSR_1)/m}{SSR_1/(n-2m-1)}$$

that is asymptotically distributed as χ_m^2. Note that, as it is an LM test, it is possible to resort to nR^2 that, asymptotically has the same distribution as F.

4. Monte Carlo experiment

The Monte Carlo experiment presented in this section has the aim of showing the performance in terms of power and size of the (non)linearity tests illustrated in the previous section for various data generating processes (DGPs). The considered sample

[6] Other solutions to this problem involve some sort of integrating out unidentified parameter from the test statistic. In the context of TAR models the problem has been investigated, for example, in Tong (1990) and Hansen (1996) that proposed to calculate test statistic for a grid of possible values of r and then constructing a summary statistic. Tsay (1989) makes use of arranged autoregression and recursive estimation to derive an alternative test for threshold nonlinearity. Tsay test seeks to transform testing threshold nonlinearity into detecting model changes. The idea behind the test is that under the null hypothesis there is no model change.

sizes are $n = 100, 250, 500, 1000$, for 2000 Monte Carlo simulations and the significance level is $\alpha = 0.05$. Simulations are conducted using the software R Development Core Team (2011).

As both size and power are investigated for all considered tests, the Monte Carlo experiment is two-fold. The considered linear DGPs are listed below, for all models the innovations are distributed as $N(0, 1)$:

1. White Noise
2. AR(1), where $\phi = -0.9, -0.5, 0.5, 0.9$
3. MA(1), where $\theta = -0.9, -0.5, 0.5, 0.9$
4. ARMA(1,1), where $\phi = 0.6$, $\theta = 0.3$
5. ARFIMA(0,d,0), where $d = 0.1, 0.3, 0.45$

To study the power of the tests, data are generated under the alternative hypothesis of nonlinearity. The following nonlinear DGPs are considered, innovations are distributed as $N(0, 1)$:

1. ARCH(1), where $X_t = \sigma_t a_t$, $\sigma_t^2 = 0.01 + \alpha X_{t-1}^2$, $\alpha = 0.3, 0.6, 0.9$
2. ARCH(2), where $X_t = \sigma_t a_t$, $\sigma_t^2 = 0.01 + 0.8X_{t-1}^2 + 0.025X_{t-2}^2$
3. GARCH(1), where $X_t = \sigma_t a_t$, $\sigma_t^2 = 0.011 + 0.12X_{t-1}^2 + 0.85\sigma_{t-1}^2$
4. TAR(1,1), where

$$X_t = \begin{cases} -0.5X_{t-1} + a_t & X_{t-1} \leq 1 \\ 0.4X_{t-1} + a_t & X_{t-1} > 1 \end{cases}$$

$$X_t = \begin{cases} 2 + 0.5X_{t-1} + a_t & X_{t-1} \leq 1 \\ 0.5 - 0.4X_{t-1} + a_t & X_{t-1} > 1 \end{cases}$$

$$X_t = \begin{cases} 1 - 0.5X_{t-1} + a_t & X_{t-1} \leq 1 \\ 1 + a_t & X_{t-1} > 1 \end{cases}$$

5. MS(1), where

$$X_t = \begin{cases} -0.5X_{t-1} + a_t & s_t = 1 \\ 0.4X_{t-1} + a_t & s_t = 2 \end{cases}$$

with $p_{11} = p_{22} = 0.5,\ 0.9$.

As for the implementation of the tests, a few remarks are in order. The Tsay test and Keenan test have been conducted for $p = 2, 4$. The BDS test has been implemented for $m = 2, 3$ and $\epsilon = 1$. For the McLeod-Li test the parameter m has been set to $\sqrt{n}$ rounded to the closest integer. The Engle LM test has been run for $m = 5$. Finally the TAR test has been implemented for $d = 1$ and $a = 0.25$, $b = 0.75$.

The results are presented in Tables 1-9 in the Appendix and several comments can be made about both size and power performance of the tests.

