A Flexible Coefficient Smooth Transition Time Series Model

Marcelo C. Medeiros and Álvaro Veiga

Abstract—**In this paper, we consider a flexible smooth transition autoregressive (STAR) model with multiple regimes and multiple transition variables. 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 neural network (AR-NN), 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. A model building procedure 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.**

Index Terms—**Neural networks, smooth transition models, threshold models, time series.**

I. 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 [1] and further developed in [2] and [3], has found a number of successful applications [4]. The term "smooth transition" in its present meaning first appeared in [5]. 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. [6, 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 between them 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. The model is called the neuro-coefficient smooth transition autoregressive (NCSTAR) model and was introduced in [7] and [8].

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 [9] with multiple regimes, the autoregressive neural network (AR-NN) model [10], [11], and the logistic STAR model [3]. The proposed model is also able to fit time series were the true generating process is an exponential STAR (ESTAR) model [3]. Furthermore, our model can be also compared to the functional coefficient autoregressive (FAR) model of [12], and the single-index coefficient regression model of [13].

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 behavior of macro-economic variables over the business cycle. Recent studies conclude that a two-regime modeling of the business cycle is rather limited. See, for example, [14], where a multiple regime STAR (MRSTAR) model is proposed and applied to describe the behavior of the U.S. gross national product (GNP) and U.S. unemployment rate [15], where an additive logistic STAR model is applied to describe business cycle nonlinearity in U.K. macroeconomic time series, or [16] where a regression tree approach is used to model multiple regimes in the U.S. industrial production. In the framework of the SETAR model, modeling multiple regimes is a well established methodology [9], [17].

Second, multiple transition variables are useful in describing complex nonlinear behavior and allow for different sources of nonlinearity. Several papers concerning multiple transition variables 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, [18], where the thresholds are controlled by two lagged values of a transformed U.S. GNP series reflecting the situation of the economy or [14]. In the present framework, we adopt a less restrictive formulation, assuming that the linear combination of variables is unknown and is estimated jointly 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 [19]–[21]).[1]

A modeling cycle procedure based on the work in [22]–[24], consisting of the stages of model specification and parameter estimation, is developed, allowing the practitioner to choose among different model specifications during the modeling

Manuscript received December 9, 2002; revised April 15, 2003. This work was supported by CNPq. This work is partly based on the doctoral dissertation of M. C. Medeiros.

M. C. Medeiros is with the Department of Economics, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, RJ 22451-900 Brazil.

Á. Veiga is with the Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, RJ 22451-900 Brazil.

Digital Object Identifier 10.1109/TNN.2004.836246

[1] It is worth mentioning that the proposal of [21] is a special case of the MRSTAR model proposed by [14].

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 model evaluation step of the modeling cycle is developed in [25].

The plan of the paper is as follows. Section II presents the model. Section III deals with the specification. Section IV analyzes the estimation procedures. Section V presents a Monte Carlo experiment to find out the behavior of the proposed tests and Section VI shows some examples with real data. Concluding remarks are made in Section VII.

II. NCSTAR Model

One important class of STAR models is the logistic STAR model of order p, LSTAR(p), proposed by [2] and defined as

$$y_t = \boldsymbol{\alpha}'\mathbf{z}_t + \boldsymbol{\lambda}'\mathbf{z}_t F\left(\gamma(y_{t-d} - c)\right) + \varepsilon_t \tag{1}$$

where ε_t is a normally distributed white noise with variance σ^2, $\mathbf{z}_t = [1, \tilde{\mathbf{z}}_t']'$, $\tilde{\mathbf{z}}_t \in \mathbb{R}^p$ is formed by a set of lagged values of y_t, and $F(\cdot)$ is the logistic function

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

The 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 $\gamma \to \infty$, model (1) becomes a two-regime SETAR model [9, p. 183].

In the present paper, we consider an additive logistic STAR model with multiple regimes and multivariate transition variables. 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 [7] and [8].

Consider a linear model with time-varying coefficients expressed as

$$y_t = \boldsymbol{\phi}_t'\mathbf{z}_t + \varepsilon_t \tag{3}$$

where $\boldsymbol{\phi}_t = [\phi_t^{(0)}, \phi_t^{(1)}, \ldots, \phi_t^{(p)}]' \in \mathbb{R}^{p+1}$ is a vector of coefficients and ε_t and $\mathbf{z}_t$ are defined as before. 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

$$\phi_t^{(j)} = \sum_{i=1}^{h} \lambda_{ji} F\left(\boldsymbol{\omega}_i'\mathbf{x}_t - \beta_i\right) - \lambda_{j0}, \quad j = 0, \cdots, p \tag{4}$$

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

The function $F(\boldsymbol{\omega}_i'\mathbf{x}_t - \beta_i)$ is the logistic function, where $\mathbf{x}_t \in \mathbb{R}^q$ is a vector of input variables, $\boldsymbol{\omega}_i = [\omega_{1i}, \ldots, \omega_{qi}]' \in \mathbb{R}^q$ and $\beta_i \in \mathbb{R}$ are parameters. The norm of $\boldsymbol{\omega}_i$ is called the *slope parameter*. In the limit, when the slope parameter approaches infinity, the logistic function becomes a step function. The elements of $\mathbf{x}_t$, called the transition variables, is formed by lagged values of y_t.[2] 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

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

or in vector notation

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

where $\boldsymbol{\psi} = [\boldsymbol{\alpha}', \boldsymbol{\lambda}_1', \ldots, \boldsymbol{\lambda}_h', \boldsymbol{\omega}_1', \ldots, \boldsymbol{\omega}_h', \beta_1, \ldots, \beta_h]' \in \mathbb{R}^r$, $r = (q+1) \times h + (p+1) \times (h+1)$, is a parameter vector, $\boldsymbol{\alpha} = [\alpha_0, \ldots, \alpha_p]' = [-\lambda_{00}, \ldots, -\lambda_{p0}]'$, and $\boldsymbol{\lambda}_i = [\lambda_{0i}, \ldots, \lambda_{pi}]'$.

Note that model (6) is, in principle, neither globally nor locally identified. There are three characteristics of neural networks which cause nonidentifiability. The first one is due to the symmetries in the neural network architecture. The value of 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. Finally, the presence of irrelevant hidden units (overparametrized model) is a problem. If model (6) has at least one hidden unit with $\boldsymbol{\lambda}_i = \mathbf{0}$, then parameters $\boldsymbol{\omega}_i$ and β_i are unidentified. On the other hand, if $\boldsymbol{\omega}_i = \mathbf{0}$, then $\boldsymbol{\lambda}_i$ and β_i can take any value without changing the value of the likelihood function.

The first problem is solved by imposing the restrictions $\beta_1 \leq \ldots \leq \beta_h$. The second problem can be circumvented, for example, by imposing the restriction $\omega_{1i} > 0$, $i = 1, \ldots, h$. To remedy the third problem, it is necessary to ensure that the model contains no irrelevant hidden units. This is tackled with the tests described in Section III. For further discussion of the identifiability concepts see, e.g., [26]–[29].

For estimation purposes it is often useful to reparametrize model (6) as

$$\begin{aligned} y_t &= G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}) + \varepsilon_t \\ &= \boldsymbol{\alpha}'\mathbf{z}_t + \sum_{i=1}^{h} \boldsymbol{\lambda}_i'\mathbf{z}_t F\left[\gamma_i\left(\tilde{\boldsymbol{\omega}}_i'\mathbf{x}_t - c_i\right)\right] + \varepsilon_t \end{aligned} \tag{7}$$

where $\gamma_i > 0$ and $||\tilde{\boldsymbol{\omega}}_i|| = 1$ with

$$\tilde{\omega}_{i1} = \sqrt{1 - \sum_{j=2}^{q} \tilde{\omega}_{ij}^2} > 0. \tag{8}$$

The parameter vector $\boldsymbol{\psi}$ is redefined as

$$\begin{aligned} \boldsymbol{\psi} = \big[\boldsymbol{\alpha}', \boldsymbol{\lambda}', \ldots, \boldsymbol{\lambda}_h', \gamma_1, \ldots, \gamma_h, \tilde{\omega}_{12}, \ldots,& \\ \tilde{\omega}_{1q}, \ldots, \tilde{\omega}_{h2}, \ldots, \tilde{\omega}_{hq}, c_1, \ldots, c_h\big]'&. \end{aligned}$$

This reparametrization has been also applied in [24].

[2] It is important to mention that the NCSTAR model can be easily generalized to include some exogenous variables in $\mathbf{z}_t$ and/or in $\mathbf{x}_t$.

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

$$y_t = \boldsymbol{\alpha}'\mathbf{z}_t + \sum_{i=1}^{h} \left[\boldsymbol{\lambda}_i'\mathbf{z}_t F\left(\gamma_i(y_{t-d} - c_i)\right)\right] + \varepsilon_t \qquad (9)$$

It should be noticed that model (9) nests the SETAR model with $h+1$ regimes. When $\gamma_i \to \infty, i = 1, \ldots, h$ model (9) becomes a SETAR model with $h+1$ regimes.

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

$$\mathbb{H} = \left\{\mathbf{x}_t \in \mathbb{R}^q | \tilde{\boldsymbol{\omega}}_i'\mathbf{x}_t = c_i\right\}. \qquad (10)$$

The direction of $\tilde{\boldsymbol{\omega}}_i$ 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

$$\mathbb{H}^+ = \left\{\mathbf{x}_t \in \mathbb{R}^q | \tilde{\boldsymbol{\omega}}_i'\mathbf{x}_t \geq c_i\right\} \qquad (11)$$

and

$$\mathbb{H}^- = \left\{\mathbf{x}_t \in \mathbb{R}^q | \tilde{\omega}_i'\mathbf{x}_t \geq c_i\right\}. \qquad (12)$$

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 (11) and (12) of each hyperplane.

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

$$y_t = \boldsymbol{\alpha}'\mathbf{z}_t + \sum_{i=1}^{h} \boldsymbol{\lambda}_i'\mathbf{z}_t F\left(\gamma_i(\tilde{\boldsymbol{\omega}}'\mathbf{x}_t - c_i)\right) + \varepsilon_t \qquad (13)$$

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

Another interesting case is when $\boldsymbol{\lambda}_i' = [\lambda_{0i}, 0, \ldots, 0]$ in (9). Then model (7) becomes an AR-NN model. AR-NN models can be interpreted as a linear model where the intercept is time-varying and changes smoothly between regimes.

An 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 FAR model of [12], and the single-index coefficient regression model of [13].

III. Specification

From (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 and to avoid overfitting. It is well-known that for neural network models overfitting is a serious problem and as the NCSTAR model nests the neural network specification as a special case, the same problem may occur here. To avoid overfitting a coherent specific-to-general model building procedure is developed based on statistical arguments. The specification strategy adopted here is based on the linearization of the nonlinear term of model (7) and a sequence of Lagrange multiplier (LM) tests is developed to determine the number of hidden units of the model, which is carried out together with the estimation of the parameters of the model.

In order to select the variables of (7), we assume that $\mathbf{x}_t$ is formed by a subset of the elements of $\mathbf{z}_t$ This is not a to restrictive assumption because we can always augment the elements of $\mathbf{z}_t$ to include all the variables in $\mathbf{x}_t$ and then use standard hypothesis tests to test the significance of the extra parameters in the linear part of the model.

A. Variable Selection

In the context of STAR models, [3] suggests first specifying a linear autoregressive model for the data under analysis using an information criterion such as the Akaike's information criterion (AIC) [30] or the Schwarz's Bayesian information criterion (SBIC) [31]. The second step is to test the null hypothesis of linearity against the alternative of STAR nonlinearity. If linearity is rejected, select the appropriate transition variable by running the linearity test for different variables and choose the one that minimize the p-value of the test.

Another possibility is to use nonparametric methods based on local estimators [32]–[36]. However, those methods require a large number of observations.

