

A Flexible Coefficient Smooth Transition Time Series Model

Marcelo C. Medeiros

Dept. of Electrical Engineering, Catholic University of Rio de Janeiro

Dept. of Economic Statistics, Stockholm School of Economics

Alvaro Veiga

Dept. of Electrical Engineering, Catholic University of Rio de Janeiro

June 27, 2000

Abstract

In this paper, we propose a flexible smooth transition autoregressive (STAR) model with multiple regimes and multiple transition variables. We show that this formulation can be interpreted as a time varying linear model where the coefficients are the outputs of a single hidden layer feedforward neural network. This proposal has the major advantage of nesting several nonlinear models, such as, the Self-Exciting Threshold AutoRegressive (SETAR), the AutoRegressive Artificial Neural Network (AR-ANN), and the Logistic STAR models. Furthermore, if the neural network is interpreted as a nonparametric universal approximation to any Borel-measurable function, our formulation is directly comparable to the Functional Coefficient AutoRegressive (FAR) and the Single-Index Coefficient Regression models. The motivation for developing a flexible model is twofold. First, allowing for multiple regimes is important to model the dynamics of several time series, as for example, the behaviour of macro-economic variables over the business cycle. Second, multiple transition variables are useful in describing complex nonlinear behaviour and allow for different sources of nonlinearity. A model building procedure consisting of specification and estimation is developed based on statistical inference arguments. A Monte-Carlo experiment showed that the procedure works in small samples, and its performance improves, as it should, in medium size samples. Several real examples are also addressed.

Keywords: Time series, smooth transition models, threshold models, neural networks.

JEL Classification Codes: C22, C51

1 Introduction

The past few years have witnessed a vast development of nonlinear time series techniques. Among the large amount of new methodologies, the Smooth Transition AutoRegressive (STAR) model, initially proposed, in its univariate form, by Chan and Tong (1986) and further developed in the papers by Luukkonen, Saikkonen and Teräsvirta (1988) and Teräsvirta (1994), has found a number of successful applications (see van Dijk (1999, Chapter 2) for a recent review). The term “smooth transition” in its present meaning first appeared in a paper by Bacon and Watts (1971). They presented their smooth transition model as a generalization to models of two intersecting lines with an abrupt change from one linear regression to another at some unknown change-point. Goldfeld and Quandt (1972, p. 263–264) generalized the so-called two-regime switching regression model using the same idea.

This paper considers an additive smooth transition time series model with multiple regimes and transitions defined by hyperplanes in a multidimensional space. We show that this model can be interpreted as a time varying linear model where the coefficients are the outputs of a single hidden layer feedforward neural network. The proposed model allows that each regime has distinct dynamics controlled by a linear combination of known variables such as, for example, several lagged values of the time series.

This proposal can be interpreted as a generalization of the STAR model with the major advantage of nesting several nonlinear models, such as, the Self-Exciting Threshold AutoRegressive (SETAR) model (Tong 1990) with multiple regimes, the AutoRegressive Artificial Neural Network (AR-ANN) model (Leisch, Trapletti and Hornik 1999), and the Logistic STAR model (Teräsvirta 1994). The proposed model is also able to fit time series where the true generating process is an Exponential STAR (ESTAR) model (Teräsvirta 1994). Furthermore, our model can be also compared to the Functional Coefficient AutoRegressive (FAR) model of Chen and Tsay (1993), and the Single-Index Coefficient Regression model of Xia and Li (1999).

The motivation for developing a flexible model is twofold. First, allowing for multiple regimes is important to model the dynamics of several time series, as for example, the behaviour of macro-economic variables over the business cycle. Recent studies conclude that a two-regime modelling of the business cycle is rather limited. See for example, van Dijk and Franses (1999) where a Multiple Regime STAR (MRSTAR) model is proposed and applied to describe the behaviour of the US GNP and US unemployment rate and Cooper (1998) where a regression tree approach is used to model multiple regimes in the US industrial production. In the framework

of the SETAR model, modelling multiple regimes is a well established methodology (see Tong (1990) and Tsay (1989) for some examples).

Second, multiple transition variables are useful in describing complex nonlinear behaviour and allow for different sources of nonlinearity. Several papers concerning multiple transition variable have appeared in the literature during the past years. However, they assumed that the transition variable was a known linear combination of individual variables. See, for example, Tiao and Tsay (1994) where the thresholds are controlled by two lagged values of a transformed US GNP series reflecting the situation of the economy or van Dijk and Franses (1999). In the present framework, we adopt a less restrictive formulation, assuming that the linear combination of variables is unknown and is joint estimated with the others parameters of the model. This is a quite flexible approach that lets the data to “speak by themselves” (for different approaches see (Franses and Paap 1999, Lewis and Stevens 1991, Astatkie, Watts and Watt 1997)).

A modelling cycle procedure, based on the work of Eitrheim and Teräsvirta (1996) and Rech, Teräsvirta and Tschernig (1999) consisting of the stages of model specification and parameter estimation, is developed allowing the practitioner to choose among different model specifications during the modelling cycle. A Monte-Carlo experiment showed that the procedure works in small samples (100 observations), and its performance improves, as it should, in medium size samples (500 observations).

The plan of the paper is as follows. Section 2 presents the model. Section 3 deals with the specification of the model. Section 4 analyses the estimation procedures. Section 5 presents a Monte-Carlo experiment to find out the behaviour of the proposed tests and Section 6 shows an example with real data. Concluding remarks are made in Section 7.

2 The Multiple Regime STAR Model with Multivariate Thresholds

One important class of STAR models is the Logistic STAR model of order p , LSTAR(p), proposed by Luukkonen et al. (1988) and defined as

$$y_t = \phi_0^{(1)} + \sum_{i=1}^p \phi_i^{(1)} y_{t-i} + \left(\phi_0^{(2)} + \sum_{i=1}^p \phi_i^{(2)} y_{t-i} \right) F(\gamma(y_{t-d} - c)) + \varepsilon_t, \quad (1)$$

where $F(\cdot)$ is the logistic function

$$F(\gamma(y_{t-d} - c)) = \frac{1}{1 + \exp(-\gamma(y_{t-d} - c))}. \quad (2)$$

The real parameter γ , $\gamma > 0$, is responsible for the smoothness of $F(\cdot)$. The scalar c is the *location parameter* and d is known as the *delay parameter*. The variable y_{t-d} is called the *transition variable*.

It is important to notice that the LSTAR model nests the SETAR model with two regimes. When $\gamma \rightarrow \infty$, model (1) becomes a two-regime SETAR model (Tong 1990, p.183).

In the present paper, we consider an additive Logistic STAR model with multiple regimes and where the transition variable is multivariate. This can be interpreted as a linear model with time-varying coefficients given by the output of a neural network with a single hidden layer, where the transition variable is defined by the inputs of the network. This idea was first introduced in literature by Veiga and Medeiros (1998) (see also Medeiros and Veiga (1999)). We call this model the Neuro-Coefficient Smooth Transition AutoRegressive (NCSTAR) model.

Consider a linear model with time-varying coefficients expressed as

$$y_t = \phi_t' \mathbf{z}_t + \varepsilon_t, \quad (3)$$

where $\phi_t = [\phi_t^{(0)}, \phi_t^{(1)}, \dots, \phi_t^{(p)}]'$ is a $(p+1) \times 1$ vector of real coefficients and $\mathbf{z}_t = [1, \tilde{\mathbf{z}}_t']'$. $\tilde{\mathbf{z}}_t$ is a $p \times 1$ vector of lagged values of y_t and/or some exogenous variables. The random term ε_t is a normally distributed white noise with variance σ^2 . The time evolution of the coefficients $\phi_t^{(j)}$ of (3) is given by the output of a single hidden layer neural network with h hidden units

$$\phi_t^{(j)} = \sum_{i=1}^h \lambda_{ji} F(\omega_i' \mathbf{x}_t - \beta_i) - \lambda_{j0}, \quad j = 0, \dots, p, \quad (4)$$

where λ_{ji} and λ_{j0} are real coefficients.

The function $F(\omega_i' \mathbf{x}_t - \beta_i)$ is the logistic function, where \mathbf{x}_t is a $q \times 1$ vector of input variables, $\omega_i = [\omega_{1i}, \dots, \omega_{qi}]'$ and β_i are real parameters. The norm of ω_i is called the *slope parameter*. In the limit, when the slope parameter approaches infinity, the logistic function becomes a step function.

The neural network architecture representing model (3) is illustrated in Figure 1. The elements of \mathbf{x}_t , called

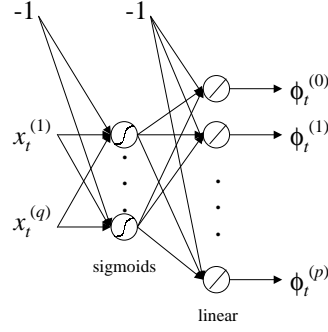


Figure 1: Architecture of the neural network.

the transition variables, can be formed by lagged values of y_t and/or any exogenous variables. Equations (3) and (4) represent a time-varying model with a multivariate smooth transition structure defined by h hidden neurons.

Equation (3) can be rewritten as

$$\begin{aligned}
 y_t = G(\mathbf{z}_t, \mathbf{x}_t; \Psi) + \varepsilon_t &= \alpha_0 + \sum_{j=1}^p \alpha_j y_{t-j} + \sum_{i=1}^h \lambda_{0i} F(\boldsymbol{\omega}'_i \mathbf{x}_t - \beta_i) \\
 &+ \sum_{j=1}^p \left\{ \sum_{i=1}^h \lambda_{ji} F(\boldsymbol{\omega}'_i \mathbf{x}_t - \beta_i) \right\} y_{t-j} + \varepsilon_t,
 \end{aligned} \tag{5}$$

or in vector notation

$$y_t = G(\mathbf{z}_t, \mathbf{x}_t; \Psi) + \varepsilon_t = \boldsymbol{\alpha}' \mathbf{z}_t + \sum_{i=1}^h \boldsymbol{\lambda}'_i \mathbf{z}_t F(\boldsymbol{\omega}'_i \mathbf{x}_t - \beta_i) + \varepsilon_t, \tag{6}$$

where $\Psi = [\boldsymbol{\alpha}', \boldsymbol{\lambda}'_1, \dots, \boldsymbol{\lambda}'_h, \boldsymbol{\omega}'_1, \dots, \boldsymbol{\omega}'_h, \beta_1, \dots, \beta_h]'$ is a parameter vector with $(q+1) \times h + (p+1) \times (h+1)$ elements, $\boldsymbol{\alpha} = [\alpha_0, \dots, \alpha_p]' = [-\lambda_{00}, \dots, -\lambda_{p0}]'$, and $\boldsymbol{\lambda}_i = [\lambda_{0i}, \dots, \lambda_{pi}]'$.

Note that model (6) is, in principle, neither globally nor locally identified. There are three characteristics of neural networks which cause the non-identifiability: the first one is due to the symmetries in the neural network architecture. The likelihood function of the model will be unchanged if we permute the hidden units, resulting in $h!$ possibilities for each one of the coefficients of the model. The second reason is caused by the fact that $F(x) = 1 - F(-x)$, where $F(\cdot)$ is the logistic function. The third reason is the mutual dependence of the parameters $\boldsymbol{\lambda}_i$, $\boldsymbol{\omega}_i$, and β_i , $i = 1, \dots, h$. If all the elements of $\boldsymbol{\lambda}_i$ equal zero, the corresponding $\boldsymbol{\omega}_i$ and β_i can assume any value without affecting the value of the likelihood function. On the other hand, if $\boldsymbol{\omega}_i = \mathbf{0}$, then $\boldsymbol{\lambda}_i$

and β_i can take any value.