As for the size (Tables 2-5), the results reveal that most of the tests have a good behaviour with respect to all considered linear models. Indeed, the percentage of rejections of the null hypothesis is often quite close to nominal level of 0.05, meaning that the examined tests tend to correctly recognize linearity of the time series. This holds with the exception of the BDS test and the TAR-LR test whose empirical size is rather large and, especially for the latter, barely reduces with the sample size. The size of the BDS test, in fact, improves for sample sizes bigger than 500, this is in line with previous results Patterson and Ashley (2000) according to which this test needs very long series to work properly. The tendency of the TAR-LR test to overeject the null hypothesis of linear model even when the DGP is in fact linear can be easily explained by considering the piecewise linear nature of the TAR models.

In terms of power (Tables 6-10), we expect that tests designed to recognize non linearity in mean (variance) perform better in case the DGP is nonlinear in mean (variance). In general, this is confirmed by the results of the experiment. In case of ARCH/GARCH DGPs the tests with the highest power are McLeod and Li test and Engle LM test, in case of TAR DGPs the test with the largest power is the LR-TAR test, followed by the tests Tsay, Keenan, Terasvirta and White. There is no big difference between the power obtained by the LR-TAR test and Tsay, Keenan, Terasvirta and White. This interesting result reveals that these tests work well in case of TAR models. The only test that exhibits large power both for ARCH/GARCH and TAR DGPs is the BDS, without forgetting that the sample size should be larger than 500.

In case of MS models, the performance of the tests changes. One could expect the power results being similar to those obtained for TAR DGPs as these models share with MSs the same regime switching nature. In fact, the responses of the tests are quite different. Tsay, Keenan, Terasvirta and White test exhibits very poor power, while the McLeod and Li test and Engle LM test (although they are designed to detect nonlinearity in variance) are characterized by extremely good power that reaches high values at the increase of the sample size.[7]

Finally, some ARFIMA models (Tables 11) have been included in the experiment to find out whether some of the tests could capture their peculiarity compared to ARMA models. In general the tests do not recognize elements of difference from the linearity. It is only when d is close to 0.5 that Keenan's test, Terasvirta and White tests, can distinguish ARFIMA from ARMA linear models.

All in all, there is a great deal of variation of the power. As shown in the synoptic

[7] These results are in line with those obtained by Patterson and Ashley (2000).

Test	Comment
BDS	very good power starting from $n = 100$ in case of ARCH/GARCH DGPs
	good power when $n > 500$ in case of TAR DPGs
	good power starting from $n = 100$ in case of MS DPGs
McLeod-Li	very good power starting from $n = 100$ in case of ARCH/GARCH DGPs
	good power starting from $n = 250$ in case of MS DPGs
EngleLM	very good power starting from $n = 100$ in case of ARCH/GARCH DGPs
	good power starting from $n = 250$ in case of MS DPGs
Tsay/Keenan	acceptably good power in case of ARCH/GARCH DGPs (better Tsay than Keenan)
	very good power starting from $n = 100$ in case of TAR DPGs
Terasvirta/White	acceptably good power in case of ARCH/GARCH DGPs (the test perform similarly)
	very good power starting from $n = 100$ in case of TAR DPGs
TAR-LR	very good power starting from $n = 100$ in case of TAR DPGs

Table 1. Summary of the power study

table below, several of the tests studied have a good power against a variety of alternatives, but no one of the tests dominates all others. The BDS test exhibits most often the highest power in detecting nonlinearity, and for this reason it should be the first to be used. On the other hand it does not provide indications about the type of nonlinearity, hence some other tests must be necessarily employed afterwards. The simulations results show that Tsay test stands as a possible marker for TAR models and has better power properties than Keenan's, hence it should be preferred. Terasvirta and White tests perform similarly to the Tsay test, except for the MS DGP.

5. Conclusion

In this paper we provide a review and a comparative analysis of the main tests to detect nonlinearity in economic time series.