In this paper, we adopt a generalization of the method considered in [3] and is based on the procedure proposed by [23]. The idea is to use a polynomial expansion of the model to select the variables in $\mathbf{z}_t$ and then, chose the elements of $\mathbf{x}_t$ among every possible combination of the elements of $\mathbf{z}_t$, by running the linearity test for each one of them. We give a brief overview of the method. For more details, see [23].

Consider model (7). The basic idea is to conduct the selection on a parametric function $g(\cdot)$ which can approximate the true function $G(\cdot)$ well but is much simpler to estimate. A well-known class of simple approximating functions are series expansions

$$g(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\xi}) = \sum_{i=1}^{L} \xi_i g_i(\mathbf{z}_{t,i}, \mathbf{x}_{t,i})$$

with parameters ξ_i, known basis functions $g_i(\cdot)$ and $\mathbf{z}_{t,i}$ and $\mathbf{x}_{t,i}$ being general subvectors of $\mathbf{z}_t$ and $\mathbf{x}_t$. Due to the linearity one can estimate the parameters ξ_i, $i = 1, \ldots, L$ by ordinary least squares. Of course, the quality of approximation depends on the choice of the basis functions $g_i(\cdot)$ and the length of the expansion L.

In order to define $g_i(\cdot)$, assume that the sample space $\mathcal{Z}$ is compact and that $G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})$ is continuous in $\mathcal{Z}$. Then it follows from the Stone-Weierstrass theorem that $G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})$ can be uniformly approximated by a polynomial in the components of $\mathbf{z}_t$ and $\mathbf{x}_t$, see [37, pp. 150–151]. Thus, using a general kth-order polynomial one obtains

$$G(\mathbf{z}_t,\mathbf{x}_t;\boldsymbol{\psi}) = \boldsymbol{\xi}_1'\mathbf{z}_t + \sum_{j_1=1}^{p}\sum_{j_2=j_1}^{p}\xi_{j_1j_2}z_{j_1,t}z_{j_2,t} + \sum_{j_1=1}^{p}\cdots\sum_{j_k=j_{k-1}}^{p}\xi_{j_1\dots j_k}z_{j_1,t}\cdots z_{j_k,t} + R(\mathbf{z}_t,\mathbf{x}_t;\boldsymbol{\psi}) \quad (14)$$

where $R(\mathbf{z}_t,\mathbf{x}_t;\boldsymbol{\psi})$ is the remainder and $\boldsymbol{\xi} = [\boldsymbol{\xi}_1', \xi_{j_1j_2}, \xi_{j_1j_2j_3}]'$ is the vector of parameters. Note that the terms involving $\mathbf{x}_t$ merged with the terms involving $\mathbf{z}_t$ as we are considering in this paper that the elements in $\mathbf{x}_t$ are a subset of the elements in $\mathbf{z}_t$.

The second step is to regress y_t on all variables in the polynomial 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 polynomial 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 expansion consists of a function of a single regressor. Choose the combination of variables that yields the lowest value of the SBIC.

If we test each possible combination of variables, we would need to estimate $\sum_{i=1}^{p}(p!/(i!(p-i)!))$ different models. If p is very large, it is not reasonable to test every possible combination. In that case, the practitioner may only estimate p models where just the set

$$\Lambda = \{z_{1,t}; z_{1,t}, z_{2,t}; z_{1,t}, z_{2,t}, z_{3,t}; \dots; z_{1,t}, \dots, z_{p,t}\}$$

is considered.[3] Not testing every possible combination of variables may cause an overparametrization of $\mathbf{z}_t$. However, this not pose serious problems as far as hypothesis tests are carried out to remove redundant variables. As suggested by one of the referees, another possibility to make the variable selection process easier is to consider only a subset of the principal components of $\mathbf{z}_t$.

B. Testing Linearity

In practical nonlinear time series modeling, 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 $\mathbf{x}_t$. After selecting the elements of $\mathbf{z}_t$ with the procedure described in Section III-A, we choose the elements of $\mathbf{x}_t$ by running the linearity test described below setting $\mathbf{x}_t$ 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, the transition function $F[\gamma_i(\tilde{\boldsymbol{\omega}}_i'\mathbf{x}_t - c_i)]$ is redefined as

$$F\left[\gamma_i\left(\tilde{\boldsymbol{\omega}}_i'\mathbf{x}_t - c_i\right)\right] = \frac{1}{1+\exp\left(-\gamma_i\left(\tilde{\boldsymbol{\omega}}_i'\mathbf{x}_t - c_i\right)\right)} - \frac{1}{2}. \quad (15)$$

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

Consider (7) 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 $\mathbf{H}_0 : \boldsymbol{\lambda}_i = \mathbf{0}$, $i = 1,\dots,h$. Note also that $F(0) = 0$. This implies another possible null hypothesis of linearity

$$\mathrm{H}_0 : \gamma_i = 0, i = 1,\dots,h. \quad (16)$$

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

$$y_t = \boldsymbol{\alpha}'\mathbf{z}_t + \boldsymbol{\lambda}'\mathbf{z}_t F\left[\gamma\left(\tilde{\boldsymbol{\omega}}_i'\mathbf{x}_t - c_i\right)\right] + \varepsilon_t. \quad (17)$$

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. [38] and [39] first discussed solutions to this problem. Following [2], [40], and [41] we solve the problem by replacing $F[\gamma(\tilde{\boldsymbol{\omega}}'\mathbf{x}_t - c)]$ by a low-order Taylor expansion approximation about $\gamma = 0$. Consider a first-order Taylor expansion of (15)

$$\begin{aligned}\tilde{F}_1(\mathbf{x}_t;\gamma,\tilde{\boldsymbol{\omega}},c) &= F(0) + \left.\frac{\partial F}{\partial \gamma}\right|_{\gamma=0}\gamma + R_1(\mathbf{x}_t;\gamma,\tilde{\boldsymbol{\omega}},c)\\ &= \frac{1}{4}\gamma(\tilde{\boldsymbol{\omega}}'\mathbf{x}_t - c) + R_1(\mathbf{x}_t;\gamma,\tilde{\boldsymbol{\omega}},c) \quad (18)\end{aligned}$$

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

$$y_t = \boldsymbol{\alpha}'\mathbf{z}_t + \frac{1}{4}\gamma\boldsymbol{\lambda}'\mathbf{z}_t(\tilde{\boldsymbol{\omega}}'\mathbf{x}_t - c) + \varepsilon_t^* \quad (19)$$

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

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

where $\mathbf{z}_t^* \in \mathbb{R}^{p-q}$ is formed by the elements of $\mathbf{z}_t$ that are not in $\mathbf{x}_t$.

Using (20) instead of (17) circumvents the identification problem, and we obtain a simple test of linearity. The null hypothesis can be defined as $\mathbf{H}_0 : \theta_{ij} = 0, \beta_{ij} = 0, \rho_{ij} = 0$. 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 the test has no power. To remedy this situation, [2] suggests

[3] Again the elements of $\mathbf{x}_t$ are omitted because we consider that $\mathbf{x}_t$ is a subset of $\mathbf{z}_t$.

a third-order Taylor approximation of the transition function, expressed as

$$\begin{aligned}\tilde{F}_3(\mathbf{x}_t;\gamma,\tilde{\boldsymbol{\omega}},c) = &\frac{1}{4}\gamma(\tilde{\boldsymbol{\omega}}'\mathbf{x}_t - c)\\ &+\frac{1}{96}\gamma^3(\tilde{\boldsymbol{\omega}}'\mathbf{x}_t - c)^3 + R_3(\mathbf{x}_t;\gamma,\tilde{\boldsymbol{\omega}},c). \quad (21)\end{aligned}$$

Replacing (15) by (21) in (17) we get

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

The null hypothesis is defined as $\mathbf{H}_0$: $\theta_{ij} = 0$, $\beta_{ij} = 0$, $\theta_{ijk} = 0$, $\beta_{ijk} = 0$, $\theta_{ijkl} = 0$ and $\beta_{ijkl} = 0$.

Now we can use (22) to test linearity. Note that $\varepsilon^*_t = \varepsilon_t$ when the null hypothesis is true. The local approximation to the log-likelihood for observation t takes the form

$$\begin{aligned}l_t = &-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\\ &\times\left\{y_t - \boldsymbol{\pi}'\mathbf{z}_t - \boldsymbol{\lambda}_1'\mathbf{z}_t\sum_{i=1}^{q}\sum_{j=i}^{q}\theta_{ij}x_{i,t}x_{j,t}\right.\\ &-\sum_{i=1}^{p-q}\sum_{j=1}^{q}\beta_{ij}z^*_{i,t}x_{j,t}\\ &-\sum_{i=1}^{q}\sum_{j=i}^{q}\sum_{k=j}^{q}\theta_{ijk}x_{i,t}x_{j,t}x_{k,t}\\ &-\sum_{i=1}^{p-q}\sum_{j=1}^{q}\sum_{k=j}^{q}\beta_{ijk}z^*_{i,t}x_{j,t}x_{k,t}\\ &-\sum_{i=1}^{q}\sum_{j=i}^{q}\sum_{k=j}^{q}\sum_{l=k}^{q}\theta_{ijkl}x_{i,t}x_{j,t}x_{k,t}x_{l,t}\\ &\left.-\sum_{i=1}^{p-q}\sum_{j=1}^{q}\sum_{k=j}^{q}\sum_{l=k}^{q}\beta_{ijkl}z^*_{i,t}x_{j,t}x_{k,t}x_{l,t}\right\}^2. \quad (23)\end{aligned}$$

At this point we make the following assumptions.

Assumption 1: The $((r+1)\times 1)$ parameter vector $\boldsymbol{\psi}^* = [\psi', \sigma^2]'$ is an interior point of the compact parameter space $\boldsymbol{\Psi}$ which is a subspace of $\mathbb{R}^r \times \mathbb{R}^+$, the r-dimensional Euclidean space.

Assumption 2: Under the null the data generating process (DGP) for the sequence of scalar real valued observations $\{y_t\}_{t=1}^T$ is an ergodic stochastic process, with true parameter vector $\boldsymbol{\psi}^* \in \boldsymbol{\Psi}$.

Assumption 3: $E|z_{t,i}|^\delta < \infty, i = 1,\ldots,p$ for some $\delta > 8$.

Assumption 2 implies that, under the null, the linear autoregressive process $y_t = \boldsymbol{\alpha}'\mathbf{z}_t + \varepsilon_t$ is ergodic.

Under $\mathbf{H}_0$ and Assumptions 1–3 the standard LM or score type test statistic

$$\begin{aligned}\mathrm{LM} = \frac{1}{\widehat{\sigma}^2}\sum_{t=1}^{T}\widehat{\varepsilon}_t\widehat{\boldsymbol{\nu}}_t' \times &\left\{\sum_{t=1}^{T}\widehat{\boldsymbol{\nu}}_t\widehat{\boldsymbol{\nu}}_t' - \sum_{t=1}^{T}\widehat{\boldsymbol{\nu}}_t\widehat{\mathbf{h}}_t'\right.\\ &\left.\times\left(\sum_{t=1}^{T}\widehat{\mathbf{h}}_t\widehat{\mathbf{h}}_t'\right)^{-1}\sum_{t=1}^{T}\widehat{\mathbf{h}}_t\widehat{\boldsymbol{\nu}}_t'\right\}^{-1} \times\sum_{t=1}^{T}\widehat{\boldsymbol{\nu}}_t\widehat{\varepsilon}_t \quad (24)\end{aligned}$$

where $\mathbf{h}_t = \mathbf{z}_t$, $\widehat{\varepsilon}_t = y_t - \widehat{\boldsymbol{\pi}}'\mathbf{z}_t$ and $\boldsymbol{\nu}_t$ is formed by all nonlinear regressors in (22), has an asymptotic χ^2 distribution with m degrees of freedom when the null hypothesis holds, where m is the number of elements in $\boldsymbol{\nu}_t$ (see [42] for details on LM type tests).