The first problem is solved by imposing the restrictions $\beta_1 \leq \dots \leq \beta_h$. The second problem can be circumvented, for example, by imposing the restriction $\omega_{1i} > 0, i = 1, \dots, h$. To remedy the third problem, it is necessary to ensure that the model contains no irrelevant hidden units. This is solved with the tests described in Section 3. For further discussion of the identifiability concepts see, e. g., Sussman (1992), Kurková and Kainen (1994), Hwang and Ding (1997), and Anders and Korn (1999).

Sometimes, specially when deriving the tests in Section 3, is useful to write model (6) as

$$y_t = G(\mathbf{z}_t, \mathbf{x}_t; \Psi) + \varepsilon_t = \boldsymbol{\alpha}'\mathbf{z}_t + \sum_{i=1}^h \boldsymbol{\lambda}'_i \mathbf{z}_t F[\gamma_i(\tilde{\boldsymbol{\omega}}'_i \mathbf{x}_t - c_i)] + \varepsilon_t, \quad (7)$$

where $\gamma_i = \|\boldsymbol{\omega}_i\|$, $\tilde{\boldsymbol{\omega}}_i = \boldsymbol{\omega}_i/\gamma_i$, and $c_i = \beta_i/\gamma_i, i = 1, \dots, h$. Note however, that we are not including a new parameter in the model (the parameter vector Ψ is the same as in (6)), but just using a convenient notation to derive the hypothesis tests. We may combine the notations used in models (6) and (7) whenever necessary.

The choice of the elements of \mathbf{x}_t , which determines the dynamics of the process, allows a number of special cases. An important one is where $\mathbf{x}_t = y_{t-d}$. In this case, model (7) becomes a LSTAR(p) model with $h + 1$ regimes, expressed as

$$y_t = \boldsymbol{\alpha}'\mathbf{z}_t + \sum_{i=1}^h [\boldsymbol{\lambda}'_i \mathbf{z}_t F(\gamma_i (y_{t-d} - c_i))] + \varepsilon_t, \quad (8)$$

It should be noticed that model (8) nests the SETAR model with $h + 1$ regimes. When $\gamma_i \rightarrow \infty, i = 1, \dots, h$, model (8) becomes a SETAR model with $h + 1$ regimes.

Another important case is where $\mathbf{x}_t = t$. In this case the the parameters of a linear model change smoothly as a function of time, and contains as a special case a linear model with h structural breaks, which has been the most popular alternative to parameter constancy in econometrics since its introduction by Chow (1960) and Quandt (1960).

When \mathbf{x}_t is a q -dimensional vector, the dynamic properties of (7) become rather more complex. When $\tilde{\boldsymbol{\omega}}'_i \mathbf{x}_t = c_i$, the parameters $\tilde{\boldsymbol{\omega}}_i$ and c_i define a hyperplane in a q -dimensional Euclidean space

$$\mathbb{H} = \{\mathbf{x}_t \in \mathbb{R}^q \mid \tilde{\boldsymbol{\omega}}'_i \mathbf{x}_t = c_i\}. \quad (9)$$

The direction of $\tilde{\omega}_i$ determines the orientation of the hyperplane and the scalar term c_i determines the position of the hyperplane in terms of its distance from the origin.

A hyperplane induces a partition of the space into two regions defined by the halfspaces

$$\mathbb{H}^+ = \{\mathbf{x}_t \in \mathbb{R}^q \mid \tilde{\omega}'_i \mathbf{x}_t \geq c_i\} \quad (10)$$

and

$$\mathbb{H}^- = \{\mathbf{x}_t \in \mathbb{R}^q \mid \tilde{\omega}'_i \mathbf{x}_t < c_i\}. \quad (11)$$

With h hyperplanes, a q -dimensional space will be split into several polyhedral regions. Each region is defined by the nonempty intersection of the halfspaces (10) and (11) of each hyperplane.

One particular case is when the hyperplanes are parallel to each other. In this case, equation (7) becomes

$$y_t = \boldsymbol{\alpha}' \mathbf{z}_t + \sum_{i=1}^h \lambda'_i \mathbf{z}_t F(\gamma_i (\tilde{\omega}'_i \mathbf{x}_t - c_i)) + \varepsilon_t, \quad (12)$$

and the input space will be split in $h + 1$ regions.

Another interesting case is when $\boldsymbol{\lambda}'_i = [\lambda_{0i}, 0, \dots, 0]$ in (8). Then model (7) becomes an AR-ANN model. AR-ANN models can be interpreted as a linear model where the intercept is time-varying and changes smoothly between regimes

Another important point to mention is that if the neural network is interpreted as a nonparametric universal approximation to any Borel-measurable function to any degree of accuracy, model (7) is directly comparable to the Functional Coefficient AutoRegressive (FAR) model of Chen and Tsay (1993), and the Single-Index Coefficient Regression model of Xia and Li (1999).

3 Specification

In this section a specific-to-general specification strategy is developed. From equation (7) two specification problems require special care. The first one is the variable selection, that is, the correct selection of elements of \mathbf{z}_t and \mathbf{x}_t . The problem of selecting the right subset of variables is very important because selecting a too small

subset leads to misspecification whereas choosing too many variables aggravates the “curse of dimensionality”.

The second problem is the selection of the correct number of hidden units, which is essential to guarantee the identifiability of the model.

The specification strategy adopted here is based on the linearization of the nonlinear term of model (7).

In order to select the variables of (7), it is useful to define $\mathbf{x}'_t = [\mathbf{x}_t^{(1)'}, \mathbf{x}_t^{(2)'}]$ and $\mathbf{z}'_t = [1, \mathbf{z}_t^{(1)'}, \mathbf{z}_t^{(2)'}]$, where $\mathbf{x}_t^{(1)}$ and $\mathbf{z}_t^{(1)}$ are $q_1 \times 1$ vectors with typical element $x_{i,t}^{(1)} = z_{i,t}^{(1)}$, $\mathbf{x}_t^{(2)}$ is $q_2 \times 1$ vector with all elements different from the elements of \mathbf{z}_t , and $\mathbf{z}_t^{(2)}$ is $(p - q_1) \times 1$ vector with all elements different from the elements of \mathbf{x}_t .

The procedure described in Section 3.1 is able to identify the elements of \mathbf{z}_t and the elements of $\mathbf{x}_t^{(2)}$. In order to determine the elements of $\mathbf{x}_t^{(1)}$ we use the linearity test of Section 3.2. After estimating the models, the standard error of the parameter estimates will help to refine the correct set of elements of $\mathbf{x}_t^{(1)}$ and \mathbf{z}_t and to test restrictions on the autoregressive parameters, to test for example, equality of models in different regimes.

3.1 Variable Selection

In the context of STAR models, Teräsvirta (1994) suggests the use of an information criterion such as the AIC (Akaike 1974) or SBIC (Schwarz 1978) to select the order of the autoregression. Then the delay parameter is chosen as to minimize the p -value of the linearity test. The main drawback of this approach is that when the true data generating process is nonlinear, the algorithm tends to select more lags than necessary.

Another possibility is to use nonparametric methods based on local estimators (Vieu 1995, Tjøstheim and Auestad 1994, Yao and Tong 1994, Auestad and Tjøstheim 1990). However, this method requires a large number of observations.

In this paper we adopt the simple procedure proposed by Rech et al. (1999). Their proposal uses global parametric least squares estimation and is based on the Taylor expansion of the model. We give a brief overview of the method. For more details, see Rech et al. (1999).

Consider model (7). The first step is to expand function $G(\mathbf{z}_t, \mathbf{x}_t; \Psi)$ into a k -order Taylor expansion around an arbitrary fixed point in the sample space. This requires that $G(\mathbf{z}_t, \mathbf{x}_t; \Psi)$ has a converging Taylor

expansion. After merging terms, one obtains

$$\begin{aligned}
G(\mathbf{z}_t, \mathbf{x}_t; \Psi) &= \boldsymbol{\pi}'\mathbf{z}_t + \boldsymbol{\xi}'\mathbf{x}_t^{(2)} + \sum_{j_1=1}^q \sum_{j_2=j_1}^q \theta_{j_1 j_2} x_{j_1, t}^{(2)} x_{j_2, t}^{(2)} + \sum_{j_1=1}^p \sum_{j_2=j_1}^p \rho_{j_1 j_2} z_{j_1, t} z_{j_2, t} \\
&+ \sum_{j_1=1}^q \sum_{j_2=1}^p \mu_{j_1 j_2} x_{j_1, t}^{(2)} z_{j_2, t} + \sum_{j_1=1}^q \sum_{j_2=j_1}^q \sum_{j_3=j_2}^q \theta_{j_1 j_2 j_3} x_{j_1, t}^{(2)} x_{j_2, t}^{(2)} x_{j_3, t}^{(2)} + \dots \\
&+ \sum_{j_1=1}^p \dots \sum_{j_k=j_{k-1}}^p \theta_{j_1 \dots j_k} z_{j_1, t} \dots z_{j_k, t} + R(\mathbf{z}_t, \mathbf{x}_t; \Psi),
\end{aligned} \tag{13}$$

where $R(\mathbf{z}_t, \mathbf{x}_t; \Psi)$ is the remainder and the θ 's, ρ 's, and μ 's are real parameters. $\boldsymbol{\pi}$ and $\boldsymbol{\xi}$ are, respectively, $(p+1) \times 1$ and $q_2 \times 1$ vectors of real parameters. Note that the terms involving $\mathbf{x}_t^{(1)}$ merged with the terms involving \mathbf{z}_t .

The second step is to regress y_t on all variables in the Taylor expansion and compute the value of a model selection criterion, AIC or SBIC for example. In this paper we use the SBIC, which is a rather parsimonious criterion. After that, remove one variable from the original model and regress y_t on all the remaining terms in the Taylor expansion and compute the value of SBIC. Repeat this procedure by omitting each variable in turn. Continue by simultaneously omitting two regressors of the original model and proceed in that way until the Taylor expansion consists of a function of a single regressor. Choose the combination of variables that yields the lowest value of the SBIC.

3.2 Testing Linearity

In practical nonlinear time series modelling, testing linearity plays an important role. In the context of model (7), testing linearity has two objectives. The first one is to verify if a linear model is able to adequately describe the data generating process. The second one refers to the variable selection problem. The linearity test is used to determine the elements of $\mathbf{x}_t^{(1)}$. After selecting the elements of \mathbf{z}_t and $\mathbf{x}_t^{(2)}$ with the procedure described in Section 3.1, we choose the elements of $\mathbf{x}_t^{(1)}$ by running the linearity test described below setting $\mathbf{x}_t^{(1)}$ equal to each possible subset of the elements of \mathbf{z}_t and choosing the one that minimize the p -value of the test.

In order to test for linearity, equation (7) is rewritten as

$$y_t = \boldsymbol{\alpha}'\mathbf{z}_t + \sum_{i=1}^h \lambda_i' \mathbf{z}_t F(\gamma(\tilde{\gamma}_i' \tilde{\mathbf{x}}_t)) + \varepsilon_t, \tag{14}$$

where $\tilde{\gamma}'_i = [-c_i, \tilde{\omega}'_i]$ and $\tilde{\mathbf{x}}'_t = [1, \mathbf{x}'_t]$. The transition function $F(\gamma(\tilde{\gamma}'_i \tilde{\mathbf{x}}_t))$ is redefined as

$$F(\tilde{\gamma}'_i \tilde{\mathbf{x}}_t) = \frac{1}{1 + \exp(\gamma(-\tilde{\gamma}'_i \tilde{\mathbf{x}}_t))} - \frac{1}{2}. \quad (15)$$

Subtracting one-half from the logistic function is useful just in deriving linearity tests where it simplifies notation but not affect the argument. The models estimated in this paper do not contain that term.