As emphasized by Giannerini (2013), it is difficult to offer a unified framework where all nonlinearity tests can be included. At the end of this comparative analysis work, we can conclude almost all tests are influenced by the specific hypothesis under which they have been conceived and there are few complementaries among the tests. Every single test, in fact, works properly only in specific cases, in which, on the other hand very high power is reached. Testing the presence of specific nonlinear features by means of more than one test to detect nonlinearity, starting from the BDS test, appears to be the safest strategy.

Some results have not been included in this paper. In particular, we did not cover tests in the frequency domain, e.g. Hinich (1982), whose advantage is in their generality, but they are relatively underused and for this reason we gave more space to other tests that are effectively more utilized by practitioners.

References

Andrews, D. and Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica*, 62, 1383–1414.

Beran, J. (1994). Statistics for Long-Memory Processes. Chapman and Hall, London. Bollerslev, T. (1986). Generalized autorregressive conditional heteroskedasticity. *Journal of Econometrics*, 31, 307–327.

Boutahar, M., Marimoutou, V., and Mouira, L. (2007). Estimation methods of the long memory parameter: Monte Carlo analysis and application. *Journal of Applied Statistics*, 34, 261–301.

Box, G. and Jenkins, G. (1970). *Time series analysis: forecasting and control.* Holden Day, San Francisco.

Brock, W., Dechert, W., and Scheinkman, J. (1987). Test for independence based on the correlation dimension. mimeo.

Brock, W., Hsieh, D., and LeBaron, B. (1991). Nonlinear Dynamics, Chaos, and Insta- bility: Statistical Theory and Economic Evidence. MIT Press.

Brock, W., Dechert, W., and Scheinkman, J. (1996). A test for independence based on the correlation dimension. *Econometric Reviews*, 15, 197–235.

Chan, K. (1991). Percentage points of likelihood ratio tests for threshold autoregression. *Journal of the Royal Statistical Society*, Series B, 53, 691–696.

Chan, K. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *The Annals of Statistics*, 21, 520–533.

Chan, K. and Tong, H. (1986). On estimating thresholds in autoregressive models. *Journal of Time Series Analysis*, 7, 178–190.

Chan, K. and Tong, H. (1990). On likelihood ratio tests for threshold autoregression. *Journal of the Royal Statistical Society*, Series B, 52, 469–476.

Chan, K. and Tsay, R. (1998). Limiting properties of the least squares estimator of a continuous threshold autoregressive model. *Biometrika*, 85, 413–426.

Chan, W., Cheung, S., Chow, W., and Zhang, L. (2015). A robust test for threshold-type nonlinearity in multivariatetime series analysis. *Journal of Forecasting* (in press).

Corduas, M. (1994). Nonlinearity tests in time series analysis. *Statistical Methods and Applications*, 3, 291–313.

Davies, N. and Petruccelli, J. (1986). Detecting non-linearity in time series. *The statis- tician*, 35, 271–280.

Davies, R. (1987). Hypothesis testing when a nuisance parameter is present under the alternative. *Biometrika*, 74, 33–43.

Engle, R. (1982). Autorregressive conditional heteroskedasticity with estimates of united kingdom inflation. *Econometrica*, 50, 987–1008.

Fan, J. and Yao, Q. (2003). Nonlinear time series: nonparametric and parametric meth- ods. Springer-Verlag, New York.

Giannerini, S. (2012). The quest for nonlinearity in time series. *Handbook of Statistics: Time Series*, 30, 43–63.

Giannerini, S., Maasoumi, E., and Dagum, E. B. (2015). Entropy testing for nonlinear serial dependence in time series. Biometrika (in press).

Granger, C. and Andersen, A. (1978). *An introduction to bilinear time series models*. Vandenhoeck and Ruprecht, Gottingen.

Granger, C. and Joyeux, R. (1980). An introduction to long-range time series models and fractional differencing. *Journal of Time Series Analysis*, 1, 15–30.