The test can be carried out in stages as follows:

1) regress y_t on $\mathbf{z}_t$ and compute $SSR_0 = \sum_{t=1}^{T}\widehat{\varepsilon}_t^2$;
2) regress $\widehat{\varepsilon}_t$ on $\mathbf{z}_t$ and on the m nonlinear regressors of (22). Compute the residual sum of squares $SSR_1 = \sum_{t=1}^{T}\widehat{\nu}_t^2$;
3) compute the χ^2 statistic

$$\mathrm{LM}^l_{\chi^2} = T\frac{SSR_0 - SSR_1}{SSR_0}eqno(25)$$

or the F version of the test

$$\mathrm{LM}^l_F = \frac{\frac{(SSR_0 - SSR_1)}{m}}{\frac{SSR_1}{(T-p-1-m)}} \quad (26)$$

where T is the number of observations.

When $\mathbf{z}_t$ and 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 [43, Ch. 7]) that an F-approximation works much better. Another possibility to improve the power of the test is to follow the idea of [29] 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. [2] suggests 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).

C. 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.

A way of doing this is applying popular methods such as pruning, in which a neural network model with a large number of hidden units is estimated first, and the size of the model is subsequently reduced. Another possibility is to sequentially add hidden units to the model based on the use of model a selection criterion such as SBIC or AIC.

However, this technique has a major drawback. Suppose the data have been generated by a NCSTAR model with h hidden units. Applying, for example, to SBIC to decide if another hidden unit should be added requires estimation of a model with $h+1$ hidden neurons. In this situation, the larger model is not identified and its parameters cannot be estimated consistently. This is likely to cause numerical problems in maximum likelihood estimation. Besides, even when convergence is achieved, lack of identification causes problems in interpreting the SBIC. A comparison of the two models based on the SBIC is then equivalent to a likelihood ratio test of h units against $h+1$ ones; see, for example, [44] for discussion. But then, when the larger model is not identified under the null hypothesis, the likelihood ratio statistic does not have its standard asymptotic χ^2 distribution when the null holds.

In this paper, we also select the hidden units sequentially but circumvent the identification problem in a way that enables us to control the significance level of the tests in the sequence and, thus, also the overall significance level of the procedure. This can be done combining the ideas of the neural network test of [41], the test of remaining nonlinearity of [22] and the results in [24] and [45]. The basic idea is to start using the test of Section III-B 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. At every step we halve the significance level of the test. This way we avoid overfitting and control the overall significance level of the procedure. An upper bound for the overall significance level may be obtained using the Bonferroni bound; see [46, p. 59].

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

$$\begin{aligned} y_t = \boldsymbol{\alpha}'\mathbf{z}_t + \boldsymbol{\lambda}_1'\mathbf{z}_t F\left[\gamma_1\left(\tilde{\boldsymbol{\omega}}_1'\mathbf{x}_t - c_1\right)\right] \\ + \boldsymbol{\lambda}_2'\mathbf{z}_t F\left[\gamma_2\left(\tilde{\boldsymbol{\omega}}_2'\mathbf{x}_t - c_2\right)\right] + \varepsilon_t. \end{aligned} \quad (27)$$

If we want to test for the second hidden unit in (27), an appropriate null hypothesis is

$$\mathrm{H}_0 : \gamma_2 = 0 \quad (28)$$

whereas the alternative is $\mathbf{H}_1 : \gamma^2 \neq 0$. We assume that under this null hypothesis the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}_1$, γ_1, $\tilde{\boldsymbol{\omega}}_1$ and c_1 can be consistently estimated and that the estimators are asymptotically normal. Note that (27) is only identified under the alternative. We may solve this problem in the same fashion we did in Section III-B, using a low-order Taylor expansion of $F[\gamma_2(\tilde{\boldsymbol{\omega}}_2'\mathbf{x}_t - c_2)]$ about $\gamma_2 = 0$. Using a third-order expansion and after rearranging terms, the resulting model is

$$\begin{aligned} y_t =& \boldsymbol{\pi}'\mathbf{z}_t + \boldsymbol{\lambda}_1'\mathbf{z}_t F\left[\gamma_1\left(\tilde{\boldsymbol{\omega}}_1'\mathbf{x}_t - c_1\right)\right] \\ &+ \sum_{i=1}^{q}\sum_{j=i}^{q}\theta_{ij}x_{i,t}x_{j,t} + \sum_{i=1}^{p-q}\sum_{j=1}^{q}\beta_{ij}z_{i,t}^{*}x_{j,t} \\ &+ \sum_{i=1}^{q}\sum_{j=i}^{q}\sum_{k=j}^{q}\theta_{ijk}x_{i,t}x_{j,t}x_{k,t} \\ &+ \sum_{i=1}^{p-q}\sum_{j=1}^{q}\sum_{k=j}^{q}\beta_{ijk}z_{i,t}^{*}x_{j,t}x_{k,t} \\ &+ \sum_{i=1}^{q}\sum_{j=i}^{q}\sum_{k=j}^{q}\sum_{l=k}^{q}\theta_{ijkl}x_{i,t}x_{j,t}x_{k,t}x_{l,t} \\ &+ \sum_{i=1}^{p-q}\sum_{j=1}^{q}\sum_{k=j}^{q}\sum_{l=k}^{q}\beta_{ijkl}z_{i,t}^{*}x_{j,t}x_{k,t}x_{l,t} + \varepsilon_t^{*}. \end{aligned} \quad (29)$$

The null hypothesis is defined as $\mathbf{H}_0$:, $\theta_{ij} = 0$, $\beta_{ij} = 0$, $\theta_{ijk} = 0$, $\beta_{ijk} = 0$, $\theta_{ijkl} = 0$, and $\beta_{ijkl} = 0$. We define the residuals estimated under the null hypothesis as $\widehat{\varepsilon}_t = y_t - \widehat{\boldsymbol{\pi}}'\mathbf{z}_t - \boldsymbol{\lambda}_1'\mathbf{z}_t F[\widehat{\gamma}_1(\widehat{\tilde{\boldsymbol{\omega}}}_1'\mathbf{x}_t - \widehat{c}_1)]$.

The local approximation to the normal log likelihood function in a neighborhood of $\mathbf{H}_0$ for observation t and ignoring the remainder is

$$\begin{aligned} l_t =& -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2} \\ &\times \Bigg\{ y_t - \boldsymbol{\pi}'\mathbf{z}_t - \boldsymbol{\lambda}_1'\mathbf{z}_t F\left[\gamma_1\left(\tilde{\boldsymbol{\omega}}_1'\mathbf{x}_t - c_1\right)\right] \\ &- \sum_{i=1}^{q}\sum_{j=i}^{q}\theta_{ij}x_{i,t}x_{j,t} - \sum_{i=1}^{p-q}\sum_{j=1}^{q}\beta_{ij}z_{i,t}^{*}x_{j,t} \\ &- \sum_{i=1}^{q}\sum_{j=i}^{q}\sum_{k=j}^{q}\theta_{ijk}x_{i,t}x_{j,t}x_{k,t} \\ &- \sum_{i=1}^{p-q}\sum_{j=1}^{q}\sum_{k=j}^{q}\beta_{ijk}z_{i,t}^{*}x_{j,t}x_{k,t} \\ &- \sum_{i=1}^{q}\sum_{j=i}^{q}\sum_{k=j}^{q}\sum_{l=k}^{q}\theta_{ijkl}x_{i,t}x_{j,t}x_{k,t}x_{l,t} \\ &- \sum_{i=1}^{p-q}\sum_{j=1}^{q}\sum_{k=j}^{q}\sum_{l=k}^{q}\beta_{ijkl}z_{i,t}^{*}x_{j,t}x_{k,t}x_{l,t} \Bigg\}^2. \end{aligned} \quad (30)$$

The LM statistic is given by (24) with

$$\begin{aligned} \widehat{\mathbf{h}}_t =& \nabla G(\mathbf{z}_t, \mathbf{x}_t; \widehat{\boldsymbol{\psi}}) \\ =& \Bigg[\mathbf{z}_t', \mathbf{z}_t' F\left[\widehat{\gamma}_1\left(\widehat{\tilde{\boldsymbol{\omega}}}_1'\mathbf{x}_t - \widehat{c}_1\right)\right], \\ & \widehat{\boldsymbol{\lambda}}_1'\mathbf{z}_t \frac{\partial F\left[\widehat{\gamma}_1\left(\widehat{\tilde{\boldsymbol{\omega}}}_1'\mathbf{x}_t - \widehat{c}_1\right)\right]}{\partial\gamma_1} \\ & \widehat{\boldsymbol{\lambda}}_1'\mathbf{z}_t \frac{\partial F\left[\widehat{\gamma}_1\left(\widehat{\tilde{\boldsymbol{\omega}}}_1'\mathbf{x}_t - \widehat{c}_1\right)\right]}{\partial\tilde{\omega}_{12}}, \ldots \\ & \widehat{\boldsymbol{\lambda}}_1'\mathbf{z}_t \frac{\partial F\left[\widehat{\gamma}_1\left(\widehat{\tilde{\boldsymbol{\omega}}}_1'\mathbf{x}_t - \widehat{c}_1\right)\right]}{\partial\tilde{\omega}_{1q}} \\ & \widehat{\boldsymbol{\lambda}}_1'\mathbf{z}_t \frac{\partial F\left[\widehat{\gamma}_1\left(\widehat{\tilde{\boldsymbol{\omega}}}_1'\mathbf{x}_t - \widehat{c}_1\right)\right]}{\partial c_1}\Bigg]'. \end{aligned}$$

Under $\mathbf{H}_0$ and Assumptions 1–3, the LM statistic has an asymptotic χ^2 distribution with m degrees of freedom and m is the number of nonlinear regressors in (29).

In the present case, Assumption 2 implies that the NCSTAR model under the null is ergodic.

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 $G(\mathbf{z}_t, \mathbf{x}_t; \widehat{\boldsymbol{\psi}})$. This has an adverse effect on the empirical size of the test. To circumvent this problem, we follow [22] and regress the residuals $\widehat{\varepsilon}_t$ on $\nabla G(\mathbf{z}_t, \mathbf{x}_t; \widehat{\boldsymbol{\psi}})$, and compute the residual sum of squares $SSR_0 = \sum_{t=1}^{T} \tilde{\varepsilon}_t^2$.
2) Regress $\tilde{\varepsilon}_t$ on $\widehat{\mathbf{h}}_t$ and $\widehat{\boldsymbol{\nu}}_t$. Compute the residual sum of squares $SSR_1 = \sum_{t=1}^{T} \widehat{v}_t^2$
3) Compute the χ^2 statistic

$$\mathrm{LM}_{\chi^2}^{hn} = T \frac{SSR_0 - SSR_1}{SSR_0} \tag{31}$$

or the F version of the test

$$\mathrm{LM}_{F}^{hn} = \frac{\frac{(SSR_0 - SSR_1)}{m}}{\frac{SSR_1}{(T-n-m)}} \tag{32}$$

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

Under $\mathbf{H}_0$, $\mathrm{LM}_{\chi^2}^{hn}$ is approximately distributed as a χ^2 with m degrees of freedom and LM_{F}^{hn} has approximately an F distribution with m and $T-n-m$ degrees of freedom.

When applying the test a special care should be taken. If $\widehat{\gamma}_1$ is very large, the gradient matrix becomes near-singular and the test statistic numerically unstable, which distorts the size of the test. The reason is that the vectors corresponding to the partial derivatives with respect to γ_i, $\boldsymbol{\omega}_i$, and c_i, respectively, tend to be almost perfectly linearly correlated. This is due to the fact that the time series of those elements of the gradient resemble dummy variables being constant most of the time and nonconstant simultaneously. In those cases, a solution is to omit the terms that depend on the derivatives of the logistic function from the regression in step 2; see [22] for a complete discussion. 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.