Consider (14) with (15) and the testing of the hypothesis that y_t is a linear process, i. e. $y_t = \boldsymbol{\alpha}' \mathbf{z}_t + \varepsilon_t$, assuming that it is stationary. The null hypothesis may be defined as $H_0 : \boldsymbol{\lambda}_i = \mathbf{0}, i = 1, \dots, h$. Note also that $F(0) = 0$. This implies another possible null hypothesis of linearity

$$H_0 : \gamma_i = 0, i = 1, \dots, h. \quad (16)$$

Hypothesis (16) offers a convenient starting point for studying the linearity problem in the LM (score) testing framework. First, consider $h = 1$. Equation (7) becomes

$$y_t = \boldsymbol{\alpha}' \mathbf{z}_t + \boldsymbol{\lambda}' \mathbf{z}_t F(\gamma(\tilde{\gamma}' \tilde{\mathbf{x}}_t)) + \varepsilon_t. \quad (17)$$

Note that model (17) is only identified under the alternative $\gamma \neq 0$. A consequence of this complication is that the standard asymptotic distribution theory for the likelihood ratio or other classical test statistics for testing (16) is not available. Davies (1977) and Davies (1987) first discussed solutions to this problem. Following Saikkonen and Luukkonen (1988), Luukkonen et al. (1988), and Teräsvirta, Lin and Granger (1993) we solve the problem by replacing $F(\gamma(\tilde{\gamma}' \tilde{\mathbf{x}}_t))$ by a low order Taylor expansion approximation about $\gamma = 0$. Consider a first-order Taylor expansion of (15)

$$t_1 = F(0) + \left. \frac{\partial F}{\partial \gamma} \right|_{\gamma=0} \gamma + R_1(\tilde{\mathbf{x}}_t; \gamma, \tilde{\gamma}) = \frac{1}{4} \gamma \left(\sum_{i=0}^q \tilde{\gamma}_i \tilde{x}_{i,t} \right) + R_1(\tilde{\mathbf{x}}_t; \gamma, \tilde{\gamma}), \quad (18)$$

where $R_1(\tilde{\mathbf{x}}_t; \gamma, \tilde{\gamma})$ is the remainder of the expansion. Replacing (15) by (18) in (17) we get

$$y_t = \boldsymbol{\alpha}' \mathbf{z}_t + \frac{1}{4} \gamma \boldsymbol{\lambda}' \mathbf{z}_t \left(\sum_{i=0}^q \tilde{\gamma}_i \tilde{x}_{i,t} \right) + \varepsilon_t^*, \quad (19)$$

where $\varepsilon_t^* = \varepsilon_t + \boldsymbol{\lambda}' \mathbf{z}_t R_1(\tilde{\mathbf{x}}_t; \gamma, \tilde{\gamma})$. Rearranging terms, (19) becomes

$$y_t = \boldsymbol{\pi}' \mathbf{z}_t + \boldsymbol{\xi}' \mathbf{x}_t^{(2)} + \sum_{i=1}^{q_1} \sum_{j=i}^{q_1} \theta_{ij} z_{i,t}^{(1)} x_{j,t}^{(1)} + \sum_{i=1}^{p-q_1} \sum_{j=1}^{q_1} \beta_{ij} z_{i,t}^{(2)} x_{j,t}^{(1)} + \sum_{i=1}^p \sum_{j=1}^{q_2} \rho_{ij} z_{i,t} x_{j,t}^{(2)} + \varepsilon_t^*. \quad (20)$$

Using (20) instead of (17) circumvents the identification problem, and we obtain a simple test of linearity. The null hypothesis can be defined as $H_0 : \boldsymbol{\xi} = \mathbf{0}, \theta_{ij} = 0, \beta_{ij} = 0, \rho_{ij} = 0$. However, the parameters θ_{ij}, β_{ij} , and ρ_{ij} do not depend on λ_0 . Thus when the only nonlinear element in (17) is the intercept and $\mathbf{x}_t^{(2)} = \mathbf{0}$ the test has no power. To remedy this situation Luukkonen et al. (1988) suggested a third-order Taylor approximation of the transition function, expressed as

$$t_3 = \frac{1}{4} \gamma \left(\sum_{i=0}^q \tilde{\gamma}_i \tilde{x}_{i,t} \right) + \frac{1}{96} \gamma^3 \left(\sum_{i=0}^q \sum_{j=0}^q \sum_{k=0}^q \tilde{\gamma}_i \tilde{\gamma}_j \tilde{\gamma}_k \tilde{x}_{i,t} \tilde{x}_{j,t} \tilde{x}_{k,t} \right) + R_3(\mathbf{x}_t; \gamma, \boldsymbol{\omega}, c). \quad (21)$$

Replacing (15) by (21) in (17) and setting $\mathbf{x}_t^{(2)} = \mathbf{0}$ we get

$$\begin{aligned} y_t = & \boldsymbol{\pi}' \mathbf{z}_t + \sum_{i=1}^{q_1} \sum_{j=i}^{q_1} \theta_{ij} z_{i,t}^{(1)} x_{j,t}^{(1)} + \sum_{i=1}^{p-q_1} \sum_{j=1}^{q_1} \beta_{ij} z_{i,t}^{(2)} x_{j,t}^{(1)} + \sum_{i=1}^{q_1} \sum_{j=i}^{q_1} \sum_{k=j}^{q_1} \theta_{ijk} z_{i,t}^{(1)} x_{j,t}^{(1)} x_{k,t}^{(1)} \\ & + \sum_{i=1}^{p-q_1} \sum_{j=1}^{q_1} \sum_{k=j}^{q_1} \sum_{l=k}^{q_1} \beta_{ijkl} z_{i,t}^{(2)} x_{j,t}^{(1)} x_{k,t}^{(1)} x_{l,t}^{(1)} \\ & + \sum_{i=1}^{p-q_1} \sum_{j=1}^{q_1} \sum_{k=j}^{q_1} \sum_{l=k}^{q_1} \beta_{ijkl} z_{i,t}^{(2)} x_{j,t}^{(1)} x_{k,t}^{(1)} x_{l,t}^{(1)} + \varepsilon_t^*. \end{aligned} \quad (22)$$

The null hypothesis is defined as $H_0 : \theta_{ij} = 0, \beta_{ij} = 0, \theta_{ijk} = 0, \beta_{ijk} = 0, \theta_{ijkl} = 0$, and $\beta_{ijkl} = 0$.

Now we can use (20) or (22) to test linearity. Note that $\varepsilon_t^* = \varepsilon_t$ when the null hypothesis is true. Under H_0 the standard Lagrange multiplier (LM) or score type test statistic has an asymptotic χ^2 distribution with $m + q_2$ degrees of freedom when the null hypothesis holds, where m is the number of nonlinear regressors in (20) or (22). The asymptotic theory requires that the linear autoregressive (null) model is stationary and ergodic. We define the residuals estimated under the null hypothesis as $\hat{\varepsilon}_t = y_t - \hat{\boldsymbol{\pi}}' \mathbf{z}_t$.

The test can be carried out in stages as follows:

1. Regress y_t on \mathbf{z}_t and compute $SSR_0 = \left(\frac{1}{T}\right) \sum_{t=1}^T \hat{\varepsilon}_t^2$.
2. Regress $\hat{\varepsilon}_t$ on $\mathbf{z}_t, \mathbf{x}_t^{(2)}$, and on the m nonlinear regressors of (20) or (22). Compute the residual sum of

squares $SSR_1 = \left(\frac{1}{T}\right) \sum_{t=1}^T \hat{v}_t^2$.

3. Compute the χ^2 statistic

$$LM_{\chi^2}^l = T \frac{SSR_0 - SSR_1}{SSR_0}, \quad (23)$$

or the F version of the test

$$LM_F^l = \frac{(SSR_0 - SSR_1)/(m + q_2)}{SSR_1/(T - p - 1 - m - q_2)}, \quad (24)$$

where T is the number of observations. Note that if we use (22) to test for linearity then $q_2 = 0$.

When \mathbf{z}_t and (or) $\mathbf{x}_t^{(2)}$ have a large number of elements, the number of auxiliary null hypothesis will sometimes be large compared to the sample size. In that case the asymptotic χ^2 distribution is likely to be a poor approximation to the actual small sample distribution. It has been found (see Granger and Teräsvirta (1993, Chapter 7)) that an F-approximation works much better. Another possibility to improve the power of the test is to follow the idea of Anders and Korn (1999) and replace the variables present only under the alternative hypothesis by their most important principal components. The number of principal components to use can be chosen such that a high proportion of the total variance is explained. Using the principal components not only reduces the number of summands, but also remove multicollinearity amongst the regressors. Luukkonen et al. (1988) suggested to augment the first-order Taylor expansion only by the terms that are functions of λ_0 , and this is called the “economy version” of the test. In the present framework, this means removing the fourth order terms in (22).

3.3 Determining the Number of Hidden Neurons

In a practical situation we want to be able to test for the number of hidden units of the neural network. This can be done combining the ideas of the neural network test of Teräsvirta et al. (1993), the test of remaining nonlinearity of Eitheim and Teräsvirta (1996) and the results in Teräsvirta and Lin (1993). The basic idea is to start using the test of Section 3.2 and test the linear model against the nonlinear alternative with only one hidden neuron. If the null hypothesis is rejected, then fit the model with one hidden unit and test for the second one. Proceed in that way until the first acceptance of the null hypothesis. The individual tests are based on

linearizing the nonlinear contribution of the additional hidden neuron. Consider first the simplest case in which the model contains one hidden unit, and we want to know whether an additional unit is required or not. Write the model as

$$y_t = \boldsymbol{\alpha}' \mathbf{z}_t + \boldsymbol{\lambda}'_1 \mathbf{z}_t F(\gamma_1(\tilde{\boldsymbol{\gamma}}'_1 \tilde{\mathbf{x}}_t)) + \boldsymbol{\lambda}'_2 \mathbf{z}_t F(\gamma_2(\tilde{\boldsymbol{\gamma}}'_2 \tilde{\mathbf{x}}_t)) + \varepsilon_t. \quad (25)$$

If we want to test for the second hidden unit in (25), an appropriate null hypothesis is

$$H_0 : \gamma_2 = 0, \quad (26)$$

whereas the alternative is $H_1 : \gamma_2 \neq 0$. We assume that under this null hypothesis the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}_1$, $\boldsymbol{\omega}_1$, and β can be consistently estimated and that the estimators are asymptotically normal. Note that (25) is only identified under the alternative. We may solve this problem in the same fashion we did in Section 3.2, using a low order Taylor expansion of $F(\gamma_2(\tilde{\boldsymbol{\gamma}}'_2 \tilde{\mathbf{x}}_t))$ about $\gamma_2 = 0$. Using a first order expansion and after rearranging terms, the resulting model is

$$\begin{aligned} y_t = & \boldsymbol{\pi}' \mathbf{z}_t + \boldsymbol{\xi}' \mathbf{x}_t^{(2)} + \boldsymbol{\lambda}'_1 \mathbf{z}_t F(\gamma_1(\tilde{\boldsymbol{\gamma}}'_1 \tilde{\mathbf{x}}_t)) \\ & + \sum_{i=1}^{q_1} \sum_{j=i}^{q_1} \theta_{ij} z_{i,t}^{(1)} x_{j,t}^{(1)} + \sum_{i=1}^{p-q_1} \sum_{j=1}^{q_1} \beta_{ij} z_{i,t}^{(2)} x_{j,t}^{(1)} + \sum_{i=1}^p \sum_{j=1}^{q_2} \rho_{ij} z_{i,t} x_{j,t}^{(2)} + \varepsilon_t^*. \end{aligned} \quad (27)$$

The null hypothesis is defined as $H_0 : \boldsymbol{\xi} = \mathbf{0}, \theta_{ij} = 0, \beta_{ij} = 0, \rho_{ij} = 0$. We define the residuals estimated under the null hypothesis as $\hat{\varepsilon}_t = y_t - \hat{\boldsymbol{\pi}}' \mathbf{z}_t - \hat{\boldsymbol{\lambda}}'_1 \mathbf{z}_t \hat{F}(\gamma_1(\tilde{\boldsymbol{\gamma}}'_1 \tilde{\mathbf{x}}_t))$.

