TEXTO PARA DISCUSSÃO

No. 470

Local-global neural networks: a new approach for nonlinear time series modelling

Mayte Suarez Fariñas
Carlos E. Pedreira
Marcelo C. Medeiros

DEPARTAMENTO DE ECONOMIA
www.econ.puc-rio.br

DEPARTAMENTO DE ECONOMIA
PUC-RIO

TEXTO PARA DISCUSSÃO
Nº. 470

LOCAL-GLOBAL NEURAL NETWORKS: A NEW APPROACH FOR NONLINEAR TIME SERIES MODELLING

MAYTE SUAREZ FARIÑAS
CARLOS E. PEDREIRA
MARCELO C. MEDEIROS

JANEIRO 2003

LOCAL-GLOBAL NEURAL NETWORKS: A NEW APPROACH FOR NONLINEAR TIME SERIES MODELLING

MAYTE SUAREZ FARIÑAS, CARLOS E. PEDREIRA, AND MARCELO C. MEDEIROS

ABSTRACT. In this paper, the Local Global Neural Networks model is proposed within the context of time series models. This formulation encompasses some already existing nonlinear models and also admits the Mixture of Experts approach. We place emphasis on the linear expert case and extensively discuss the theoretical aspects of the model: stationarity conditions, existence, consistency and asymptotic normality of the parameter estimates, and model identifiability. The proposed model consists of a mixture of stationary or non-stationary linear models and is able to describe "intermittent" dynamics: the system spends a large fraction of the time in a bounded region, but, sporadically, it develops an instability that grows exponentially for some time and then suddenly collapses. Intermittency is a commonly observed behavior in ecology and epidemiology, fluid dynamics and other natural systems. A model building strategy is also considered and the parameters are estimated by concentrated maximum likelihood. The whole procedure is illustrated with two real time-series.

KEYWORDS. Neural networks, nonlinear models, time-series, model identifiability, parameter estimation, model building, sunspot number.

Forthcoming in the *Journal of the American Statistical Association: Theory and Methods*

1. INTRODUCTION

The past few years have witnessed a vast development of nonlinear time series techniques (Tong, 1990; Granger and Teräsvirta, 1993). Among them, nonparametric models that do not make assumptions about the parametric form of the functional relationship between the variables to be modelled have become widely applicable due to computational advances. For some references on nonparametric time series models see Härdle (1990), Härdle, Lütkepohl, and Chen (1997), Heiler (1999), and Fan and Yao (2003). Another class of models, the flexible functional forms, offers an alternative that in fact also leaves the functional form of the relationship partially unspecified. While these models do contain parameters, often a large number of them, the parameters are not globally identified. Identification, if achieved, is local at best without imposing restrictions on the parameters. Usually, the parameters are not interpretable either as they often are in parametric models.

Date: February, 09 2004.

The artificial neural network (ANN) model is a prominent example of such a flexible functional form. It has found applications in a number of fields, including economics, finance, energy, epidemiology, etc. The use of the ANN model in applied work is generally motivated by the mathematical result stating that under mild regularity conditions, a relatively simple ANN model is capable of approximating any Borel-measurable function to any given degree of accuracy (Funahashi, 1989; Cybenko, 1989; Hornik, Stinchombe, and White, 1989, 1990; White, 1990; Gallant and White, 1992).