IV. Estimation Procedures and Parameter Inference

As selecting the number of hidden units requires estimation of neural network models, we now turn to this problem. A large number of algorithms for estimating the parameters of neural network type models are available in the literature. In this paper, we estimate the parameters of our NCSTAR model by maximum likelihood. This is because our modeling procedure is built on the use of statistical inference, and most of the algorithms applied to the estimation of neural network type models do not allow that. As a by-product, the use of maximum likelihood also makes it possible to obtain an idea of the uncertainty in the parameter estimates through asymptotic standard deviation estimates. It may be argued that maximum likelihood estimation of neural network models is most likely to lead to convergence problems, and that penalizing the log-likelihood function one way or the other is a necessary precondition for satisfactory results. Two things can be said in favor of maximum likelihood here. First, in this paper, model building proceeds from specific-to-general (small to large) models, so that estimation of unidentified or nearly unidentified models, a major reason for penalizing the log-likelihood, is avoided. Second, the starting values are chosen carefully.

In the case where ε_t is a Gaussian white noise with zero mean and finite variance, $\varepsilon_t \sim \mathrm{NID}(0, \sigma^2)$, maximum likelihood is equivalent to nonlinear least squares. Hence, the parameter vector $\boldsymbol{\psi}$ of (7) is estimated as

$$\begin{aligned} \widehat{\boldsymbol{\psi}} &= \underset{\boldsymbol{\psi}}{\operatorname{argmin}}\, Q_T(\boldsymbol{\psi}) \\ &= \underset{\boldsymbol{\psi}}{\operatorname{argmin}} \sum_{t=1}^{T} \left(y_t - G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})\right)^2 . \end{aligned} \tag{33}$$

Consider the following additional assumptions.

Assumption 4: The parameters satisfy the conditions $\beta_1 \leq \ldots \leq \beta_h$, $\gamma_i > 0$, $i = 1, \ldots, h$ and $\widehat{\omega}_{i1}$ is defined as in (8) for $i = 1, \ldots, h$.

Assumption 5: The NCSTAR model has no irrelevant hidden units.

Assumptions 4 and 5 guarantees the global identifiability of the NCSTAR model.

Theorem 1: Under Assumptions 1, 2, 4, and 5 the maximum likelihood estimator $\widehat{\boldsymbol{\psi}}$ is almost surely consistent for $\boldsymbol{\psi}$ and

$$\sqrt{T}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \xrightarrow{D} \mathrm{N}\left(0, -\operatorname*{plim}_{T \to \infty} \mathbf{A}(\boldsymbol{\psi}^{-1})\right) \tag{34}$$

where $\mathbf{A}(\boldsymbol{\psi}) = (1/\sigma^2 T)(\partial^2 Q_T(\boldsymbol{\psi}) / \partial\boldsymbol{\psi}\partial\boldsymbol{\psi}')$.

Proof: To prove consistency we use [47, Th. 3.5], showing that the assumptions stated therein are fulfilled.

Assumptions 2.1 and 2.3, related to the probability space and to the density functions, are trivial.

Let $q(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}) = [y_t - G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})]^2$. Assumption 3.1a states that for each $\boldsymbol{\psi} \in \Psi$, $-\mathrm{E}(q(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}))$ exists and is finite for $t = 1, \ldots, T$. Under Assumption 2 and the fact that ε_t is a zero mean normally distributed random variable with finite variance, hence, k-integrable, Assumption 3.1a in [47] follows.

Assumption 3.1b states that $-\mathrm{E}(q(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}))$ is continuous in Ψ, $t = 1, \ldots, T$. Let $\boldsymbol{\psi} \to \boldsymbol{\psi}^*$, since for any t, $G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})$ is continuous on Ψ, then $q(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}) \to q(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}^*)$, $\forall t$ (pointwise convergence). From the continuity of $G(\mathbf{z}_t, \mathbf{x}_t, \boldsymbol{\psi})$ on the compact set Ψ, we have uniform continuity and we obtain that $q(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})$ is dominated by an integrable function dF. Then, by Lebesgue's dominated convergence theorem, we get $\int q(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}) dF \to \int q(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}^*) dF$, and $\mathrm{E}(q(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}))$ is continuous.

Assumption 3.1c states that $-\mathrm{E}(q(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}))$ obeys the strong (weak) uniform law of large numbers (ULLN). [48, Lemma A2] guarantees that $\mathrm{E}(q(\mathbf{z}_t; \mathbf{x}_t; \boldsymbol{\psi}))$ obeys the strong law of large numbers. The set of hypothesis (b) of this lemma is satisfied:

1) we are working with an ergodic process;
2) from the continuity of $\mathrm{E}(q(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}))$ and from the compactness of Ψ we have that $\inf \mathrm{E}(q(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})) =$

$\mathrm{E}(q(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}^*))$ for $\boldsymbol{\psi}^* \in \Psi$, and with Assumption 3.1a in [47] we may guarantee that $\mathrm{E}(q(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}^*))$ exists and is finite, getting that $\inf \mathrm{E}(q(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})) > -\infty$.

Assumption 3.2 is related to the unique identifiability of $\boldsymbol{\psi}^*$. Under Assumptions 4 and 5 the NCSTAR model is globally identifiable.

To prove normality, we use [47, Th. 6.4] and check its assumptions.

Assumptions 2.1, 2.3, and 3.1 follow from the proof of consistency showed above.

Assumptions 3.2 and 3.6 follow from the fact that $G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})$ is continuously differentiable of order 2 on $\boldsymbol{\psi}$ in the compact space Ψ.

In order to check Assumptions 3.7a and 3.8a we have to prove that $\mathrm{E}(\nabla Q_T)(\boldsymbol{\psi})) < \infty$ and $\mathrm{E}(\nabla^2 Q_T(\boldsymbol{\psi})) < \infty$, $\forall T$. The expected gradient and the expected Hessian of $Q_T(\boldsymbol{\psi})$ are given by

$$\mathrm{E}\left(\nabla Q_T(\psi)\right) = -2\mathrm{E}\left(\nabla G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})\left(y_t - G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})\right)\right)$$

and

$$\mathrm{E}\left(\nabla^2 Q_T(\boldsymbol{\psi})\right) = 2\mathrm{E}\left(\nabla G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})\nabla G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})' - \nabla^2 G(\mathbf{z}_t, \mathbf{x}_t; \psi)\left(y_t - G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})\right)\right)$$

respectively.

Assumptions 3.7a and 3.8a follow considering the normality condition on ε_t, the properties of the function $G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})$, and the fact that $\nabla G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})$ and $\nabla^2 G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})$ contains at most terms of order $z_{i,t}x_{j,t}$, $i = 1, \ldots, p$, $i = 1, \ldots, q$.

Assumption 3.8b: Under Assumption 1, the fact that the function $G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})$ is continuous, and dominated convergence, Assumption 3.8b follows.

Assumption 3.8c: The proof of consistency and the ULLN from [48] yields the result.

Assumption 3.9: White's $A_T^* \equiv \mathrm{E}(\nabla^2 Q(\boldsymbol{\psi}^*)) = 2\mathrm{E}(\nabla G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}^*)\nabla' G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}^*))$ is $O(1)$ in our setup. Assumption 5, the properties of function $G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi})$, and the unique identification of $\boldsymbol{\psi}$ imply the nonsingularity of $\mathrm{E}(\nabla G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}^*)\nabla' G(\mathbf{z}_t, \mathbf{x}_t, \boldsymbol{\psi}^*))$.

Assumption 6.1: Using [49, Th. 2.4] we can show $2\boldsymbol{\xi}'\nabla G(\mathbf{z}_t, \mathbf{x}_t, \boldsymbol{\psi}^*)\varepsilon_t$ obeys the central limit theorem (CLT) for some $(r \times 1)$ vector $\boldsymbol{\xi}$, such that $\boldsymbol{\xi}'\boldsymbol{\xi} = 1$. Assumptions A(i) and A(iii) of [49] hold because ε_t is a Gaussian white noise. Assumption A(ii) holds with $V = 4\sigma^2\boldsymbol{\xi}'\mathrm{E}(\nabla G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}^*)\nabla' G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}^*))$. Furthermore, since any measurable transformation of mixing processes is itself mixing (see [49, Lemma 2.1]), $2\boldsymbol{\xi}'\nabla G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}^*)\varepsilon_t$ is a strong mixing sequence and obeys the CLT. By using the Cramér–Wold device $\nabla Q_T(\boldsymbol{\psi})$ also obeys the CLT with covariance matrix $B_T^* = 4\sigma^2\mathrm{E}(\nabla G(\mathbf{x}_t; \boldsymbol{\psi}^*)\nabla' G(\mathbf{z}_t, \mathbf{x}_t; \boldsymbol{\psi}^*)) = 2\sigma^2 A_T^*$ which is $O(1)$ and nonsingular. ■

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 [50] 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 interpolation 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 with several simulated data-sets, using the SQP algorithm turns the estimation process rather slow and does not improve the precision of the estimation.

A. Concentrated Least-Squares

In order to reduce the computational burden we apply concentrated maximum likelihood to estimate $\boldsymbol{\psi}$ as follows. Consider the ith iteration and rewrite model (7) as

$$\mathbf{y} = \mathbf{Z}(\boldsymbol{\phi})\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{35}$$

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

$$\mathbf{Z}(\boldsymbol{\phi}) = \begin{pmatrix} \mathbf{z}_1' & F\left[\gamma_1(\boldsymbol{\omega}_1'\mathbf{X}_1 - c_1)\right]\mathbf{z}_1' & \ldots & F\left[\gamma_h(\boldsymbol{\omega}_h'\mathbf{x}_1 - c_h)\right]\mathbf{z}_1' \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{z}_T' & F\left[\gamma_1(\boldsymbol{\omega}_1'\mathbf{x}_T - c_1)\right]\mathbf{z}_T' & \ldots & F\left[\gamma_h(\boldsymbol{\omega}_h'\mathbf{x}_T - c_h)\right]\mathbf{z}_T' \end{pmatrix}$$

with $\boldsymbol{\phi} = [\gamma_1, \ldots, \gamma_h, \tilde{\boldsymbol{\omega}}_1', \ldots, \tilde{\boldsymbol{\omega}}_h', c_1, \ldots, c_h]'$. Assuming $\boldsymbol{\phi}$ fixed, the parameter vector $\boldsymbol{\theta}$ can be estimated analytically by

$$\widehat{\boldsymbol{\theta}} = \left(\mathbf{Z}(\boldsymbol{\phi})'\mathbf{Z}(\boldsymbol{\phi})\right)^{-1}\mathbf{Z}(\boldsymbol{\phi})'\mathbf{y}. \tag{36}$$

The remaining parameters are estimated conditionally on $\boldsymbol{\theta}$ by applying the Levenberg–Marquadt algorithm which completes the ith iteration. This form of concentrated maximum likelihood was proposed by [51]. It reduces the dimensionality of the iterative estimation problem considerably.

B. Starting Values

The iterative optimization algorithms are often sensitive to the choice of starting values, and this is certainly so in the case of NCSTAR models. Besides, a NCSTAR model with h hidden units contains h, parameters, γ_i, $i = 1, \ldots, h$, that are not scale-free. Our first task is, thus, to rescale the input variables such that they have the standard deviation equal to unity. In the univariate NCSTAR case, this simply means normalizing y_t. If the model contains exogenous variables, they are normalized separately. This, together with the fact that $\|\tilde{\boldsymbol{\omega}}_h\| = 1$, gives us a basis for discussing the choice of starting values of γ_i, $i = 1, \ldots, h$. Furthermore, in the multivariate case normalizing generally makes numerical optimization easier as all variables have the same standard deviation. Then we draw K sets of values $\gamma_h^{(k)}$, $\tilde{\boldsymbol{\omega}}_h^{(k)}$ and $c_h^{(k)}$, $k = 1, \ldots, K$ for the parameters γ_h, $\tilde{\boldsymbol{\omega}}_h$, and c_h, compute the value of the log-likelihood, and select the values for which the log-likelihood is maximized. This is done as follows.