The local approximation to the normal log likelihood function in a neighborhood of H_0 for observation t and ignoring the remainder is

$$\begin{aligned} l_t = & -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \left\{ y_t - \boldsymbol{\pi}' \mathbf{z}_t - \boldsymbol{\xi}' \mathbf{x}_t^{(2)} - \boldsymbol{\lambda}'_1 \mathbf{z}_t F(\gamma_1(\tilde{\boldsymbol{\gamma}}'_1 \tilde{\mathbf{x}}_t)) \right. \\ & \left. - \sum_{i=1}^{q_1} \sum_{j=i}^{q_1} \theta_{ij} z_{i,t}^{(1)} x_{j,t}^{(1)} - \sum_{i=1}^{p-q_1} \sum_{j=1}^{q_1} \beta_{ij} z_{i,t}^{(2)} x_{j,t}^{(1)} - \sum_{i=1}^p \sum_{j=1}^{q_2} \rho_{ij} z_{i,t} x_{j,t}^{(2)} \right\}^2. \end{aligned} \quad (28)$$

The LM statistic, assuming σ^2 constant, is

$$LM = \frac{1}{\hat{\sigma}^2} \sum_{t=1}^T \hat{\varepsilon}_t \hat{\boldsymbol{\nu}}_t' \left\{ \sum_{t=1}^T \hat{\boldsymbol{\nu}}_t \hat{\boldsymbol{\nu}}_t' - \sum_{t=1}^T \hat{\boldsymbol{\nu}}_t \hat{\mathbf{h}}_t' \left(\sum_{t=1}^T \hat{\mathbf{h}}_t \hat{\mathbf{h}}_t' \right)^{-1} \sum_{t=1}^T \hat{\mathbf{h}}_t \hat{\boldsymbol{\nu}}_t' \right\} \sum_{t=1}^T \hat{\boldsymbol{\nu}}_t \hat{\varepsilon}_t, \quad (29)$$

where

$$\begin{aligned} \hat{\mathbf{h}}_t &= \nabla G(\mathbf{z}_t, \mathbf{x}_t; \hat{\boldsymbol{\Psi}}) \\ &= \left[\mathbf{z}_t', \mathbf{z}_t' \hat{F}(\boldsymbol{\omega}'_1 \mathbf{x}_t - \beta_1), \hat{\boldsymbol{\lambda}}_1' \mathbf{z}_t \frac{\partial \hat{F}(\boldsymbol{\omega}'_1 \mathbf{x}_t - \beta_1)}{\partial \boldsymbol{\omega}'_1}, \hat{\boldsymbol{\lambda}}_1' \mathbf{z}_t \frac{\partial \hat{F}(\boldsymbol{\omega}'_1 \mathbf{x}_t - \beta_1)}{\partial \beta_1} \right]', \end{aligned}$$

and the vector $\hat{\boldsymbol{\nu}}_t$ is formed by $\mathbf{x}_t^{(2)}$ and the products $z_{i,t}^{(1)} x_{j,t}^{(1)}$, $i = 1, \dots, q_1$; $j = 1, \dots, q_1$, $z_{i,t}^{(2)} x_{j,t}^{(1)}$, $i = 1, \dots, p - q_1$; $j = 1, \dots, q_1$, and $z_{i,t} x_{j,t}^{(2)}$, $i = 1, \dots, p$; $j = 1, \dots, q_2$.

The test can be carried out in stages as follows:

1. Estimate model (7) with only one hidden neuron. If the sample size is small and the model is difficult to estimate, then numerical problems in applying the nonlinear least squares routine may lead to a solution such that the residual vector is not precisely orthogonal to the gradient matrix of $G(\mathbf{z}_t, \mathbf{x}_t; \hat{\boldsymbol{\Psi}})$. This has an adverse effect on the empirical size of the test. To circumvent this problem, we regress the residuals $\hat{\varepsilon}_t$ on $\nabla G(\mathbf{z}_t, \mathbf{x}_t; \hat{\boldsymbol{\Psi}})$, and compute the residual sum of squares $SSR_0 = (\frac{1}{T}) \sum_{t=1}^T \tilde{\varepsilon}_t^2$.
2. Regress $\tilde{\varepsilon}_t$ on $\hat{\mathbf{h}}_t$ and $\hat{\boldsymbol{\nu}}_t$. Compute the residual sum of squares $SSR_1 = (\frac{1}{T}) \sum_{t=1}^T \hat{v}_t^2$.
3. Compute the χ^2 statistic

$$LM_{\chi^2}^{hn} = T \frac{SSR_0 - SSR_1}{SSR_0}, \quad (30)$$

or the F version of the test

$$LM_F^{hn} = \frac{(SSR_0 - SSR_1)/(m + q_2)}{SSR_1/(T - n - m - q_2)}, \quad (31)$$

where m and n are, respectively, the number of elements of $\hat{\boldsymbol{\nu}}_t$ and $\hat{\mathbf{h}}_t$.

Under H_0 , $LM_{\chi^2}^{hn}$ is approximately distributed as a χ^2 with $m + q_2$ degrees of freedom and LM_F^{hn} has approximately an F distribution with $m + q_2$ and $T - n - m - q_2$ degrees of freedom.

When applying the test a special care should be taken. If $\hat{\gamma}_1$ is very large, we may have some numerical problems when carrying out the test in small samples. A solution is to omit $\hat{\lambda}'_1 \mathbf{z}_t \frac{\partial \hat{F}(\omega'_1 \mathbf{x}_t - \beta_1)}{\partial \omega'_1}$ and $\hat{\lambda}'_1 \mathbf{z}_t \frac{\partial \hat{F}(\omega'_1 \mathbf{x}_t - \beta_1)}{\partial \beta_1}$ from the regression in step 2. This can be done without significantly affecting the value of the test statistic.

Note that the same comments about the power of the linearity test of the previous section apply here and a test using a third-order Taylor expansion can be developed using the same arguments.

4 Estimation Procedures and Parameter Inference

After specifying the model, the parameters should be estimated by nonlinear least squares (NLS) or maximum likelihood (ML). In the case where $\varepsilon_t \sim \text{NID}(0, \sigma^2)$, both methods are equivalent. Hence the parameter vector Ψ of (7) is estimated as

$$\hat{\Psi} = \underset{\Psi}{\operatorname{argmin}} Q_T(\Psi) = \underset{\Psi}{\operatorname{argmin}} \sum_{t=1}^T (y_t - G(\mathbf{z}_t, \mathbf{x}_t; \Psi))^2. \quad (32)$$

Under some regularity conditions the estimates are consistent and asymptotically normal, that is,

$$\sqrt{T} \left(\hat{\Psi} - \Psi^* \right) \rightarrow \text{N}(0, C), \quad (33)$$

where Ψ^* is the true parameter vector and C is the covariance matrix of the estimates. Following Davidson and MacKinnon (1993, Chapter 5), C can be consistently estimated as

$$C = \hat{\sigma}^2 \left(\hat{\mathbf{H}}' \hat{\mathbf{H}} \right)^{-1}, \quad (34)$$

where $\hat{\sigma}^2$ is the estimated variance of the residuals and $\hat{\mathbf{H}}$ is a matrix with T rows given by $\nabla G(\mathbf{z}_t, \mathbf{x}_t; \hat{\Psi})$.

The estimation of the parameters is not easy, and in general the optimization algorithm is very sensitive to the choice of the starting values of the parameters. The use of algorithms like the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm or the Levenberg-Marquardt are strongly recommended. See Bertsekas (1995) for details about the optimization algorithms. Another important question that should be addressed is the choice of the linear search procedure to select the size of the step. Cubic or quadratic interpolation are usually a good choice. All the models in this paper are estimated with the Levenberg-Marquardt algorithm with cubic inter-

polation linear search. Another possibility is to use constrained optimization techniques, such the Sequential Quadratic Programming (SQP) algorithm and impose the identification restrictions. However, by our own experience, using the SQP algorithm turns the estimation process rather slow and does not affect the quality of the solution.

The estimation procedure is carried together with the test for the number of hidden units. First we test for linearity against a model given by (7) with $h = 1$. If linearity is rejected we estimate the parameters of the nonlinear model and test for the second hidden unit. If we reject the null hypothesis, we use the estimated values for the first hidden unit as starting values and use the procedure below to compute initial values for the second hidden unit. We proceed in that way until the first acceptance of the null hypothesis.

Concerning the selection of the starting values, we propose the following algorithm.

Rewrite model (7) as

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad (35)$$

where $\mathbf{y}' = [y_1, y_2, \dots, y_T]$, $\boldsymbol{\varepsilon}' = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T]$, $\boldsymbol{\theta}' = [\boldsymbol{\alpha}', \boldsymbol{\lambda}'_1, \dots, \boldsymbol{\lambda}'_h]$,

$$\mathbf{Z} = \begin{pmatrix} \mathbf{z}'_1 & \tilde{\mathbf{z}}'_{1,1} & \dots & \tilde{\mathbf{z}}'_{h,1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{z}'_T & \tilde{\mathbf{z}}'_{1,T} & \dots & \tilde{\mathbf{z}}'_{h,T} \end{pmatrix},$$

and $\tilde{\mathbf{z}}_{i,t} = F(\gamma_i(\tilde{\boldsymbol{\omega}}'_i \mathbf{x}_t - c_i))\mathbf{z}_t$, $i = 1, \dots, h$.