Another example of a flexible model, derived from ANNs, is the mixture-of-experts. The idea is to "divide and conquer" and was proposed by Jacobs, Jordan, Nowlan, and Hinton (1991). The motivation for the development of this model is twofold: first, the ideas of Nowlan (1990), viewing competitive adaptation in unsupervised learning as an attempt to fit a mixture of simple probability distributions into a set of data points; and the ideas developed in Jacobs (1990) using a similar modular architecture but a different cost function. Jordan and Jacobs (1994) generalized the above ideas by proposing the so-called hierarchical mixture-of-experts. Both the mixture-of-experts and the hierarchical mixture-of-experts have been applied with success in different areas. In terms of mixtures-of-experts of time series models, the literature focuses mainly on mixtures of Gaussian processes. For example, Weigend, Mangeas, and Srivastava (1995) show an application to financial time series forecasting. Good applications of hierarchical mixtures-of-experts in time series are given by Huerta, Jiang, and Tanner (2001) and Huerta, Jiang, and Tanner (2003). Carvalho and Tanner (2002a) and Carvalho and Tanner (2002b) proposed the mixture of generalized linear time series models and derived several asymptotic results. It would worth mentioning the Mixture Autoregressive model proposed by Wong and Li (2000) and its generalization developed in Wong and Li (2001).

This paper proposes a new model, based on ANNs and partly inspired by the ideas form the mixture-of-experts literature, named Local Global Neural Networks (LGNN). The main idea is to locally approximate the original function by a set of very simple approximation functions. The input-output mapping is expressed by a piecewise structure. The network output is constituted by a combination of several pairs, each of those, composed by an approximation function and by an activation-level function. The activation-level function defines the role of an associated approximation function, for each subset of the domain. Partial superposition of activation level functions is allowed. In this way, modelling is approached by the specialization of neurons in each of the sectors of the domain. In other words, the neurons are formed by pairs of activation level and approximation functions, which emulate the generator function in different sub-sets of the domain. The level of specialization in a given sector is proportional to the value of the activation-level function. This formulation encompasses some already existing nonlinear models and can be interpreted as a mixture of experts model. We place emphasis on the linear expert case. The model is then called the

Linear Local Global Neural Network (L^2GNN) model. A geometric interpretation of the model is given and the conditions under which the proposed model is asymptotically stationary are carefully studied. We show that the L^2GNN model consists of a mixture of stationary or non-stationary linear models, being able to describe "intermittent" dynamics: the system spends a large fraction of the time in a bounded region, but, sporadically, it develops an instability that grows exponentially for some time and then suddenly collapses. Furthermore, based on Trapletti, Leisch, and Hornik (2000), we extensively discuss the existence, consistency, and asymptotic normality of the parameter estimates. Conditions under which the L^2GNN model is identifiability are also carefully considered. Identification is essential for consistency and asymptotic normality of the parameter estimates. A model building strategy is developed and the parameters are estimated by concentrated maximum likelihood, which reduces dramatically the computational burden. The whole procedure is illustrated with two real time-series. Similar proposals are the Stochastic Neural Network (SNN) model developed in Lai and Wong (2001) and the Neuro-Coefficient Smooth Transition Autoregressive (NCSTAR) model of Medeiros and Veiga (2000a).

The paper proceeds as follows. Section 2 presents the model and Section 3 discuss the geometric interpretation for it. Section 4 presents some probabilistic properties of the L^2GNN model. Parameter estimation is considered in Section 5. A model building strategy is discussed in Section 6. Section 7 shows examples with real time-series and finally, Section 8 briefly summarizes our results. A technical appendix provides the proofs of the main results.

2. MODEL FORMULATION

The Local Global Neural Network (LGNN) model describes a stochastic process $y_t \in \mathbb{R}$ through the following nonlinear model:

$$y_t = G(\mathbf{x}_t; \boldsymbol{\psi}) + \varepsilon_t, \quad t = 1, \ldots, T, \tag{1}$$

where $\mathbf{x}_t \in \mathbb{R}^q$ represents a vector of lagged values of y_t and/or some exogenous variables, $\{\varepsilon_t\}$ is sequence of independently and identically distributed random variables with zero mean and variance $\sigma^2 < \infty$. The function $G(\mathbf{x}_t; \boldsymbol{\psi})$ is a nonlinear function of $\mathbf{x}_t$, with the vector of parameters $\boldsymbol{\psi}$ belonging to a compact subspace Ψ of the Euclidean space, and is defined as:

$$G(\mathbf{x}_t; \boldsymbol{\psi}) = \sum_{i=1}^{m} L\left(\mathbf{x}_t; \boldsymbol{\psi}_{L_i}\right) B\left(\mathbf{x}_t; \boldsymbol{\psi}_{B_i}\right), \tag{2}$$