1) For $k = 1, \ldots, K$:
   a) construct a vector $\mathbf{v}_h^{(k)} = [v_{1h}^{(k)}, \ldots, v_{qh}^{(k)}]'$ such that $v_{1h}^{(k)} \in (0,1]$ and $v_{jh}^{(k)} \in [-1,1]$, $j = 2, \ldots, q$. The values for $v_{1h}^{(k)}$ are drawn from a uniform (0, 1] distribution and the ones for ${}^{(k)}v_{jh}$, $j = 2, \ldots, q$ from a uniform $[-1, 1]$ distribution;
   b) define $\tilde{\boldsymbol{\omega}}_h^{(k)} = \mathbf{v}_h^{(k)} \|\mathbf{v}_h^{(k)}\|^{-1}$, which guarantees $\|\tilde{\boldsymbol{\omega}}_h^{(k)}\| = 1$;
   c) let $c_h^{(k)} = \mathrm{med}(\tilde{\boldsymbol{\omega}}_h^{(k)\prime}\mathbf{x})$, where $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$.
2) Define a grid of N positive values $\gamma_h^{(n)}$, $n = 1, \ldots, N$ for the slope parameter. This need not be done randomly. As the changes in γ_h have a small effect of the slope when γ_h is large, only a small number of large values are required.
3) For $k = 1, \ldots, K$ and $n = 1, \ldots, N$, compute the value of $Q_T(\boldsymbol{\psi})$ for each combination of starting values. Choose the values of the parameters that maximize the concentrated log-likelihood function as starting values.

After selecting the starting values of the hth hidden unit we have to reorder the units if necessary in order to ensure that the identifying restrictions are satisfied.

Typically, K =1000 and N =20 will ensure good estimates of the parameters. We should stress, however, that K is a nondecreasing function of the number of input variables. If the latter is large we have to select a large K as well.

C. *Estimation of the Slope Parameter*

Concerning the slope parameter, we should stress that it is very difficult to have a precise estimate of γ_i, $i = 1, \ldots, h$. One of the reasons is that for large γ_i, the derivatives of the transition function, as already mentioned in Section III-C, 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. Another reason for not considering low values of the t-statistic is that under the null hypothesis $\gamma_i = 0$, because of the identification problem, it does not have the usual t-distribution. Again, see [22] for discussion.

V. Monte Carlo Experiment

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

Model 1:

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

Model 2:

$$\begin{aligned} y_t = {} & 1.8y_{t-1} - 1.06y_{t-2} + (0.02 - 0.90y_{t-1} + 0.795y_{t-2}) \\ & \times F\left(20(y_{t-1} - 0.02)\right) + \varepsilon_t,\ \varepsilon_t \sim \mathrm{NID}(0, 0.02^2). \end{aligned} \tag{38}$$

Model 3:

$$\begin{aligned} y_t = {} & -0.1 + 0.3y_{t-1} + 0.2y_{t-2} + (-1.2y_{t-1} + 0.5y_{t-2}) \\ & \times F\left(20(y_{t-1} + 0.6)\right) + (1.8y_{t-1} - 1.2y_{t-2}) \\ & \times F\left(20(y_{t-1} - 0.6)\right) + \varepsilon_t,\ \varepsilon_t \sim \mathrm{NID}(0, 0.5^2). \end{aligned} \tag{39}$$

Model 4:

$$\begin{aligned} y_t = {} & 0.5 + 0.8y_{t-1} - 0.2y_{t-2} + (-0.5 - 1.2y_{t-1} + 0.8y_{t-2}) \\ & \times F\left(11.31(0.7071y_{t-1} - 0.7071y_{t-2} - 0.1414)\right) \\ & + \varepsilon_t,\ \varepsilon_t \sim \mathrm{NID}(0, 0.5^2). \end{aligned} \tag{40}$$

Model 5:

$$\begin{aligned} y_t = {} & 0.5 + 0.8y_{t-1} - 0.2y_{t-2} + (1.5 - 0.6y_{t-1} - 0.3y_{t-2}) \\ & \times F\left(8.49(0.7071y_{t-1} - 0.7071y_{t-2} + 1.0607)\right) \\ & + (-0.5 - 1.2y_{t-1} + 0.7y_{t-2}) \\ & \times F\left(8.49(0.7071y_{t-1} - 0.7071y_{t-2} - 1.0607)\right) \\ & + \varepsilon_t,\ \varepsilon_t \sim \mathrm{NID}(0, 1^2). \end{aligned} \tag{41}$$

Model 1 is a stationary linear autoregressive model and is just used to check the empirical size of the linearity test. Models 2–5 are all different specifications of the NCSTAR model and have distinct dynamic properties. Considering Model 2, [3] discussed a similar specification. The only difference is that in his paper $\gamma = 100$ instead of 20. Model 2 is a logistic STAR model of order 2 with two extreme regimes. The "lower regime" of the process, corresponding to $F(20(y_{t-1} - 0.02)) = 0$, is such that the roots of the characteristic polynomial $g(w) = w^2 - 1.8w + 1.06$ are complex pair with modulus 1.03, so that the regime is explosive. The roots of the characteristic polynomial $g(w) = w^2 - 0.92w + 0.265$ corresponding to the "upper regime," $F(20(y_{t-1} - 0.02)) = 1$, are also a complex pair with modulus 0.51, so the regime is not explosive. As to the long-term behavior, the model has a unique stable stationary point, $y_\infty = 0.036$. Model 3 has three limiting regimes. The "lower regime," corresponding to $F(20(y_{t-1} + 0.06)) = 0$ and $F(20(y_{t-1} - 0.06)) = 0$, has a characteristic polynomial with roots equal to 0.62 and -0.32, so the regime is stationary. The characteristic equation in the "middle regime," $F(20(y_{t-1} + 0.06)) = 1$ and $F(20(y_{t-1} - 0.06)) = 0$, has roots -1.4 and 0.5, thus, the regime is explosive. Finally, the "upper regime," $F(20(y_{t-1} + 0.06)) = 1$ and $F(20(y_{t-1} - 0.06)) = 1$, is also explosive with the roots of the characteristic polynomial being 1.33 and -0.43. Considering the long-term behavior, the model has a limit cycle with a period of 8 time units. Model 4 has two extreme regimes. The first one, $F(11.31(0.7071y_{t-1} - 0.7071y_{t-2} - 0.1414)) = 0$, has a characteristic equation with a complex pair of roots with modulus 0.45, so the regime is stable. The characteristic polynomial of the second regime, $F(11.31(0.7071y_{t-1} - 0.7071y_{t-2} - 0.1414)) = 1$, has roots -1 and 0.6, so the regime is nonstationary. However, considering the long-term behavior, the process has two stable stationary points, 0.38 and -0.05. Finally, Model 6 has three limit regimes. In the "lower regime," $F(8.49(0.7071y_{t-1} - 0.7071y_{t-2} + 1.0607)) = 0$

TABLE I
MEDIAN AND MAD OF THE NLS ESTIMATES OF THE PARAMETERS. TRUE VALUES BETWEEN PARENTHESES

	100 observations							
Parameter	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
$\hat{\hat{\omega}}_{11}$	–	–	–	–	0.7310 (0.7071)	0.0906	0.7193 (0.7071)	0.0531
$\hat{\hat{\omega}}_{21}$	–	–	–	–	–	–	0.7160 (0.7071)	0.0283
$\hat{\hat{\omega}}_{12}$	–	–	–	–	-0.6829 (-0.7071)	0.0956	-0.6938 (-0.7071)	0.0553
$\hat{\hat{\omega}}_{22}$	–	–	–	–	–	–	-0.6955 (-0.7071)	0.0289
$\hat{c}_1$	0.0236 (0.02)	0.0344	-0.5578 (-0.6)	0.1869	0.1422 (0.1414)	0.1261	-0.3252 (-1.0607)	1.0763
$\hat{c}_2$	–	–	0.5853 (0.6)	0.0785	–	–	1.0971 (-1.0607)	0.3120

	500 observations							
Parameter	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
$\hat{\hat{\omega}}_{11}$	–	–	–	–	0.7162 (0.7071)	0.0335	0.7103 (0.7071)	0.0134
$\hat{\hat{\omega}}_{21}$	–	–	–	–	–	–	0.7074 (0.7071)	0.0036
$\hat{\hat{\omega}}_{12}$	–	–	–	–	-0.6979 (-0.7071)	0.0338	-0.7039 (-0.7071)	0.0133
$\hat{\hat{\omega}}_{22}$	–	–	–	–	–	–	-0.7068 (0.7071)	0.0036
$\hat{c}_1$	0.0202 (0.02)	0.0289	-0.6038 (-0.6)	0.0285	0.1469 (0.1414)	0.0433	-1.0307 (-1.0607)	0.1387
$\hat{c}_2$	–	–	0.6025 (0.6)	0.0166	–	–	1.0635 (1.0607)	0.0480

and $F(8.49(0.7071y_{t-1} - 0.7071y_{t-2} - 1.0607)) = 0$, the characteristic equation has a complex pair of roots with modulus 0.45. The "middle regime," $F(8.49(0.7071y_{t-1} - 0.7071y_{t-2} + 1.0607)) = 1$ and $F(8.49(0.7071y_{t-1} - 0.7071y_{t-2} - 1.0607)) = 0$, is stable and the characteristic equation has also a complex pair of roots with modulus 0.71. The "upper regime," $F(8.49(0.7071y_{t-1} - 0.7071y_{t-2} + 1.0607)) = 1$ and

TABLE II
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 POLYNOMIAL EXPANSION

	100 observations					
Model	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

	500 observations					
Model	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 III
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

	100 observations					
Model	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

	500 observations					
Model	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

$F(8.49(0.7071y_{t-1} - 0.7071y_{t-2} - 1.0607)) = 1$, has a characteristic equation with roots 0.56 and -0.36. The process has only one stable stationary point, $y_\infty = 0.99$.

A. 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 I shows the median and the median absolute deviation (MAD) of the estimates, defined as

$$\mathrm{MAD}(\hat{\psi}) = \mathrm{median}\left(\left|\hat{\psi} - \mathrm{median}(\hat{\psi})\right|\right). \tag{42}$$

The true value of the parameters are shown between parentheses.

Reporting the median and MAD was suggested by [52] 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. Considering Model 2, it is interesting to notice that γ_1 is strongly overestimated when only 100 observations are considered. When the number of observations is increased the estimation of the parameter γ_1 improves substantially.

B. Model Selection Tests

1) Variable Selection: Tables II and III show, respectively, the results of the variable selection procedure using a third-order polynomial expansion in (14) and using only the linear term (no cross-products) in (14). 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 $\mathbf{z}_t$. The cases where the number of variables is correct but the combination is not the correct one appear under the heading "U."

Observing Table II, 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 polynomial expansion than in the linear case (Table III). Further simulation results can be found in [23].

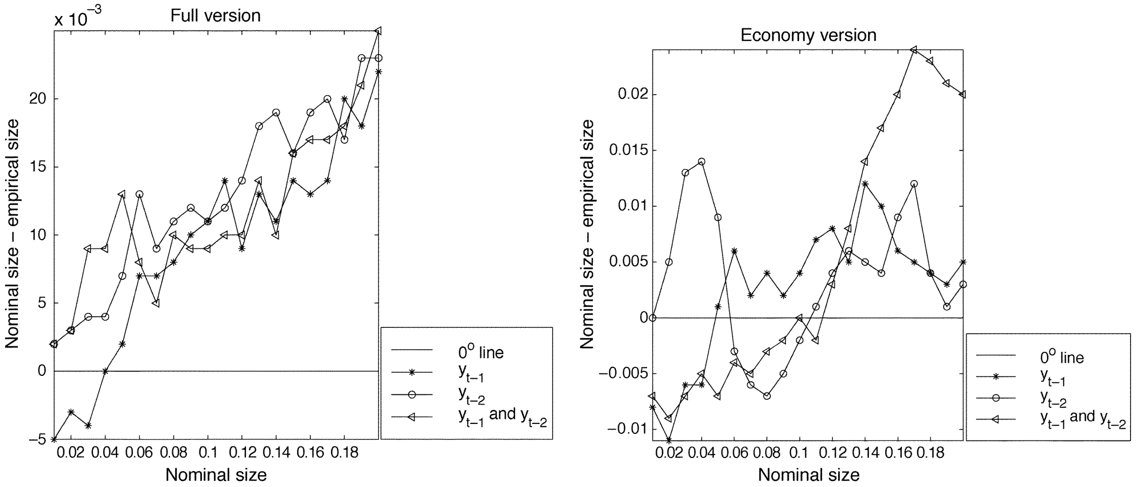
2) Linearity Tests: Concerning the size of the linearity test developed in Section III-B, hereafter LM_F^l and its "economy version," $\mathrm{LM}_F^{l,e}$, we show the plot of the deviation of empirical size from the nominal size versus the nominal size. The results are shown in Fig. 1. The results 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 Figs. 2–5.