Once γ_i , $\tilde{\boldsymbol{\omega}}_i$, and c_i have been determined, the parameter vector $\boldsymbol{\theta}$ can be estimated by

$$\hat{\boldsymbol{\theta}} = (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{y}. \quad (36)$$

The selection of the parameters γ_h , $\tilde{\boldsymbol{\omega}}_h$, and c_h of the h^{th} hidden unit is divided in the following steps:

1. Draw K possible values for $\tilde{\boldsymbol{\omega}}_h$ such as $\tilde{\omega}_{1h} \in (0, 1)$ and $\tilde{\omega}_{jh} \in [-1, 1]$, $j = 2, \dots, q$, and call them \mathbf{v}_k , $k = 1, \dots, K$. We recommend to choose the values for $\tilde{\omega}_{1h}$ from a uniform random distribution over the interval $(0, 1]$ and for $\tilde{\omega}_{jh}$, $j = 2, \dots, q$, from a uniform random distribution over the interval $[-1, 1]$.
2. For $k = 1, \dots, K$:

- (a) Normalize each vector \mathbf{v}_k and compute the projection $\mathbf{x}_{/\mathbf{v}_k}$ of \mathbf{x} ($\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_T]$) in the direction of \mathbf{v}_k .
 - (b) Compute the median of $\mathbf{x}_{/\mathbf{v}_k}$ and call it b_k .
 - (c) Draw a grid of N possible values for the slope and call them $\gamma_{h,n}$, $n = 1, \dots, N$.
3. For $k = 1, \dots, K$ and $n = 1, \dots, N$, set $\gamma_h = \gamma_{h,n}$, $\tilde{\omega}_h = \mathbf{v}_k$, and $c_h = b_k$ and compute the value of $Q_T(\Psi)$. Choose the values of the parameters so that they minimize the objective function and call them γ^* , \mathbf{v}^* , and b^* .
 4. Set $\gamma_h = \gamma^*$, $\tilde{\omega}_h = \mathbf{v}_k^*$, and $c_h = b^*$ and use them as the starting-values.

After selecting the starting values of the h^{th} hidden unit we have to reorder the units.

Concerning the slope parameter, we should stress that it is very difficult to have a precise estimate of γ_i , $i = 1, \dots, h$. One of the reasons is that for large γ_i , the derivatives of the transition function, as already mentioned in Section 3.3, approach to degenerate functions. Hence to obtain an accurate estimate of γ_i one needs a large number of observations in the neighborhood of c_i . In general we have only few observations near c_i and rather imprecise estimates of the slope parameter, causing that the parameters of the logistic function to have t -statistics very close to zero. In that sense, the model builder should thus not automatically take a low absolute value of the t -statistic of the parameters of the transition function as an evidence against the estimated nonlinear model.

5 Monte-Carlo Experiment

In this section we report the results of a simulation study designed to find out the behaviour of the proposed tests and the variable selection procedure. We simulated the following models, discarding the first 500 observations to avoid any initialization effects.

- Model 1:

$$y_t = 0.8 - 0.5y_{t-1} + 0.3y_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 1^2). \quad (37)$$

- Model 2:

$$y_t = 1.8y_{t-1} - 1.06y_{t-2} + (0.02 - 0.90y_{t-1} + 0.795y_{t-2})F(20(y_{t-1} - 0.02)) + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 0.02^2), \quad (38)$$

- Model 3:

$$y_t = -0.1 + 0.3y_{t-1} + 0.2y_{t-2} + (-1.2y_{t-1} + 0.5y_{t-2})F(20(y_{t-1} + 0.6)) + (1.8y_{t-1} - 1.2y_{t-2})F(20(y_{t-1} - 0.6)) + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 0.5^2). \quad (39)$$

- Model 4:

$$y_t = 0.5 + 0.8y_{t-1} - 0.2y_{t-2} + (-0.5 - 1.2y_{t-1} + 0.8y_{t-2})F(11.31(0.7071y_{t-1} - 0.7071y_{t-2} - 0.1414)) + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 0.5^2). \quad (40)$$

- Model 5:

$$y_t = 0.5 + 0.8y_{t-1} - 0.2y_{t-2} + (1.5 + 0.6y_{t-1} - 0.3y_{t-2})F(8.49(0.7071y_{t-1} - 0.7071y_{t-2} + 1.0607)) + (-0.5 - 1.2y_{t-1} + 0.7y_{t-2})F(8.49(0.7071y_{t-1} - 0.7071y_{t-2} - 1.0607)) + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 1^2). \quad (41)$$

5.1 Estimation Algorithm

To evaluate the performance of the estimation algorithm in small samples, we simulated 1000 replications of models (38)–(41) each of which with 100 and 500 observations. We estimated the parameters for each replication, with \mathbf{z}_t and \mathbf{x}_t correctly specified. Table (1) shows the median and the median absolute deviation (MAD) of the estimates, defined as

$$\text{MAD}(\hat{\Psi}) = \text{median}(|\hat{\Psi} - \text{median}(\hat{\Psi})|). \quad (42)$$

The true value of the parameters are shown between parentheses.

Reporting the median and MAD was suggested by van Dijk (1999) and can be interpreted as measures that are robust to outliers.

In small samples, the discrepancies between the estimates and their true values are small, except for the case of slope parameter, and when we increase the sample size we obtain rather precise estimates.

5.2 Model Selection Tests

5.2.1 Variable Selection

Tables 2 and 3 show, respectively, the results of the variable selection procedure using a third order Taylor expansion in (13) and using only the linear term (no cross-products) in (13). The selection was made among the first five lags of y_t . We report only the results concerning the nonlinear models. The column C indicates the relative frequency of correctly selecting the elements of \mathbf{z}_t . The columns U and O indicate, respectively, the relative frequency of underfitting and overfitting the dimension of \mathbf{z}_t .

Observing Table 2, we can see that the SBIC outperforms the AIC in most of the cases. With a sample size of 500 observations the SBIC always find the correct set of variables, and in small samples the SBIC has a satisfactory performance with models (38) and (41), but underfits models (39) and (40) in more than 50% of the replications. As we expected, the algorithm works better when we use the third-order Taylor expansion than in the linear case (Table 3). Further simulation results can be found in Rech et al. (1999).

5.2.2 Linearity Tests

Concerning the size of the linearity test developed in Section 3.2, hereafter LM_F^l and its “economy version”, $LM_F^{l,e}$, we show the plot of the deviation of empirical size from the nominal size versus the nominal size. The results are shown in Figure 2. The results are based on 1000 replications of model (37). Observing the plots we can see that the size is acceptable and the distortions seem smaller at low levels of significance.

In power simulations of the linearity test the data were generated from models (38)–(41). The results are shown in Figures 3–6.

In both size and power simulations we assume that \mathbf{z}_t is correctly specified. In power simulations we also tested the ability of the linearity test to identify the correct set of elements of \mathbf{x}_t . We expect that when \mathbf{x}_t is correctly defined, the power increases.

Table 1: Median and MAD of the NLS estimates of the parameters. True values between parentheses

Parameter	100 observations							
	Model 2		Model 3		Model 4		Model 5	
	Median	MAD	Median	MAD	Median	MAD	Median	MAD
$\hat{\alpha}_0$	0.0045 (0)	0.0042	-0.1394 (-0.1)	0.5738	0.5125 (0.5)	0.1417	0.6990 (0.5)	1.3776
$\hat{\alpha}_1$	1.6019 (1.8)	0.2054	0.2895 (0.3)	0.4362	0.8309 (0.8)	0.1549	0.8099 (0.8)	0.5729
$\hat{\alpha}_2$	-0.9548 (-1.06)	0.1932	0.1513 (0.2)	0.1946	-0.2301 (-0.2)	0.1438	-0.2367 (-0.2)	0.4140
$\hat{\lambda}_{01}$	0.0184 (0.02)	0.0335	0.0616 (0)	0.8141	0.6066 (-0.5)	0.4255	1.7964 (1.5)	2.9473
$\hat{\lambda}_{02}$	-	-	-0.0530 (0)	0.8242	-	-	-0.3629 (-0.5)	2.4208
$\hat{\lambda}_{11}$	-0.5937 (-0.9)	0.3618	-0.9468 (-1.2)	0.7731	-1.1014 (-1.2)	0.3467	-0.2052 (0.6)	1.5347
$\hat{\lambda}_{12}$	-	-	1.8792 (1.8)	0.8985	-	-	-0.7638 (-1.2)	1.3172
$\hat{\lambda}_{21}$	0.6167 (0.795)	0.3111	0.7014 (0.5)	0.3124	0.6957 (0.8)	0.3423	0.3573 (-0.3)	1.4217
$\hat{\lambda}_{22}$	-	-	-1.3381 (-1.2)	0.3015	-	-	0.2850 (0.7)	1.1932
$\hat{\gamma}_1$	106.9324 (20)	99.2520	20.1428 (20)	18.1140	19.4749 (11.31)	15.9591	3.5715 (8.49)	2.5729
$\hat{\gamma}_2$	-	-	29.4483 (20)	25.6801	-	-	8.2832 (8.49)	6.4536
$\hat{\omega}_{11}$	-	-	-	-	0.7310 (0.7071)	0.0906	0.7193 (0.7071)	0.0531
$\hat{\omega}_{21}$	-	-	-	-	-	-	0.7160 (0.7071)	0.0283
$\hat{\omega}_{12}$	-	-	-	-	-0.6829 (-0.7071)	0.0956	-0.6938 (-0.7071)	0.0553
$\hat{\omega}_{22}$	-	-	-	-	-	-	-0.6955 (-0.7071)	0.0289
\hat{e}_1	0.0236 (0.02)	0.0344	-0.5578 (-0.6)	0.1869	0.1422 (0.1414)	0.1261	-0.3252 (-1.0607)	1.0763
\hat{e}_2	-	-	0.5853 (0.6)	0.0785	-	-	1.0971 (-1.0607)	0.3120

Parameter	500 observations							
	Model 2		Model 3		Model 4		Model 5	
	Median	MAD	Median	MAD	Median	MAD	Median	MAD
$\hat{\alpha}_0$	0.0025 (0)	0.0085	-0.1295 (-0.1)	0.1541	0.5041 (0.5)	0.0521	0.4597 (0.5)	0.3095
$\hat{\alpha}_1$	1.7070 (1.8)	0.1204	0.2739 (0.3)	0.1307	0.8063 (0.8)	0.0590	0.7868 (0.8)	0.1224
$\hat{\alpha}_2$	-1.0429 (-1.06)	0.1630	0.1860 (0.2)	0.0519	-0.2029 (-0.2)	0.0536	-0.1877 (-0.2)	0.0948
$\hat{\lambda}_{01}$	0.0163 (0.02)	0.0226	0.0276 (0)	0.1551	-0.4904 (-0.5)	0.1618	1.5557 (1.5)	0.3163
$\hat{\lambda}_{02}$	-	-	0.0184 (0)	0.1616	-	-	-0.5596 (-0.5)	0.5317
$\hat{\lambda}_{11}$	-0.7601 (-0.9)	0.2636	-1.1898 (-1.2)	0.1601	-1.1852 (-1.2)	0.1269	0.5797 (0.6)	0.1813
$\hat{\lambda}_{12}$	-	-	1.7880 (1.8)	0.1467	-	-	-1.1959 (-1.2)	0.2117
$\hat{\lambda}_{21}$	0.7817 (0.795)	0.3248	0.5084 (0.5)	0.0651	0.7965 (0.8)	0.1361	-0.2908 (-0.3)	0.1626
$\hat{\lambda}_{22}$	-	-	-1.2083 (-1.2)	0.0657	-	-	0.6937 (0.7)	0.2184
$\hat{\gamma}_1$	25.4119 (20)	15.6414	22.7696 (20)	11.4805	13.2183 (11.31)	5.6405	8.8183 (8.49)	4.3173
$\hat{\gamma}_2$	-	-	21.2108 (20)	6.6001	-	-	8.5442 (8.49)	1.5223
$\hat{\omega}_{11}$	-	-	-	-	0.7162 (0.7071)	0.0335	0.7103 (0.7071)	0.0134
$\hat{\omega}_{21}$	-	-	-	-	-	-	0.7074 (0.7071)	0.0036
$\hat{\omega}_{12}$	-	-	-	-	-0.6979 (-0.7071)	0.0338	-0.7039 (-0.7071)	0.0133
$\hat{\omega}_{22}$	-	-	-	-	-	-	-0.7068 (-0.7071)	0.0036
\hat{e}_1	0.0202 (0.02)	0.0289	-0.6038 (-0.6)	0.0285	0.1469 (0.1414)	0.0433	-1.0307 (-1.0607)	0.1387
\hat{e}_2	-	-	0.6025 (0.6)	0.0166	-	-	1.0635 (1.0607)	0.0480

Table 2: Relative frequency of selecting correctly the variables of the model at sample sizes 100 and 500 observations based on 1000 replications among the first 5 lags and using a third order Taylor expansion.