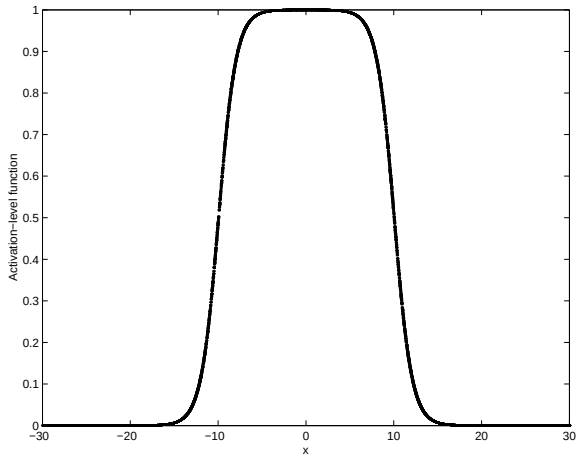


FIGURE 1. Example of an activation-level function with $\mathbf{x}_t \sim \text{Unif}(-30, 30)$, $\gamma = 1$, $\mathbf{d} = 1$, $\beta^{(1)} = -10$, and $\beta^{(2)} = 10$.

where $\boldsymbol{\psi} = \left[\boldsymbol{\psi}_L', \boldsymbol{\psi}_B'\right]'$, $\boldsymbol{\psi}_L = \left[\boldsymbol{\psi}_{L_1}', \ldots, \boldsymbol{\psi}_{L_m}'\right]'$, $\boldsymbol{\psi}_B = \left[\boldsymbol{\psi}_{B_1}', \ldots, \boldsymbol{\psi}_{B_m}'\right]'$, and the functions $B\left(\mathbf{x}_t; \boldsymbol{\psi}_{B_i}\right)$: $\mathbb{R}^q \to \mathbb{R}$ and $L\left(\mathbf{x}_t; \boldsymbol{\psi}_{L_i}\right) : \mathbb{R}^q \to \mathbb{R}$ are named, respectively, as activation-level and approximation functions. Furthermore, $B\left(\mathbf{x}_t; \boldsymbol{\psi}_{B_i}\right)$ is defined as:

$$B\left(\mathbf{x}_t; \boldsymbol{\psi}_{B_i}\right) = -\left[\frac{1}{1 + \exp\left(\gamma_i\left(\langle \mathbf{d}_i, \mathbf{x}_t\rangle - \beta_i^{(1)}\right)\right)} - \frac{1}{1 + \exp\left(\gamma_i\left(\langle \mathbf{d}_i, \mathbf{x}_t\rangle - \beta_i^{(2)}\right)\right)}\right], \tag{3}$$

where

$$\boldsymbol{\psi}_{B_i} = \left[\gamma_i, d_{i1}, \ldots, d_{iq}, \beta_i^{(1)}, \beta_i^{(2)}\right]',$$

$\langle \cdot, \cdot \rangle$ denotes the internal product in Euclidean space, $\gamma_i \in \mathbb{R}$, $\mathbf{d}_i \in \mathbb{R}^q$, $\beta_i^{(1)} \in \mathbb{R}$, and $\beta_i^{(2)} \in \mathbb{R}$, $i = 1, \ldots, m$. It is clear that due to the existence of γ_i in the expression (3), the restriction $\|\mathbf{d}_i\| = 1$ can be made, without loss of model generality. Figure 1 shows an example of an activation-level function.