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.

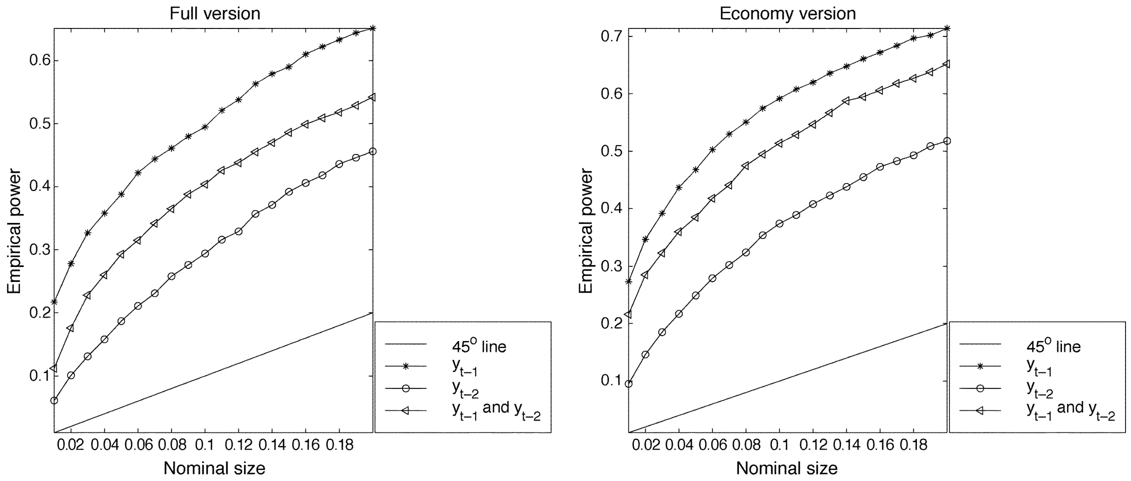
In Figs. 2 and 3 we can observe that the power of the test improves when we select y_{t-1} as the transition variable and in Fig. 4 the power increases when we use y_{t-1} and y_{t-2} as transition variables. With model (41) the power is always 1 when the transition variable is correctly chosen.

3) Tests for the Number of Hidden Units: To study the behavior 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 Figs. 6 and 7. As we can see the test is conservative with the empirical size well below the corresponding nominal one. However, the test has good power when model (38) is considered. An interesting point to mention is the relatively low power of the additional hidden unit test when model (40) is considered, despite the fact that the power of the linearity test is always one when the correct transition variables are selected; see Fig. 4. A possible explanation is that although the model is strongly



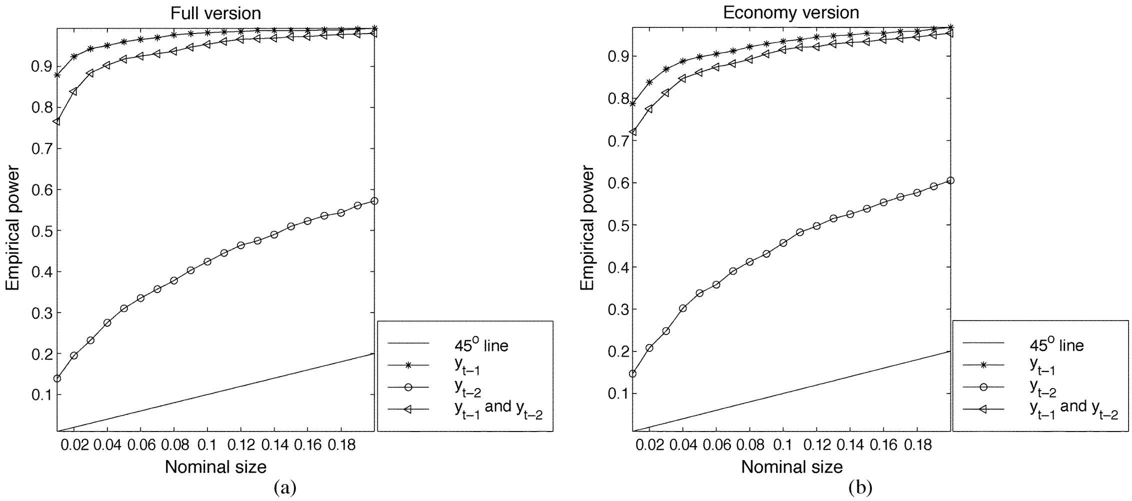
(a) (b)

Fig. 1. 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). (a) Refers to the LM_F^l test. (b) Refers to the $\mathrm{LM}_F^{l,e}$ test.



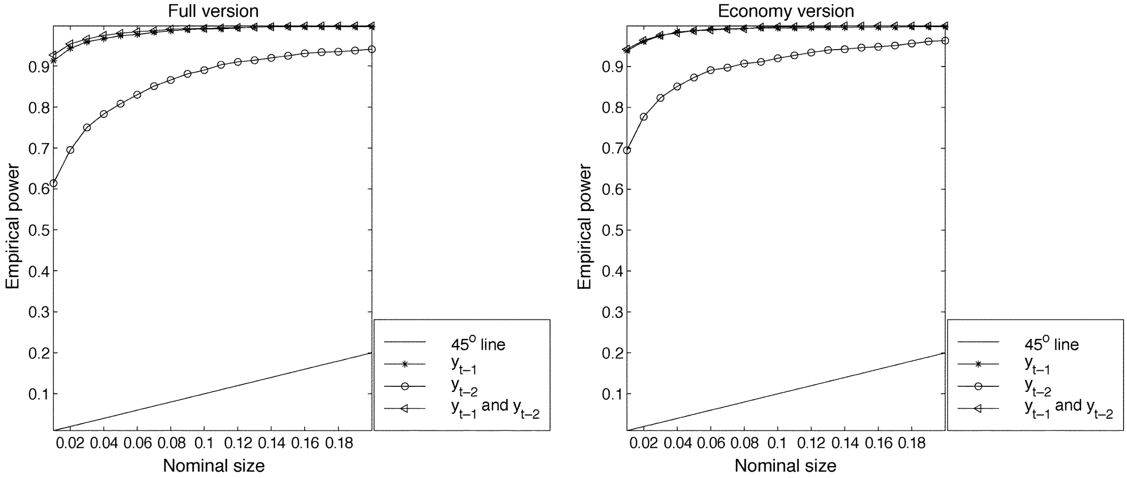
(a) (b)

Fig. 2. 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 $\mathrm{LM}_F^{l,e}$ test.



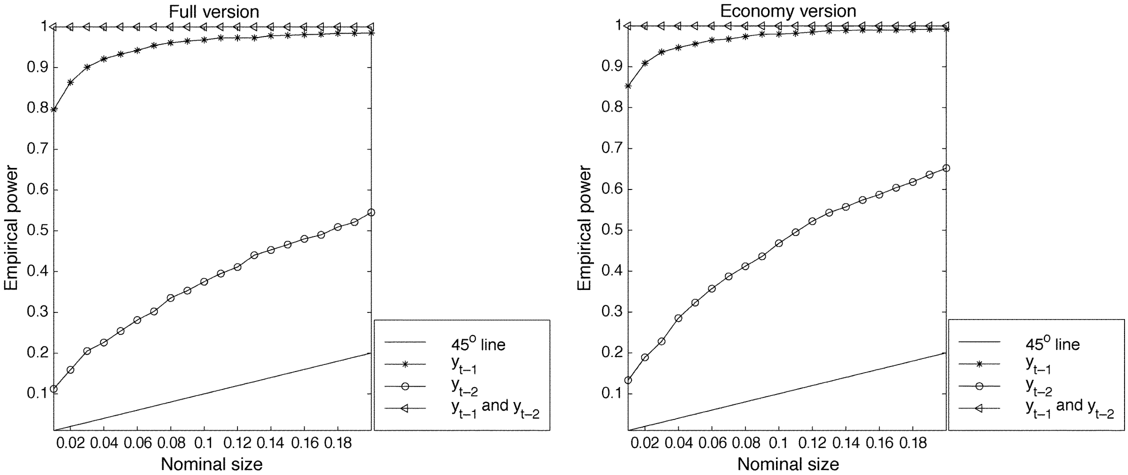
(a) (b)

Fig. 3. Power-size curve of the linearity tests at sample size of 100 observations based on 1000 replications of model (39). (a) Refers to the LM_F^l test. (b) Refers to the $\mathrm{LM}_F^{l,e}$ test.



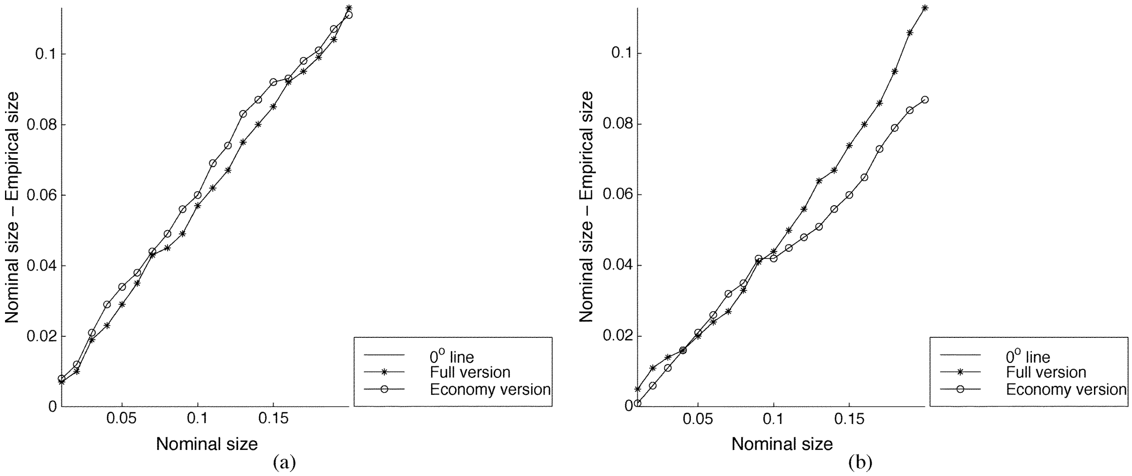
(a) (b)

Fig. 4. Power-size curve of the linearity tests at sample size of 100 observations based on 1000 replications of model (40). (a) Refers to the LM$^{l}_{F}$ test. (b) Refers to the LM$^{l,e}_{F}$ test.



(a) (b)

Fig. 5. Power-size curve of the linearity tests at sample size of 100 observations based on 1000 replications of model (41). (a) Refers to the LM$^{l}_{F}$ test. (b) Refers to the LM$^{l,e}_{F}$ test.



(a) (b)

Fig. 6. 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). (a) Refers to model (38). (b) Refers to model (40).