Granger, C. and Newbold, P. (1976). The use of r2 to determine the appropriate transformation of regression variables. Journal of Econometrics, 4, 205–210.

Hamilton, J. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57, 357–384.

Hamilton, J. (1994). *Time series analysis.* Priceton University Press.

Hansen, B. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, 64, 413–430.

Hansen, B. (1999). Testing for linearity. *Journal of economic surveys*, 13, 551–576.

Hinich, M. (1982). Testing for gaussianity and linearity of a stationary time series. Journal of time series analysis, 3, 169–176.

Hosking, J. (1981). Fractional differencing. *Biometrika*, 68, 165–176.

Keenan, D. (1985). A Tukey nonadditivity-type test for time series. *Biometrika*, 72, 39–44.

Lee, T., White, H., and Granger, C. (1993). Testing for neglected nonlinearity in time series models. *Journal of Econometrics*, 56, 269–290.

Ling, S., Peng, L., and Zhu, F. (2015). Inference for a special bilinear time-series model. *Journal of time series analysis*, 36, 61–66.

Maravall, A. (1983). An application of nonlinear time series forecasting. Journal of Business and Economic Statistics, 1, 66–74.

McCulloch, R. and Tsay, R. (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *Journal of the American Statistical Association*, 88, 968–978.

McLeod, A. and Li, W. (1983). Diagnostic checking ARMA time series models using squared-residual autocorrelations. Journal of time series analysis, 4, 269–273.

Moeanaddin, R. and Tong, H. (1988). A comparison of likelihood ratio test and cusum test for threshold autoregression. *Journal of Royal Statistical Society*, Series D: the Statistician, 37, 493–494.

Palma, W. (2007). *Long-Memory Time Series: Theory and Methods*. Wiley series in probability and statistics.

Patterson, D. and Ashley, R. (2000). *A nonlinear time series workshop*. Kluwer Academic Publishers.

Priestley, M. B. (1978). Non-linear models in time series analysis. *Journal of the Royal Statistical Society*. Series D (The Statistician), 27, 159–176.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsey, J. (1969). Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society.* Series B, 31, 350–371.

Rao, T. S. (1981). On the theory of bilinear models. *Journal of the Royal Statistical Society.* Series B, 43, 244–255.

Rao, T. S. and Gabr, M. (1984). *An introduction to bispectral analysis and bilinear time series models.* Springer-Verlag, New York.

Rao, T. S. and Terdik, G. (2003). On the theory of discrete and continuous bilinear time series models. In D. Shanbhag and C. Rao, editors, *Stochastic Processes: Modelling and Simulation, Handbook of Statistics*, vol.21, 827 –870. Elsevier.

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, 89, 208–218.

Teräsvirta, T. (1996). Power properties of linearity tests for time series. *Studies in nonlinear dynamics & econometrics*, 1, 3–10.

Teräsvirta, T. (2005). *Univariate nonlinear time series models.* WP series in Economics and Finance 593, Stockholm School of Economics.

Teräsvirta, T. (2006). *An introduction to univariate GARCH models.* WP series in Economics and Finance 646, Stockholm School of Economics.

Teräsvirta, T., Lin, C., and Granger, C. (1993). Power of the neural network linearity test. *Journal of Time Series Analysis*, 14, 209–220.

Tong, H. (1983). *Threshold models in non-linear time series analysis.* Springer-Verlag.

Tong, H. (1990). *Non-linear time series: a dynamical system approach.* Oxford University Press.

Tong, H. (2011). Threshold models in time series analysis - 30 years on. *Statistics and its Interface,* 4, 107–118.

Tong, H. (2015). Threshold models in time series analysis - some reflections. *Journal of Econometrics* (in press).

Tong, H. and Lim, K. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society,* Series B, 42, 245–292.

Tsay, R. (1986). Non-linearity tests for time series. *Biometrika*, 73, 461–466.

Tsay, R. (1989). Testing and modeling threshold autoregressive processes. *Journal of the American Statistical Association*, 84, 231–240.