Model	100 observations					
	C		U		O	
	SBIC	AIC	SBIC	AIC	SBIC	AIC
2	0.9280	0.6630	0.0190	0	0.0530	0.3370
3	0.4670	0.5520	0.5180	0.0330	0.0150	0.4150
4	0.4760	0.6130	0.5050	0.0110	0.0190	0.3760
5	0.9980	0.5700	0	0	0.0020	0.4300

Model	500 observations					
	C		U		O	
	SBIC	AIC	SBIC	AIC	SBIC	AIC
2	1	0.9110	0	0	0	0.0890
3	1	0.7450	0	0	0	0.2550
4	1	0.8060	0	0	0	0.1940
5	1	0.5960	0	0	0	0.4040

Table 3: Relative frequency of selecting correctly the variables of the model at sample sizes 100 and 500 observations based on 1000 replications among the first 5 lags and no cross-products of the regressors.

Model	100 observations					
	C		U		O	
	SBIC	AIC	SBIC	AIC	SBIC	AIC
2	0.8380	0.5630	0	0	0.1620	0.4370
3	0.3050	0.3640	0.4620	0.1270	0.2330	0.5090
4	0.0070	0.0360	0.7790	0.4850	0.2140	0.4790
5	0.1900	0.3460	0.6590	0.2440	0.1510	0.4100

Model	500 observations					
	C		U		O	
	SBIC	AIC	SBIC	AIC	SBIC	AIC
2	0.9400	0.5970	0	0	0.0600	0.4030
3	0.7810	0.3510	0.0010	0	0.2180	0.6490
4	0.0280	0.1090	0.7860	0.2770	0.1860	0.6140
5	0.7270	0.3450	0.1260	0	0.1470	0.6550

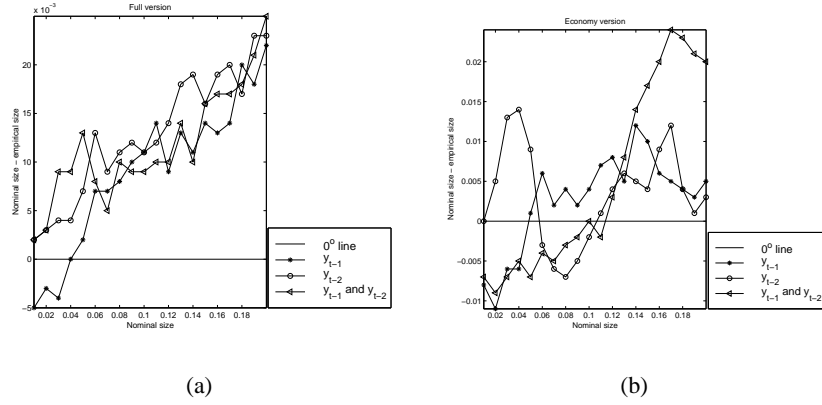


Figure 2: Discrepancy between the empirical and the nominal sizes of the linearity tests at sample size of 100 observations based on 1000 replications of model (37). Panel (a) refers to the LM_F^l test. Panel (b) refers to the $LM_F^{l,e}$ test.

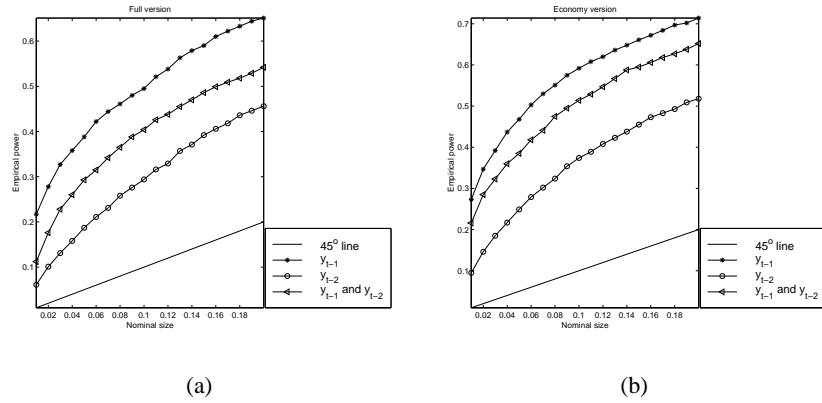


Figure 3: Power-size curve of the linearity tests at sample size of 100 observations based on 1000 replications of model (38). Panel (a) refers to the LM_F^l test. Panel (b) refers to the $LM_F^{l,e}$ test.

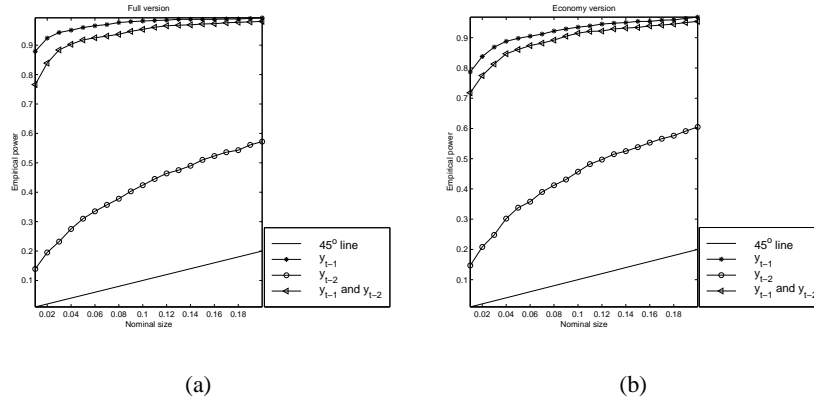


Figure 4: Power-size curve of the linearity tests at sample size of 100 observations based on 1000 replications of model (39). Panel (a) refers to the LM_F^l test. Panel (b) refers to the $LM_F^{l,e}$ test.

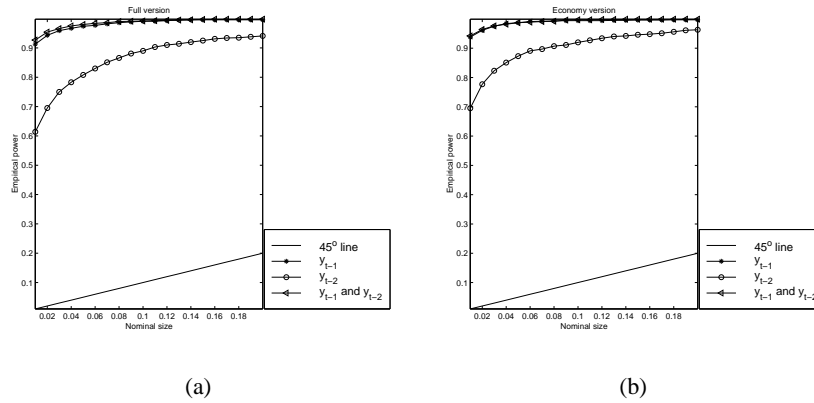


Figure 5: Power-size curve of the linearity tests at sample size of 100 observations based on 1000 replications of model (40). Panel (a) refers to the LM_F^l test. Panel (b) refers to the $LM_F^{l,e}$ test.

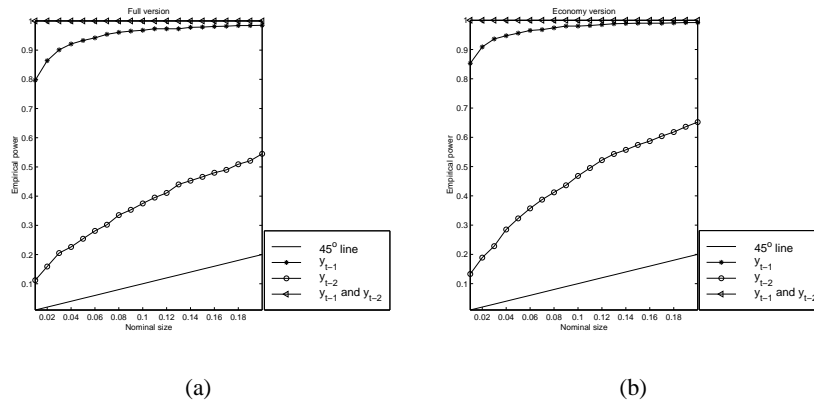


Figure 6: Power-size curve of the linearity tests at sample size of 100 observations based on 1000 replications of model (41). Panel (a) refers to the LM_F^l test. Panel (b) refers to the $LM_F^{l,e}$ test.

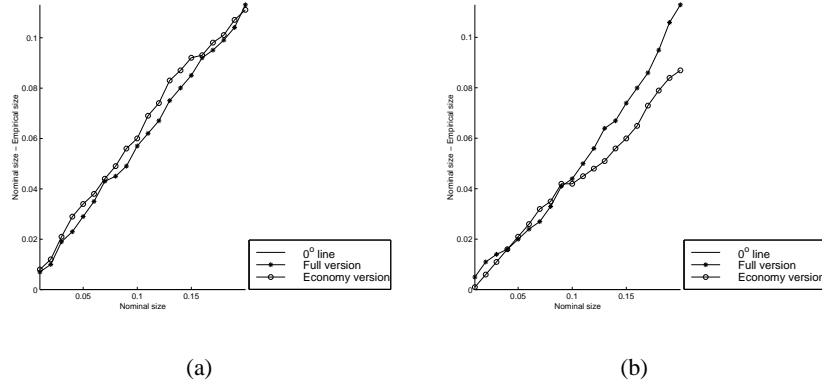


Figure 7: Discrepancy between the empirical and the nominal sizes of the additional hidden unit tests at sample size of 100 observations based on 1000 replications of model (37) and (37). Panel (a) refers to model (38). Panel (b) refers to model (40).

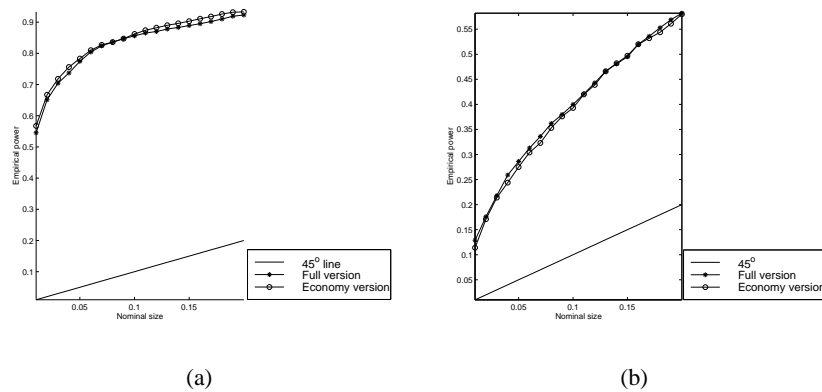


Figure 8: Power-size curve of the additional hidden unit tests at sample size of 100 observations based on 1000 replications of model (39) and (41). Panel (a) refers to model (39). Panel (b) refers to model (41).