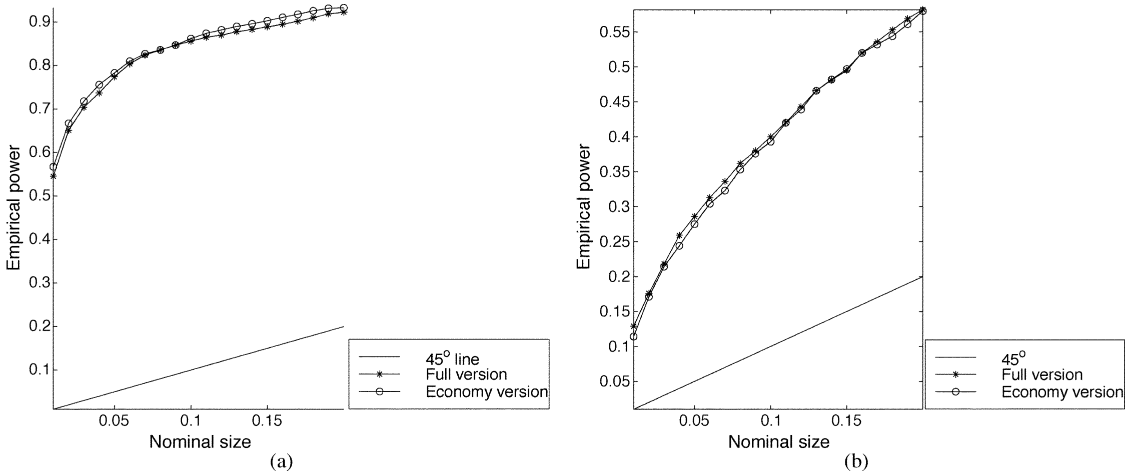


Fig. 7. Power-size curve of the additional hidden unit tests at sample size of 100 observations based on 1000 replications of model (39) and (41). (a) Refers to model (39). (b) Refers to model (41).

nonlinear, reason that makes the power being always one, it has more parameters than model (38), imposing a large number of regressors in the additional hidden unit test when the alternative hypothesis is considered even with the economy version of the test. For that reason, the test is conservative in small samples. As the sample sizes increases, the problem will vanish.

VI. Examples

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

A. 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 Northwest Canada over the period 1821–1934. For further details and a background history see Tong [9, Ch. 7]. Some previous analyses of this series can be found in [3], [9], [13], [17], and [53]. We report only results for in-sample fitting because the number of observations is rather small and also because most of the previous studies in the literature have only considered in-sample analysis.

We start selecting the variables of the model among the first 7 lags of the time series. With the procedure described in Section III-A 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 $(p - \text{value} = 0.0002)$.

The sequence of including hidden units is discontinued after adding the first hidden unit and the estimated model is

$$\begin{aligned} y_t = {} & \underset{(0.18)}{0.49} + \underset{(0.07)}{1.25}\, y_{t-1}\, \underset{(0.10)}{0.37}\, y_{t-2} \\ & - \left(\underset{(2.34)}{1.05} + \underset{(0.18)}{0.42}\, y_{t-1} - \underset{(0.56)}{0.25}\, y_{t-2}\right) \\ & \times F\left[\underset{(5.76)}{11.02}\left(y_{t-2} - \underset{(1.51)}{3.34}\right)\right] + \widehat{\varepsilon}_t. \\ \widehat{\sigma}_\varepsilon = {} & 0.198 \quad \frac{\widehat{\sigma}_\varepsilon}{\widehat{\sigma}_L} = 0.87 \quad R^2 = 0.88 \quad JB = 0.33 \\ & \text{ARCH}(1) = 0.27 \quad \text{ARCH}(2) = 0.48 \\ & \text{ARCH}(3) = 0.60 \quad \text{ARCH}(4) = 0.48 \end{aligned} \tag{43}$$

where $\widehat{\sigma}_\varepsilon$ is the residual standard deviation, $\widehat{\sigma}_\varepsilon/\widehat{\sigma}_L$ is the ratio between the standard deviation of the residuals from the nonlinear model and a linear AR(2) model, R^2 is the determination coefficient, JB is the p-value of the Jarque-Bera test of normality, and ARCH(j), $j = 1, \ldots, 4$, is the p-value of the LM test of no autoregressive conditional heteroskedasticity (ARCH) against ARCH of order j.

The estimated residual standard deviation $(\widehat{\sigma}_\varepsilon = 0.198)$ is smaller than in other models that use only the first two lags as variables. For example, the nonlinear model proposed by Tong [9, p. 410], has a residual standard deviation of 0.222, the exponential autoregressive (EXPAR) model proposed by [53] has $\widehat{\sigma}_\varepsilon = 0.208$, and for the single-index coefficient regression model of [13], $\widehat{\sigma}_\varepsilon = 0.200$. [3] found a better result $(\widehat{\sigma}_\varepsilon = 0.187)$, but he included up to lag 11 in his model. Table IV shows the results of the misspecification tests developed in [25]. They are Lagrange Multiplier tests for qth-order serial correlation in the residuals against no serial correlation, parameter constancy against smoothing changing ones, and constant error variance. The results indicate no model misspecification.

B. Example 2: Annual Sunspot Numbers

In this example we consider the annual sunspot numbers over the period 1700–1998. The observations for the period

TABLE IV
RESULTS OF MISSPECIFICATION TESTS OF THE ESTIMATED NCSTAR MODEL

				Test for q-th order serial correlation								
q	1	2	3	4	5	6	7	8	9	10	11	12
p-value	0.45	0.34	0.26	0.35	0.48	0.60	0.62	0.67	0.70	0.63	0.43	0.54

	Test for parameter constancy	Test for constant variance	Test for 2nd hidden unit
p-value	0.88	0.18	0.18

TABLE V
RESULTS OF MISSPECIFICATION TESTS OF THE ESTIMATED NCSTAR MODEL

				Test for q-th order serial correlation								
q	1	2	3	4	5	6	7	8	9	10	11	12
p-value	0.08	0.18	0.14	0.19	0.28	0.35	0.06	0.10	0.09	0.11	0.15	0.16

	Test for parameter constancy	Test for constant variance	Test for 4th hidden unit
p-value	0.83	0.02	0.04

1700–1979 were used to estimate the model and the remaining were used to forecast evaluation. We adopted the same transformation as in [9], $y_t = 2[\sqrt{(1+Nt)} - 1]$, 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. However, the residuals of the estimated linear AR model are strongly autocorrelated. The serial correlation is removed by also including y_{t-3} in the set of selected variables. 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 sequence of including hidden units is discontinued after adding the third hidden unit and the final estimated model is

$$
\begin{aligned}
y_t = & -\underset{(1.85)}{4.64} + \underset{(0.38)}{1.04}\, y_{t-1} + \underset{(0.37)}{0.13}\, y_{t-2} - \underset{(0.27)}{0.08}\, y_{t-3} + \underset{(0.12)}{0.36}\, y_{t-7} \\
& + \Bigg(- \underset{(1.62)}{0.06} + \underset{(0.22)}{0.36}\, y_{t-1} - \underset{(0.36)}{0.34}\, y_{t-2} \\
& \qquad - \underset{(0.18)}{0.08}\, y_{t-3} + \underset{(0.07)}{0.13}\, y_{t-7} \Bigg) \\
& \times F\left[\underset{(105.26)}{256.62} \left(\underset{(-)}{0.32}\, y_{t-1} - \underset{(0.70)}{0.95}\, y_{t-2} + \underset{(4.20)}{6.05} \right) \right] \\
& + \Bigg(\underset{(1.37)}{0.80} - \underset{(0.25)}{0.14}\, y_{t-1} - \underset{(0.39)}{0.32}\, y_{t-2} \\
& \qquad + \underset{(0.18)}{0.58}\, y_{t-3} + \underset{(0.08)}{0.12}\, y_{t-7} \Bigg) \\
& \times F\left[\underset{(69.12)}{129.15} \left(\underset{(-)}{0.59}\, y_{t-1} - \underset{(0.39)}{0.80}\, y_{t-2} + \underset{(0.30)}{0.62} \right) \right] \\
& + \Bigg(\underset{(1.71)}{5.38} - \underset{(0.38)}{0.08}\, y_{t-1} + \underset{(0.37)}{0.06}\, y_{t-2} \\
& \qquad - \underset{(0.27)}{0.25}\, y_{t-3} - \underset{(0.12)}{0.39}\, y_{t-7} \Bigg) \\
& \times F\left[\underset{(1.85)}{3.22} \left(\underset{(-)}{0.99}\, y_{t-1} - \underset{(0.09)}{0.11}\, y_{t-2} - \underset{(2.54)}{3.99} \right) \right] + \widehat{\varepsilon}_t. \qquad (44)
\end{aligned}
$$

$$
\begin{aligned}
& \widehat{\sigma}_\varepsilon = 1.696 \quad R^2 = 0.91 \quad JB = 0.001 \\
& \quad \text{ARCH}(1) = 0.76 \quad \text{ARCH}(2) = 0.94 \\
& \quad \text{ARCH}(3) = 0.96 \quad \text{ARCH}(4) = 0.54. \qquad (45)
\end{aligned}
$$

As in the previous example, the value of the estimated in-sample residual standard deviation ($\widehat{\sigma}_\varepsilon = 1.696$) is smaller than other nonlinear models. For example, [13] estimated a model where $\widehat{\sigma}_\varepsilon = 1.772$ and Tong [9, 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, $F(\boldsymbol{\omega}_i'\mathbf{x}_t - \beta_i)$, $i = 1, \ldots, 3$, is

$$
\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} 1 & 0.59 & -0.45 \\ 0.59 & 1. & 0.06 \\ -0.45 & 0.06 & 1 \end{pmatrix} \qquad (46)
$$

indicating that there is no irrelevant neurons in the model as none of the correlations is close to unity in absolute value. Furthermore, the results of the misspecification tests of model (44) in Table V indicate no model misspecification.

In order to assess the out-of-sample performance of the estimated model we compare our forecasting results with the ones obtained from the two SETAR models, the one reported in Tong [9, p. 420] and the other in [54], an artificial neural network (ANN) model with five hidden neurons and the first nine lags as input variables, estimated with Bayesian regularization [55], [56], and a linear model with lags selected using SBIC. The SETAR model estimated by [54] is one in which the threshold variable is a nonlinear function of lagged values of the time series whereas it is a single lag in Tong's model.

Table VI shows the one-step ahead forecasts, their root mean square errors, and mean absolute errors (MAEs) for the annual number of sunspots for the period 1980–1998.

Both the root mean squared errors (RMSE) and the MAEs of our model are lower than the ones of the other models considered here.

VII. CONCLUSION

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

TABLE VI
ONE-STEP AHEAD FORECASTS, THEIR ROOT MEAN SQUARE ERRORS, AND MEAN ABSOLUTE ERRORS FOR THE ANNUAL NUMBER OF SUNSPOTS FROM A SET OF TIME SERIES MODELS, FOR THE PERIOD 1980–1998

Year	Observation	NCSTAR		NN model		SETAR model [9]		SETAR model [54]		AR	
		Forecast	Error	Forecast	Error	Forecast	Error	Forecast	Error	Forecast	Error
1980	154.6	136.0	18.62	138.1	16.4	160.9	-6.4	134.3	20.3	159.8	-5.2
1981	140.4	133.7	6.71	114.3	26.1	137.2	3.2	125.4	15.0	123.3	17.1
1982	115.9	102.6	13.33	94.3	21.6	99.0	16.9	99.3	16.6	99.6	16.3
1983	66.6	80.9	-14.31	76.7	-10.1	75.9	-9.4	85.0	-18.4	78.9	-12.3
1984	45.9	40.1	4.81	40.8	5.1	35.6	10.2	41.3	4.7	33.9	12.0
1985	17.9	22.9	-5.02	26.1	-8.2	24.2	-6.3	29.8	-11.9	29.3	-11.4
1986	13.4	8.08	5.32	13.7	-0.3	10.7	2.7	9.8	3.6	10.7	2.7
1987	29.4	11.0	18.40	20.4	9.0	20.1	9.3	16.5	12.9	23.0	6.4
1988	100.2	75.6	24.56	79.7	20.5	54.4	45.7	66.4	33.8	61.2	38.9
1989	157.6	168.4	-10.79	170.6	-13.0	155.7	1.9	121.8	35.8	159.2	-1.6
1990	142.6	156.3	-13.68	157.6	-14.9	156.4	-13.8	152.5	-9.9	175.5	-32.9
1991	145.7	119.8	25.84	118.7	26.9	93.2	52.4	123.7	22.0	119.1	26.6
1992	94.3	104.3	-9.97	98.8	-4.5	111.3	-16.9	115.9	-21.7	118.9	-24.6
1993	54.6	66.8	-12.22	71.0	-16.4	67.8	-13.2	69.2	-14.6	57.9	-3.3
1994	29.9	29.0	0.91	27.8	2.0	27.0	2.9	35.7	-5.8	29.9	-0.1
1995	17.5	13.7	3.81	22.6	-5.1	18.4	-0.9	18.9	-1.4	17.6	-0.1
1996	8.6	7.7	0.87	12.0	-3.4	18.0	-9.4	11.6	-3.0	15.7	-7.1
1997	21.5	25.0	-3.47	18.2	3.3	12.3	9.2	11.8	9.7	16.0	5.5
1998	64.3	66.3	-2.03	70.4	-6.1	46.7	17.6	58.5	5.8	52.5	11.8
RMSE			12.66		13.8		18.7		16.9		16.5
MAE			10.23		11.2		13.1		14.1		12.4

model nests several nonlinear models, such as, for example, the SETAR, STAR, and AR-NN 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 seems to perform better than the linear model and other nonlinear specifications considered in the paper. 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.