Tsay, R. (2010). *Analysis of Financial Time Series.* J.Wiley.

van Dijk, D., Terasvirta, T., and Franses, P. (2002). Smooth transition autoregressive models - a survey of recent developments. *Econometric Reviews*, 21, 1–47.

White, H. (1989). An additional hidden unit test for neglect non-linearity in multilayer feed-forward networks. *Proceedings of the International Joint Conference on Neural Networks*, Washington DC, I, 451–455.

6. Appendix

Table 2. DGP: WN. Empirical size of tests (nominal level 0.05)

WN	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	4.8	5.0	4.5	5.5
Tsay, p=4	4.7	5.5	5.2	4.8
Keenan, p=2	4.5	6.1	4.2	5.6
Keenan, p=4	3.6	5.6	4.5	4.5
Terasvirta	5.6	4.4	3.8	5.4
White	5.5	5.1	3.5	6.0
BDS, m=2	13.9	6.9	5.5	6.3
BDS, m=3	14.0	8.0	6.6	6.5
McLeod-Li	4.4	5.1	5.2	4.7
EngleLM	2.7	4.0	3.9	4.6
TAR-LR	11.6	12.7	10.0	9.9

Table 3. DGP: AR(1). Empirical size of tests (nominal level 0.05)

AR(1)	ϕ=-0.9				ϕ=-0.5			
	$n = 100$	$n = 250$	$n = 500$	$n = 1000$	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	4.5	4.8	4.3	5.3	4.1	5.1	5.3	4.8
Tsay, p=4	4.8	4.4	5.1	4.9	3.7	4.9	4.9	4.9
Keenan, p=2	5.2	5.1	3.8	4.9	5.2	5.7	4.9	4.6
Keenan, p=4	5.1	5.2	4.0	5.0	5.0	5.7	5.9	4.3
Terasvirta	4.8	5.3	4.2	5.6	5.8	5.4	5.2	4.7
White	5.5	5.3	3.8	5.5	6.1	5.4	4.9	5.0
BDS, m=2	12.0	6.9	7.0	5.4	13.4	8.0	5.8	5.6
BDS, m=3	13.3	6.5	6.7	5.7	14.6	7.3	6.0	5.6
McLeod-Li	4.7	5.2	4.5	5.8	4.5	5.3	3.9	5.2
EngleLM	3.1	4.4	4.0	4.5	3.2	3.7	4.0	5.4
TAR-LR	11.8	10.2	9.8	10.3	12.4	11.6	10.6	10.4
AR(1)	ϕ=0.9				ϕ=0.5			
	$n = 100$	$n = 250$	$n = 500$	$n = 1000$	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	4.1	4.2	4.8	5.1	2.5	3.8	3.7	4.7
Tsay, p=4	4.1	4.8	5.4	5.0	3.8	3.3	4.2	4.3
Keenan, p=2	4.3	4.4	4.2	5.2	1.2	2.5	3.1	3.2
Keenan, p=4	4.6	4.7	4.2	5.4	1.1	2.7	3.2	3.2
Terasvirta	4.2	4.0	4.2	4.8	7.3	5.9	6.2	4.6
White	4.6	3.9	4.6	4.3	6.7	5.3	5.9	4.7
BDS, m=2	13.7	7.8	6.0	5.1	13.5	8.1	6.5	5.5
BDS, m=3	14.2	7.2	5.8	4.8	13.6	8.3	6.3	4.7
McLeod-Li	4.4	4.7	5.4	4.2	4.6	5.5	5.2	5.3
EngleLM	3.6	4.3	4.4	4.5	3.0	4.5	4.0	4.5
TAR-LR	13.5	10.9	10.4	9.5	12.1	10.9	10.9	10.3

Table 4. DGP: MA(1). Empirical size of tests (nominal level 0.05)