In Figures 3–4 we can observe that the power of the test improves when we select y_{t-1} as the transition variable and in Figure 5 the power increases when we use y_{t-1} and y_{t-2} as transition variables. With model (41) the power is always 1 when the transition variable is correctly chosen.

5.2.3 Tests for the Number of Hidden Units

To study the behaviour of the tests for the number of hidden neurons we simulated 1000 replications of models (38)–(41) at sample sizes of 100 observations. In all models we tested for the second hidden unit after estimating the first one. The results are reported in Figures 7–8. As we can see the test is conservative with the empirical size well below the corresponding nominal one. However, the test has good power.

6 Examples

In this section we present an illustration of the modelling techniques discussed in this work. The first example considers only the in-sample fitting, the second one considers one-step ahead forecasts, and the third one considers multi-step forecasts. In all cases the variables of the model were selected using the procedure described in Section 3.1 based on a third order Taylor expansion, and the transition variables were chosen according to the p -value of the linearity test (full version).

6.1 Example 1: Canadian Lynx

The first data set analyzed is the 10-based logarithm of the number of Canadian Lynx trapped in the Mackenzie River district of North-west Canada over the period 1821–1934. For further details and a background history see Tong (1990, Chapter 7). Some previous analyzes of this series can be found in Ozaki (1982), Tsay (1989), Tong (1990), Teräsvirta (1994), and Xia and Li (1999). We start selecting the variables of the model among the first 7 lags of the time series. With the procedure described in Section 3.1 and using the SBIC, we identified lags 1 and 2 and with the AIC, lags 1,2,3,5,6, and 7. We continue building a model considering only lags 1 and 2, which is more parsimonious. The p -value of the linearity test is minimized with y_{t-2} as transition variable. The estimated model is

$$y_t = \underset{(0.18)}{0.49} + \underset{(0.07)}{1.25} y_{t-1} - \underset{(0.10)}{0.37} y_{t-2} + \left(\underset{(2.34)}{1.05} + \underset{(0.18)}{0.42} y_{t-1} - \underset{(0.56)}{0.25} y_{t-2} \right) \times F [11.02 (y_{t-2} - 3.34)] + \hat{\varepsilon}_t. \quad (43)$$

The estimated residual standard deviation is $\hat{\sigma}_\varepsilon = 0.198$. This value is smaller than in other models that use only the first two lags as variables. For example, the nonlinear model proposed by Tong (1990, p. 410), has a residual standard deviation of 0.222, the Exponential AutoRegressive (EXPAR) model proposed by Ozaki (1982) has $\hat{\sigma}_\varepsilon = 0.208$, and for the Single-Index Coefficient Regression model of Xia and Li (1999), $\hat{\sigma}_\varepsilon = 0.200$. Teräsvirta (1994) found a better result ($\hat{\sigma}_\varepsilon = 0.187$), but he included up to lag 11 in his model.

6.2 Example 2: Annual Sunspot Numbers

In this example we consider the annual sunspot numbers over the period 1700–1998. The observations for the period 1700–1979 were used to estimate the model and the remaining were used to forecast evaluation. We

adopted the same transformation as in Tong (1990), $y_t = 2 \left[\sqrt{(1 + N_t)} - 1 \right]$, where N_t is the sunspot number. We selected lags 1,2, and 7 using SBIC and lags 1,2,4,5,6,7,8,9, and 10 with AIC. Choosing the lags selected by SBIC, linearity was rejected and the p -value of the linearity test was minimized with lags 1 and 2 as transition variables. The final estimated model is

$$\begin{aligned}
y_t = & -\underset{(1.12)}{0.06} + \underset{(0.14)}{0.27}y_{t-1} + \underset{(0.12)}{0.39}y_{t-2} - \underset{(0.04)}{0.04}y_{t-7} \\
& + \left(\underset{(1.31)}{-1.87} + \underset{(0.14)}{0.21}y_{t-1} - \underset{(0.20)}{0.20}y_{t-2} + \underset{(0.05)}{0.28}y_{t-7} \right) \times F \left[2.35 \times 10^4 (0.32y_{t-1} - 0.95y_{t-2} - 0.06) \right] \\
& + \left(\underset{(1.14)}{5.36} - \underset{(0.22)}{0.13}y_{t-1} - \underset{(0.23)}{0.05}y_{t-2} + \underset{(0.06)}{0.03}y_{t-7} \right) \times F \left[0.19 \times 10^4 (0.70y_{t-1} - 0.71y_{t-2} + 0.14) \right] \quad (44) \\
& + \left(\underset{(1.18)}{-1.50} + \underset{(0.19)}{0.55}y_{t-1} - \underset{(0.15)}{0.30}y_{t-2} + \underset{(0.05)}{0.10}y_{t-7} \right) \times F \left[0.19 \times 10^3 (0.91y_{t-1} - 0.40y_{t-2} - 5.02) \right] \\
& + \hat{\varepsilon}_t.
\end{aligned}$$

The estimated in-sample residual standard deviation is $\hat{\sigma}_\varepsilon = 1.719$. As in the previous example, this value is smaller than other nonlinear models. For example, Xia and Li (1999), estimated a model where $\hat{\sigma}_\varepsilon = 1.772$ and Tong (1990, p. 420) estimated a two-regime SETAR model which has residual standard deviation of 1.932. The estimated correlation matrix of the output of the hidden units is

$$\hat{\Sigma} = \begin{pmatrix} 1 & 0.42 & -0.22 \\ 0.42 & 1 & 0.37 \\ -0.22 & 0.37 & 1 \end{pmatrix}, \quad (45)$$

indicating that there is no irrelevant neurons in the model.

We continue considering the out-of-sample performance of the estimated model. Table 4 shows the results of the one-step ahead forecast computed by model (44) and the SETAR model estimated in Tong (1990, p. 420) for 1980–1998. Table 4 shows the one-step ahead forecasts, their root mean square errors, and mean absolute errors for the transformed annual number of sunspots for the period 1980-1998. Both the root mean squared errors (RMSE) and the mean absolute errors (MAE) of our model are lower than the ones of the SETAR specification. In that sense, the flexible LSTAR model outperforms the two-regime SETAR formulation.

Table 4: One-step ahead forecasts, their root mean square errors, and mean absolute errors for the transformed annual number of sunspots for the period 1980-1998.

Year	Observation	SETAR Model		NCSTAR model	
		Forecast	Error	Forecast	Error
1980	22.95	23.45	-0.50	21.31	1.64
1981	21.78	21.51	0.27	19.87	1.91
1982	19.64	18.00	1.62	18.70	0.93
1983	14.44	15.55	-1.10	16.78	-2.34
1984	11.70	10.11	1.59	12.53	-0.83
1985	6.69	8.04	-1.35	8.01	-1.31
1986	5.59	4.85	0.74	5.35	0.24
1987	9.03	7.19	1.84	7.56	1.47
1988	18.12	12.90	5.22	17.25	0.87
1989	23.19	23.04	0.15	24.12	-0.93
1990	21.97	23.09	-1.12	21.81	0.16
1991	22.22	17.42	4.81	19.18	3.04
1992	17.52	19.19	-1.67	19.00	-1.48
1993	12.91	14.59	-1.67	15.11	-2.20
1994	9.12	8.59	0.53	9.90	-0.78
1995	6.60	6.80	-0.20	6.68	-0.08
1996	4.20	6.73	-2.53	4.30	-0.10
1997	7.49	5.30	2.19	6.63	0.85
1998	14.16	11.81	2.35	16.24	-2.08
RMSE			2.14		1.47
MAE			1.65		1.22

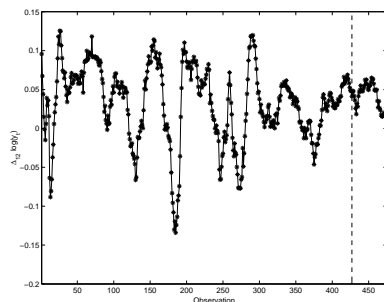


Figure 9: First difference of the logarithm of the US Industrial Production (monthly data).

Table 5: Descriptive statistics and unit root tests of the annual growth of the US Industrial Production.

Mean	σ	SK	EK	$ADF(1)$	$ADF(4)$	$PP(5)$
0.03	0.05	-0.71	0.44	-4.18 (0.0000)	-5.05 (0.0000)	-4.03 (0.0000)

6.3 Example 3: U. S. Industrial Production

We apply the flexible STAR specification to model the twelve-month difference of the logarithm of the seasonally adjusted US Industrial Production index ($\tilde{y}_t = \Delta_{12} \log(y_t)$) from 1960.1 to 1999.7 (475 observations). The data were obtained from *Economagic* (www.economagic.com). The in-sample period is from 1960.1 to 1995.7 (427 data points). The remaining points (48 observations) were used to forecast evaluation. Figure 9 shows the full sample. The dashed line indicates the split of the data.

Table 5 shows some descriptive statistics of the in-sample data, where σ is the standard deviation of the series, SK is the skewness, EK is the excess of kurtosis, $ADF(i)$ is the Augmented Dickey-Fuller unit root test with an intercept and up to lag i included in the test equation and $PP(5)$ is the Phillips-Perron unit root test with lag 5 truncation for Bartlett kernel. The p -values are given in parentheses below the statistics. As we can see, the null hypothesis of a unit root is strongly rejected.

We proceeded estimating a linear model with the lags selected with SBIC, among the the following candidate set of lags $\Upsilon = \{1, 2, 3, 4, 6, 8, 9, 12, 13, 14, 16\}$. The elements of Υ were select by observation of the partial autocorrelation function (PACF) of \tilde{y}_t . We selected the candidate lags in this way, because running the procedure described in Section 3.1 for, for example, the first 20 lags will be very time consuming. The estimated model is described in (46). Some statistics are summarized in Table 6, where $\hat{\sigma}_\varepsilon^2$ is the residual variance,

Table 6: Results of estimation of a linear model.

$\hat{\sigma}_\varepsilon$	<i>SK</i>	<i>EK</i>	<i>JB</i>	LB(12)
9.8×10^{-4}	-0.04	0.76	9.86 (7.2×10^{-4})	39.57 (0.0000)

Table 7: Results of estimation of the nonlinear model.

$\hat{\sigma}_\varepsilon$	<i>SK</i>	<i>EK</i>	<i>JB</i>
8.9×10^{-4}	-0.02	0.53	4.90 (0.09)

LB is the Ljung-Box test of no autocorrelation, and *JB* is the Jarque-Bera test of normality.

$$\tilde{y}_t = \underset{(0.01)}{0.03} + \underset{(0.02)}{1.14}\tilde{y}_{t-1} - \underset{(0.02)}{0.17}\tilde{y}_{t-4} - \underset{(0.04)}{0.40}\tilde{y}_{t-12} + \underset{(0.05)}{0.45}\tilde{y}_{t-13} - \underset{(0.02)}{0.09}\tilde{y}_{t-16} + \hat{\varepsilon}_t \quad (46)$$

We proceeded with the nonlinear modelling. First we selected the lags that compose the set of variables of the model. The variable selection procedure, based on the SBIC, identified lags 1, 4, 12, 13, and 16. We used the same set of candidate lags as in the linear case, and the same elements were chosen.