ACKNOWLEDGMENT

The authors would like to thank T. Teräsvirta, G. Rech, C. Pedreira, and two anonymous referees for valuable comments. Part of this work was done while M. C. Medeiros was a visiting graduate student at the Department of Economic Statistics, Stockholm School of Economics, whose kind hospitality is gratefully acknowledged.

REFERENCES

[1] K. S. Chan and H. Tong, "On estimating thresholds in autoregressive models," *J. Time Series Anal.*, vol. 7, pp. 179–190, 1986.

[2] R. Luukkonen, P. Saikkonen, and T. Teräsvirta, "Testing linearity against smooth transition autoregressive models," *Biometrika*, vol. 75, pp. 491–499, 1988.

[3] T. Teräsvirta, "Specification, estimation, and evaluation of smooth transition autoregressive models," *J. Amer. Statist. Assoc.*, vol. 89, no. 425, pp. 208–218, 1994.

[4] D. van Dijk, T. Teräsvirta, and P. H. Franses, "Smooth transition autoregressive models—a survey of recent developments," *Econometric Rev.*, vol. 21, pp. 1–47, 2002.

[5] D. W. Bacon and D. G. Watts, "Estimating the transition between two intersecting lines," *Biometrika*, vol. 58, pp. 525–534, 1971.

[6] S. M. Goldfeld and R. Quandt, *Nonlinear Methods in Econometrics*. Amsterdam, The Netherlands: North Holland, 1972.

[7] A. Veiga and M. Medeiros, "A hybrid linear-neural model for time series forecasting," in *Proc. NEURAP*, Marseilles, 1998, pp. 377–384.

[8] M. C. Medeiros and A. Veiga, "A hybrid linear-neural model for time series forecasting," *IEEE Trans. Neural Netw.*, vol. 11, no. 6, pp. 1402–14 012, Nov. 2000.

[9] H. Tong, *Non-linear Time Series: A Dynamical Systems Approach*, ser. Oxford Statistical Science Series. Oxford, U.K.: Oxford Univ. Press, 1990, vol. 6.

[10] F. Leisch, A. Trapletti, and K. Hornik, "Stationarity and stability of autoregressive neural network processes," in *Advances in Neural Information Processing Systems*, M. S. Kearns, S. A. Solla, and D. A. Cohn, Eds. Cambridge, MA: MIT Press, 1999, vol. 11.

[11] A. Trapletti, F. Leisch, and K. Hornik, "Stationary and integrated autoregressive neural network processes," *Neural Computat.*, vol. 12, pp. 2427–2450, 2000.

[12] R. Chen and R. S. Tsay, "Functional coefficient autoregressive models," *J. Amer. Statist. Assoc.*, vol. 88, pp. 298–308, 1993.

[13] Y. Xia and W. K. Li, "On single-index coefficient regression models," *J. Amer. Statist. Assoc.*, vol. 94, no. 448, pp. 1275–1285, 1999.

[14] D. van Dijk and P. H. Franses, "Modeling multiple regimes in the business cycle," *Macroeconom. Dynam.*, vol. 3, no. 3, pp. 311–340, 1999.

[15] N. Öcal and D. Osborn, "Business cycle nonlinearities in uk consumption and production," *J. Appl. Econometrics*, vol. 15, pp. 27–43, 2000.

[16] S. J. Cooper, "Multiple regimes in US output fluctuations," *J. Bus. Econom. Statist.*, vol. 16, no. 1, pp. 92–100, 1998.

[17] R. Tsay, "Testing and modeling threshold autoregressive processes," *J. Amer. Statist. Assoc.*, vol. 84, pp. 431–452, 1989.

[18] G. C. Tiao and R. S. Tsay, "Some advances in nonlinear and adaptive modeling in time-series," *J. Forecasting*, vol. 13, pp. 109–131, 1994.

[19] P. H. Franses and R. Paap, "Censored latent effects autoregression with an application to us unemployment," Econometric Institute, Erasmus Univ., Econometric Inst. Rep. 9841/A, 1999.

[20] P. A. W. Lewis and J. G. Stevens, "Nonlinear modeling of time series using multivariate adaptive regression splines," *J. Amer. Statist. Assoc.*, vol. 86, pp. 864–877, 1991.

[21] T. Astatkie, D. G. Watts, and W. E. Watt, "Nested threshold autoregressive (NeTAR) models," *Int. J. Forecasting*, vol. 13, pp. 105–116, 1997.

[22] Ø. Eitrheim and T. Teräsvirta, "Testing the adequacy of smooth transition autoregressive models," *J. Econometrics*, vol. 74, pp. 59–75, 1996.

[23] G. Rech, T. Teräsvirta, and R. Tschernig, "A simple variable selection technique for nonlinear models," *Commun. Statist., Theory and Methods*, vol. 30, pp. 1227–1241, 2001.

[24] M. C. Medeiros, T. Teräsvirta, and G. Rech, Building neural network models for time series: A statistical approach, Stockholm School Economics, ser. Working Paper Series in Economics and Finance 508, 2002.

[25] M. C. Medeiros and A. Veiga, "Diagnostic checking in a flexible nonlinear time series model," *J. Time Series Anal.*, vol. 24, pp. 461–482, 2003.

[26] H. J. Sussman, "Uniqueness of the weights for minimal feedforward nets with a given input-output map," *Neural Netw.*, vol. 5, pp. 589–593, 1992.

[27] V. Kurková and P. C. Kainen, "Functionally equivalent feedforward neural networks," *Neural Computat.*, vol. 6, pp. 543–558, 1994.

[28] J. T. G. Hwang and A. A. Ding, "Prediction intervals for artificial neural networks," *J. Amer. Statist. Assoc.*, vol. 92, no. 438, pp. 109–125, 1997.

[29] U. Anders and O. Korn, "Model selection in neural networks," *Neural Netw.*, vol. 12, pp. 309–323, 1999.

[30] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.

[31] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.

[32] R. Tschernig and L. Yang, "Nonparametric lag selection for time series," *J. Time Series Anal.*, vol. 21, pp. 457–487, 2000.

[33] P. Vieu, "Order choice in nonlinear autoregressive models," *Statist.*, vol. 26, pp. 307–328, 1995.

[34] T. Tjøstheim and B. Auestad, "Nonparametric identification of nonlinear time series: selecting significant lags," *J. Amer. Statist. Assoc.*, vol. 89, pp. 1410–1419, 1994.

[35] Q. Yao and H. Tong, "On subset selection in nonparametric stochastic regression," *Statistica Sinica*, vol. 4, pp. 51–70, 1994.

[36] B. Auestad and D. Tjøstheim, "Identification of nonlinear time series: first order characterization and order determination," *Biometrika*, vol. 77, pp. 669–687, 1990.

[37] H. Royden, *Real Analysis*. New York: Macmillan, 1963.

[38] R. B. Davies, "Hypothesis testing when the nuisance parameter in present only under the alternative," *Biometrika*, vol. 64, pp. 247–254, 1977.

[39] ——, "Hypothesis testing when the nuisance parameter in present only under the alternative," *Biometrika*, vol. 74, pp. 33–44, 1987.

[40] P. Saikkonen and R. Luukkonen, "Lagrange multiplier tests for testing nonlinearities in time series models," *Scandinavian J. Statist.*, vol. 15, pp. 55–68, 1988.

[41] T. Teräsvirta, C. F. Lin, and C. W. J. Granger, "Power of the neural network linearity test," *J. Time Series Anal.*, vol. 14, no. 2, pp. 309–323, 1993.

[42] L. G. Godfrey, *Misspecification Tests in Econometrics*, 2nd ed, ser. Econometric Society Monographs. Cambridge, U.K.: Cambridge Univ. Press, 1988, vol. 16.

[43] C. W. J. Granger and T. Teräsvirta, *Modeling Nonlinear Economic Relationships*. Oxford, U.K.: Oxford Univ. Press, 1993.

[44] T. Teräsvirta and I. Mellin, "Model selection criteria and model selection tests in regression models," *Scandinavian J. Statist.*, vol. 13, pp. 159–171, 1986.

[45] T. Teräsvirta and C.-F. J. Lin, "Determining the number of hidden units in a single hidden-layer neural network model," Bank of Norway, 1993.

[46] E. Dudewicz and S. Mishra, *Modern Mathematical Statistics*. New York: Wiley, 1988.

[47] H. White, *Estimation, Inference and Specification Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1994.

[48] B. M. Pötscher and I. R. Prucha, "A class of partially adaptive one-step m-estimators for the nonlinear regression model with dependent observations," *J. Econometrics*, vol. 32, pp. 219–251, 1986.

[49] H. White and I. Domowitz, "Nonlinear regression with dependent observations," *Econometrica*, vol. 52, pp. 143–162, 1984.

[50] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 1995.

[51] S. Leybourne, P. Newbold, and D. Vougas, "Unit roots and smooth transitions," *J. Time Series Anal.*, vol. 19, pp. 83–97, 1998.

[52] D. van Dijk, "Smooth transition models: Extensions and outlier robust inference," Ph.D. dissertation, Tinbergen Inst., Rotterdam, The Netherlands, 1999.

[53] T. Ozaki, "The statistical analysis of perturbed limit cycle process using nonlinear time series models," *Journal of Time Series Analysis*, vol. 3, pp. 29–41, 1982.

[54] R. Chen, "Threshold variable selection in open-loop threshold autoregressive models," *J. Time Series Anal.*, vol. 16, no. 5, pp. 461–481, 1995.

[55] D. J. C. MacKay, "Bayesian interpolation," *Neural Computat.*, vol. 4, pp. 415–447, 1992.

[56] ——, "A practical bayesian framework for backpropagation networks," *Neural Computat.*, vol. 4, pp. 448–472, 1992.

Marcelo C. Medeiros was born in Rio de Janeiro, Brazil, in 1974. He received the B.S. degree in electrical engineering (systems) and the M.Sc. and Ph.D. degrees in electrical engineering (statistics) from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil, in 1996, 1998, and 2000, respectively.

His main research interest is nonlinear time series analysis and the link between econometrics and machine learning.

Álvaro Veiga was born in Florianópolis, Brazil, in 1955. He received the B.S. degree in electrical engineering (systems) from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil, in 1978, the M.Sc. degree from COPPE-UFRJ, Rio de Janeiro, Brazil, in 1982, and the Ph.D. degree from École Nationale Supérieure des Télécommunications, Paris, France, in 1989.

His main research interests include nonlinear time series modeling, finance, and econometrics.