MA(1)	θ=-0.9				θ=-0.5			
	$n = 100$	$n = 250$	$n = 500$	$n = 1000$	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	3.4	4.1	3.8	3.2	4.6	4.9	3.9	4.4
Tsay, p=4	4.8	5.7	4.6	5.0	4.9	5.2	4.8	6.2
Keenan, p=2	1.5	2.0	1.5	1.4	3.4	3.4	3.5	3.7
Keenan, p=4	1.3	1.1	1.5	2.0	4.4	4.5	4.3	4.7
Terasvirta	8.0	7.0	6.1	5.9	5.9	6.4	5.3	6.3
White	7.2	7.2	6.9	5.7	6.3	6.6	6.0	5.8
BDS, m=2	13.8	7.0	6.2	5.0	13.6	7.7	6.6	5.5
BDS, m=3	14.2	7.1	6.4	4.9	14.7	8.6	6.6	5.7
McLeod-Li	4.3	4.6	4.8	4.7	5.0	4.9	5.7	5.3
EngleLM	3.4	5.3	4.7	5.1	3.1	4.1	4.6	5.3
TAR-LR	13.0	12.8	11.7	11.1	11.4	12.6	11.6	10.4
MA(1)	θ=0.9				θ=0.5			
	$n = 100$	$n = 250$	$n = 500$	$n = 1000$	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	4.9	5.6	6.4	5.3	4.2	6.1	4.5	5.7
Tsay, p=4	4.7	4.4	5.6	3.6	4.9	5.2	5.3	6.5
Keenan, p=2	4.6	6.0	5.4	5.7	4.3	5.4	4.7	5.0
Keenan, p=4	4.6	4.9	5.8	5.2	5.1	5.4	4.9	4.0
Terasvirta	2.9	2.7	2.6	2.9	1.3	0.9	1.1	0.9
White	3.5	2.8	2.7	2.7	1.4	1.0	1.0	1.2
BDS, m=2	13.7	7.9	5.2	6.3	12.2	7.9	6.9	5.2
BDS, m=3	13.7	8.5	5.9	6.2	13.1	7.5	5.4	5.9
McLeod-Li	4.6	4.8	5.2	4.6	3.9	5.0	4.7	5.0
EngleLM	3.4	5.0	5.5	5.4	3.1	4.6	4.2	4.3
TAR-LR	13.3	10.5	8.5	8.9	13.0	10.8	9.2	9.6

Table 5. DGP: ARMA(1,1). Empirical size of tests (nominal level 0.05)

ARMA(1,1)	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	4.0	5.3	5.3	5.7
Tsay, p=4	4.3	4.8	4.7	5.4
Keenan, p=2	3.5	5.4	5.7	6.3
Keenan, p=4	3.3	5.0	5.2	5.2
Terasvirta	1.9	1.4	1.8	1.7
White	1.8	1.5	2.1	2.2
BDS, m=2	13.6	8.2	6.1	5.4
BDS, m=3	12.5	8.1	6.7	5.3
McLeod-Li	4.7	5.1	5.7	5.1
EngleLM	3.3	4.1	4.4	4.3
TAR-LR	10.8	9.6	9.3	8.9