The linearity test rejected the null hypothesis at 0.01 level and the *p*-value of the test was minimized with lags 1 and 13 as transition variables. The final estimated nonlinear model, with two hidden units is

$$\begin{aligned} \tilde{y}_t = & \underset{(0.01)}{0.02} + \underset{(0.07)}{1.19}\tilde{y}_{t-1} - \underset{(0.07)}{0.07}\tilde{y}_{t-4} - \underset{(0.14)}{0.56}\tilde{y}_{t-12} + \underset{(0.17)}{0.25}\tilde{y}_{t-13} - \underset{(0.06)}{0.17}\tilde{y}_{t-16} \\ & + \left(\underset{(0.01)}{-0.02} + \underset{(0.13)}{0.10}\tilde{y}_{t-1} - \underset{(0.10)}{0.14}\tilde{y}_{t-4} - \underset{(0.19)}{0.31}\tilde{y}_{t-12} + \underset{(0.24)}{0.63}\tilde{y}_{t-13} + \underset{(0.11)}{0.12}\tilde{y}_{t-16} \right) \\ & \times F [181 (0.99\tilde{y}_{t-1} - 0.17\tilde{y}_{t-13} + 0.03)] \\ & + \left(\underset{(2.3 \times 10^{-3})}{3.9 \times 10^{-3}} - \underset{(0.12)}{0.22}\tilde{y}_{t-1} + \underset{(0.07)}{0.10}\tilde{y}_{t-4} + \underset{(0.12)}{0.5458}\tilde{y}_{t-12} - \underset{(0.16)}{0.50}\tilde{y}_{t-13} - \underset{(0.09)}{0.03}\tilde{y}_{t-16} \right) \\ & \times F [6.82 \times 10^4 (0.85\tilde{y}_{t-1} + 0.52\tilde{y}_{t-13} - 0.02)] + \hat{\varepsilon}_t. \end{aligned} \quad (47)$$

Table 7 presents some statistics of the estimated model. The estimated correlation coefficient of the two hidden units is 0.58, indicating that there is no redundant neurons in the model. The estimation results showed

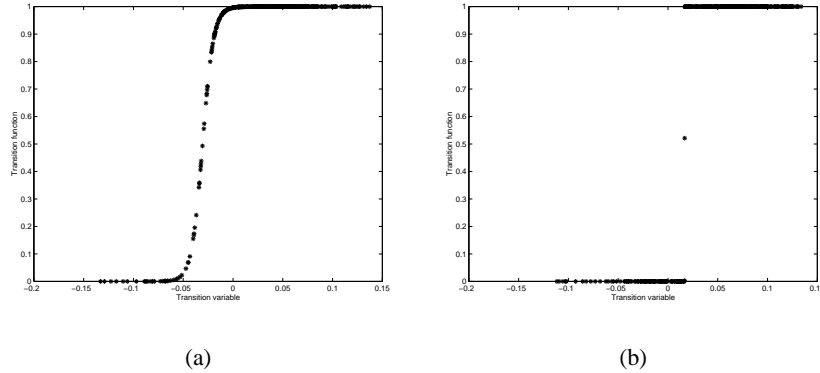


Figure 10: Transition functions versus transition variables.

Table 8: Multi-step forecast evaluation of the estimated linear and nonlinear models.

Forecast horizon	AR Model		NCSTAR model	
	RMSE	MAE	RMSE	MAE
1	0.1761	0.0040	0.1741	0.0042
2	0.2583	0.0068	0.2575	0.0066
3	0.3216	0.0086	0.3202	0.0083
4	0.3852	0.0098	0.3825	0.0099

that the flexible LSTAR model has better in-sample fit than the linear model. Normality of the residuals is not strongly rejected.

The transition functions are illustrated in Figure 10. As we can see, the first transition function is quite smooth and the second one is very sharp.

A useful diagnostic check is to consider the long-run properties of the skeleton of the estimated model (47). Since we cannot generally tackle this problem analytically, we simulated (47), with different sets of starting values, with the noise suppressed and observe how the process develops as $t \rightarrow \infty$. In the present case, y_∞ tends to an stationary point valued 6.67×10^{-2} .

The next step is to evaluate the forecasting performance of both models. We computed the 1- to 4-step ahead forecasts of the out-of-sample observations. The results are illustrated in Table 8. RMSE is the root mean squared forecasting error and MAE is the mean absolute error. In order to compute the multi-step ahead forecasts of the nonlinear model we used the Monte-Carlo technique based on 2000 replications. Both models have almost the same forecasting performance, with the flexible STAR model being slightly better.

7 Conclusions

In this paper we consider a generalization of the logistic STAR model in order to deal with multiple regimes and to obtain a flexible specification of the transition variables. Furthermore, the results presented here can be easily generalized into a multivariate framework with exogenous variables. The proposed model nests several nonlinear models, such as, for example, the SETAR, STAR, and AR-ANN models, and thus is very flexible. Even more, if the neural network is interpreted as a nonparametric universal approximation to any Borel-measurable function, the proposed model is comparable to the FAR model, and the Single-Index Coefficient Regression model. A model specification procedure based on statistical inference is developed and the results of a simulation experiment showed that the proposed tests are well sized and have good power in small samples. When put into test in real experiments, the proposed model outperforms the linear model and other nonlinear specifications. Finally, both the simulation study and the real examples suggest that the theory developed here is useful and the proposed model thus seems to be a useful tool for the practicing time series analysts.

Acknowledgments

The authors would like to thank Timo Teräsvirta, Gianluigi Rech, and Carlos Pedreira for valuable comments, and the CNPq for the financial support. Part of this work was done while the first author was a visiting graduate student at the Department of Economic Statistics, Stockholm School of Economics, whose kind hospitality is gratefully acknowledged.

References

- Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**: 716–723.
- Anders, U. and Korn, O. (1999). Model selection in neural networks, *Neural Networks* **12**: 309–323.
- Astatkie, T., Watts, D. G. and Watt, W. E. (1997). Nested threshold autoregressive (NeTAR) models, *International Journal of Forecasting* **13**: 105–116.

- Auestad, A. and Tjøstheim, D. (1990). Identification of nonlinear time series: First order characterization and order determination, *Biometrika* **77**: 669–687.
- Bacon, D. W. and Watts, D. G. (1971). Estimating the transition between two intersecting lines, *Biometrika* **58**: 525–534.
- Bertsekas, D. P. (1995). *Nonlinear Programming*, Athena Scientific, Belmont, MA.
- Chan, K. S. and Tong, H. (1986). On estimating thresholds in autoregressive models, *Journal of Time Series Analysis* **7**: 179–190.
- Chen, R. and Tsay, R. S. (1993). Functional coefficient autoregressive models, *Journal of the American Statistical Association* **88**: 298–308.
- Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions, *Econometrica* **28**: 591–605.
- Cooper, S. J. (1998). Multiple regimes in US output fluctuations, *Journal of Business and Economic Statistics* **16**(1): 92–100.
- Davidson, R. and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*, Oxford University Press, New York, NY.
- Davies, R. B. (1977). Hypothesis testing when the nuisance parameter is present only under the alternative, *Biometrika* **64**: 247–254.
- Davies, R. B. (1987). Hypothesis testing when the nuisance parameter is present only under the alternative, *Biometrika* **74**: 33–44.
- Eitrheim, . and Teräsvirta, T. (1996). Testing the adequacy of smooth transition autoregressive models, *Journal of Econometrics* **74**: 59–75.
- Franses, P. H. and Paap, R. (1999). Censored latent effects autoregression with an application to US unemployment, *Econometric Institute Report 9841/A*, Econometric Institute – Erasmus University.
- Goldfeld, S. M. and Quandt, R. (1972). *Nonlinear Methods in Econometrics*, North Holland, Amsterdam.

- Granger, C. W. J. and Teräsvirta, T. (1993). *Modelling Nonlinear Economic Relationships*, Oxford University Press, Oxford.
- Hwang, J. T. G. and Ding, A. A. (1997). Prediction intervals for artificial neural networks, *Journal of the American Statistical Association* **92**(438): 109–125.
- Kurková, V. and Kainen, P. C. (1994). Functionally equivalent feedforward neural networks, *Neural Computation* **6**: 543–558.
- Leisch, F., Trapletti, A. and Hornik, K. (1999). Stationarity and stability of autoregressive neural network processes, in M. S. Kearns, S. A. Solla and D. A. Cohn (eds), *Advances in Neural Information Processing Systems*, Vol. 11, MIT Press, USA.
- Lewis, P. A. W. and Stevens, J. G. (1991). Nonlinear modeling of time series using multivariate adaptive regression splines, *Journal of the American Statistical Association* **86**: 864–877.
- Luukkonen, R., Saikkonen, P. and Teräsvirta, T. (1988). Testing linearity against smooth transition autoregressive models, *Biometrika* **75**: 491–499.
- Medeiros, M. C. and Veiga, A. (1999). A hybrid linear-neural model for time series forecasting, *IEEE Transactions on Neural Networks*. forthcoming.
- Ozaki, T. (1982). The statistical analysis of perturbed limit cycle process using nonlinear time series models, *Journal of Time Series Analysis* **3**: 29–41.
- Quandt, R. E. (1960). Tests of hypothesis that a linear regression system obeys two separate regimes, *Journal of the American Statistical Association* **55**: 324–330.
- Rech, G., Teräsvirta, T. and Tschernig, R. (1999). A simple variable selection technique for nonlinear models, *Working Paper Series in Economics and Finance* 296, Stockholm School of Economics.
- Saikkonen, P. and Luukkonen, R. (1988). Lagrange multiplier tests for testing non-linearities in time series models, *Scandinavian Journal of Statistics* **15**: 55–68.
- Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics* **6**: 461–464.

- Sussman, H. J. (1992). Uniqueness of the weights for minimal feedforward nets with a given input-output map, *Neural Networks* **5**: 589–593.
- Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models, *Journal of the American Statistical Association* **89**(425): 208–218.
- Teräsvirta, T. and Lin, C.-F. J. (1993). Determining the number of hidden units in a single hidden-layer neural network model, *Research report*, Bank of Norway.
- Teräsvirta, T., Lin, C. F. and Granger, C. W. J. (1993). Power of the neural network linearity test, *Journal of Time Series Analysis* **14**(2): 309–323.
- Tiao, G. C. and Tsay, R. S. (1994). Some advances in non-linear and adaptive modelling in time-series, *Journal of Forecasting* **13**: 109–131.
- Tjøstheim, D. and Auestad, A. (1994). Nonparametric identification of nonlinear time series: Selecting significant lags, *Journal of the American Statistical Association* **89**: 1410–1419.
- Tong, H. (1990). *Non-linear Time Series: A Dynamical Systems Approach*, Vol. 6 of *Oxford Statistical Science Series*, Oxford University Press, Oxford.
- Tsay, R. (1989). Testing and modeling threshold autoregressive processes, *Journal of the American Statistical Association* **84**: 431–452.
- van Dijk, D. (1999). *Smooth Transition Models: Extensions and Outlier Robust Inference*, PhD thesis, Tinbergen Institute, Rotterdam, The Netherlands. www.few.eur.nl/few/people/djvandijk/thesis/thesis.htm.
- van Dijk, D. and Franses, P. H. (1999). Modelling multiple regimes in the business cycle, *Macroeconomic Dynamics* **3**(3): 311–340.
- Veiga, A. and Medeiros, M. (1998). A hybrid linear-neural model for time series forecasting, *Proceedings of the NEURAP 98*, Marseilles, pp. 377–384.
- View, P. (1995). Order choice in nonlinear autoregressive models, *Statistics* **26**: 307–328.
- Xia, Y. and Li, W. K. (1999). On single-index coefficient regression models, *Journal of the American Statistical Association* **94**(448): 1275–1285.

Yao, Q. and Tong, H. (1994). On subset selection in non-parametric stochastic regression, *Statistica Sinica* **4**: 51–70.