Table 6. DGP: ARCH(1). Empirical power of tests

ARCH(1) - $\alpha = 0.3$	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	14.4	20.7	25.6	30.1
Tsay, p=4	12.8	20.5	26.3	30.5
Keenan, p=2	11.5	14.0	17.0	19.3
Keenan, p=4	8.6	10.0	13.0	13.5
Terasvirta	19.3	27.2	29.9	35.4
White	15.5	21.2	21.8	23.8
BDS, m=2	52.6	85.5	99.0	100.0
BDS, m=3	49.9	81.0	98.0	100.0
McLeod-Li	24.9	64.6	93.9	99.9
EngleLM	29.0	70.8	96.1	99.4
ARCH(1) - $\alpha = 0.6$	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	32.7	45.9	56.1	66.3
Tsay, p=4	35.5	53.7	66.8	77.6
Keenan, p=2	21.2	30.1	37.4	42.8
Keenan, p=4	18.0	26.4	32.3	38.0
Terasvirta	36.7	50.5	61.2	67.8
White	29.3	39.4	47.3	53.1
BDS, m=2	**86.1**	**99.8**	**100**	**100**
BDS, m=3	**83.7**	**99.7**	**100**	**100**
McLeod-Li	**55.1**	**94.1**	**99.9**	**100**
EngleLM	**55.1**	**93.9**	**99.0**	**99.9**
ARCH(1) - $\alpha = 0.9$	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	48.1	66.1	77.2	87.8
Tsay, p=4	58.0	77.6	89.5	96.7
Keenan, p=2	30.3	43.2	55.0	63.2
Keenan, p=4	27.7	39.1	49.6	61.1
Terasvirta	51.9	68.8	76.1	85.8
White	41.4	56.1	65.9	73.8
BDS, m=2	96.9	100.0	100.0	100.0
BDS, m=3	96.2	100.0	100.0	100.0
McLeod-Li	69.1	95.9	99.8	100.0
EngleLM	66.4	93.0	98.1	99.9

Table 7. DGP: ARCH(2). Empirical power of tests

ARCH(2)	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	45.8	64.4	74.8	84.1
Tsay, p=4	53.0	75.7	87.3	94.0
Keenan, p=2	29.9	41.1	48.6	58.3
Keenan, p=4	26.6	36.5	45.6	53.1
Terasvirta	51.0	63.5	72.4	81.6
White	41.3	50.8	59.4	67.9
BDS, m=2	95.0	100.0	100.0	100.0
BDS, m=3	94.2	100.0	100.0	100.0
McLeod-Li	66.1	96.2	99.9	100.0
EngleLM	69.0	98.0	99.1	100.0

Table 8. DGP: GARCH(1,1). Empirical power of tests

GARCH(1,1)	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	11.2	19.4	26.5	33.4
Tsay, p=4	16.1	34.0	45.4	58.4
Keenan, p=2	8.8	13.5	16.1	19.9
Keenan, p=4	7.8	13.2	15.0	20.4
Terasvirta	11.1	17.4	24.4	30.9
White	10.1	12.7	17.7	20.9
BDS, m=2	30.7	**58.8**	**86.4**	**98.9**
BDS, m=3	37.7	**70.4**	**94.7**	**100**
McLeod-Li	32.5	**80.1**	**98.8**	**100**
EngleLM	34.2	**83.9**	**98.3**	**100**

Table 9. DGP: TAR(1,1). Empirical power of tests

TAR$(1,1)$	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	77.0	99.6	100.0	100.0
Tsay, p=4	50.0	96.3	100.0	100.0
Keenan, p=2	65.3	88.4	96.6	99.9
Keenan, p=4	37.7	66.5	80.1	91.7
Terasvirta	86.8	99.9	100.0	100.0
White	91.5	100.0	100.0	100.0
BDS, m=2	41.7	69.9	91.6	99.6
BDS, m=3	38.9	66.0	89.4	99.3
McLeod-Li	8.9	14.1	24.4	43.9
EngleLM	9.1	16.9	28.0	53.7
TAR-LR	90.3	99.9	100	100
TAR$(1,1)$ with constant	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	93.1	100	100	100
Tsay, p=4	73.1	99.5	100	100
Keenan, p=2	66.5	98.3	100	100
Keenan, p=4	12.2	31.1	59.6	89.3
Terasvirta	99.7	100	100	100
White	100	100	100	100
BDS, m=2	15.4	15.0	18.9	22.6
BDS, m=3	24.6	34.7	55.3	83.1
McLeod-Li	5.3	8.0	9.5	14.8
EngleLM	4.1	7.8	12.4	18.3
TAR-LR	100	100	100	100
TAR$(1,1)$ with WN	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	18.1	46.5	77.8	98.3
Tsay, p=4	10.1	26.3	55.5	91.9
Keenan, p=2	40.7	73.2	93.7	99.8
Keenan, p=4	76.0	99.6	100.0	100.0
Terasvirta	33.0	66.7	93.5	99.9
White	36.8	73.5	96.5	99.9
BDS, m=2	13.9	13.9	15.0	22.2
BDS, m=3	14.3	12.8	13.3	19.4
McLeod-Li	4.3	5.3	7.6	9.5
EngleLM	4.3	5.7	7.4	8.7
TAR-LR	36.4	75.7	98.1	100

Table 10. DGP: MS(1). Empirical power of tests

MS(1) $p = q = 0.5$	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	10.6	11.3	13.4	14.0
Tsay, p=4	9.0	10.9	12.9	14.3
Keenan, p=2	8.1	10.1	9.9	11.5
Keenan, p=4	6.5	7.6	8.6	10.5
Terasvirta	15.5	17.5	19.4	20.6
White	12.1	12.8	13.2	14.3
BDS, m=2	**41.8**	**73.2**	**95.6**	**100**
BDS, m=3	**40.4**	**68.2**	**91.8**	**99.9**
McLeod-Li	17.3	**46.3**	**79.9**	**98.8**
EngleLM	11.2	**41.7**	**77.2**	**98.8**
TAR-LR	14.8	12.9	14.3	13.7
MS(1) $p = q = 0.9$	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	10.0	11.5	12.5	13.3
Tsay, p=4	7.6	10.7	10.3	11.7
Keenan, p=2	5.5	5.1	4.6	3.8
Keenan, p=4	3.8	5.0	4.6	4.4
Terasvirta	14.0	20.3	23.9	27.3
White	11.5	14.6	16.5	21.6
BDS, m=2	34.1	68.4	89.8	99.2
BDS, m=3	33.9	64.0	86.6	98.4
McLeod-Li	13.9	38.4	68.8	93.3
EngleLM	8.4	35.4	67.0	92.2
TAR-LR	14.7	14.5	14.9	16.5

Table 11. DGP: ARFIMA(0,d,0). Empirical size of tests (nominal level 0.05)

ARFIMA(0,d,0), $d = 0.1$	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	4.8	4.6	5.0	4.4
Tsay, p=4	4.3	4.1	4.7	4.0
Keenan, p=2	6.8	7.2	10.5	11.5
Keenan, p=4	5.3	5.8	9.1	9.4
Terasvirta	4.7	5.6	5.5	5.2
White	5.4	5.7	5.5	5.3
BDS, m=2	13.8	8.9	6.0	5.4
BDS, m=3	14.5	9.1	6.8	5.3
McLeod-Li	4.9	4.7	4.8	4.3
EngleLM	3.4	4.6	4.0	4.8
ARFIMA(0,d,0), $d = 0.3$	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	4.3	6.5	6.0	7.0
Tsay, p=4	3.5	5.2	5.0	5.8
Keenan, p=2	8.1	12.2	26.1	40.7
Keenan, p=4	5.2	6.4	15.8	27.4
Terasvirta	5.7	7.1	8.0	9.0
White	5.5	5.1	3.5	6.0
BDS, m=2	13.8	8.0	7.0	5.5
BDS, m=3	13.6	7.6	6.6	5.4
McLeod-Li	4.4	4.8	4.6	4.0
EngleLM	3.2	4.8	4.9	4.5
ARFIMA(0,d,0), $d = 0.45$	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Tsay, p=2	4.5	5.8	7.2	10.1
Tsay, p=4	4.2	4.2	4.9	6.4
Keenan, p=2	26.7	27.3	35.3	46.0
Keenan, p=4	22.5	23.8	28.1	35.6
Terasvirta	8.7	11.5	13.0	18.5
White	9.9	12.5	14.8	20.3
BDS, m=2	14.9	8.8	6.7	5.4
BDS, m=3	14.8	7.8	7.0	5.4
McLeod-Li	4.8	4.5	5.3	5.9
EngleLM	3.6	3.8	4.8	5.1