In the present paper, the approximation functions are linear, that is: $L\left(\mathbf{x}_t; \boldsymbol{\psi}_{L_i}\right) = \mathbf{a}_i'\mathbf{x}_t + b_i$, with $\mathbf{a}_i = [a_{11}, a_{12}, \ldots, a_{1q}]' \in \mathbb{R}^q$ and $b_i \in \mathbb{R}$. In that case the model is called the Linear Local-Global Neural Network (L^{2}GNN) model, where

$$y_t = \sum_{i=1}^{m} \left(\mathbf{a}_i'\mathbf{x}_t + b_i\right) B\left(\mathbf{x}_t; \boldsymbol{\psi}_{B_i}\right) + \varepsilon_t, \quad t = 1, \ldots, T, \tag{4}$$

$\boldsymbol{\psi}_{L_i} = [a_{i1}, \ldots, a_{iq}, b_i]'$, $\boldsymbol{\psi} \in \mathbb{R}^{2m(2+q)}$, and the stochastic process y_t consists of a mixture of linear processes. In (4), we consider that ε_t is a random noise normally distributed. The normality assumption can be relaxed and substituted by some moment conditions.

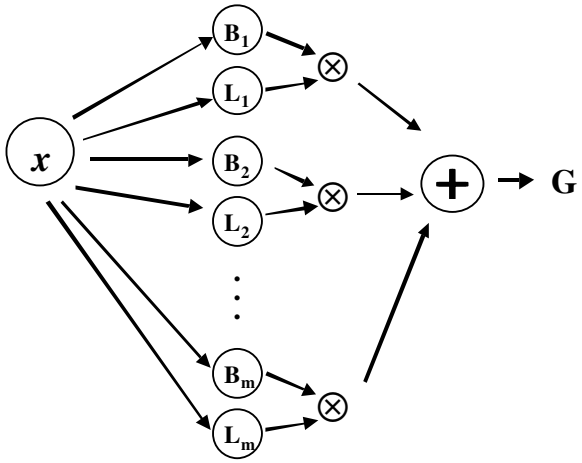


FIGURE 2. Neural network architecture.

This architecture, initially proposed by Pedreira, Pedroza, and Fariñas (2001) for the problem of approximations of L^2-integrable real functions in the univariate case, can be represented through the diagram illustrated in Figure 2. Notice that the hidden layer is formed by m pairs of neurons. Each pair of neurons is composed of the activation-level unit, represented by function $B(\mathbf{x}_t; \boldsymbol{\psi}_{B_i})$, and the approximation unit related to function $L(\mathbf{x}_t; \boldsymbol{\psi}_{L_i})$, $i = 1, \ldots, m$. We should however stress the fact that model (4) is, in principle, neither globally nor locally identified. This issue will be fully addressed in Section 5.2.

As pointed out in the introduction, the L^{2}GNN model is closely related to the NCSTAR model of Medeiros and Veiga (2000a) and the SNN model of Lai and Wong (2001). But though closely related, there are significant differences. The NCSTAR model can be written as

$$y_t = \mathbf{a}_0'\mathbf{x}_t + b_0 + \sum_{i=1}^{m} (\mathbf{a}_i'\mathbf{x}_t + b_i) F(\mathbf{x}_t; \mathbf{d}_i, \beta_i) + \varepsilon_t \tag{5}$$

where $F(\mathbf{x}_t; \mathbf{d}, \beta_i)$ is a *single* logistic function, unlike our equation (3) which is the difference between two logistic functions, defined as

$$F(\mathbf{x}_t; \mathbf{d}_i, \beta_i) = \frac{1}{1 + e^{-(\mathbf{d}_i'\mathbf{x}_t + \beta_i)}},$$

and ε_t is a Gaussian white noise. The SNN model starts from this same equation (see equation (8) in Lai and Wong (2001)), and then replaces the logistic functions $F(\cdot)$ by stochastic Bernoulli variables I_{ti}, $i = 1, \ldots, m$, whose expectation value equals $F(\mathbf{x}_t; \mathbf{d}_i, \beta_i)$ (equations (9a) and (9b) in op. cit.). There are two main implications of these differences. First, on the contrary of the NCSTAR and L^{2}GNN models, the SNN model is a stochastic linear map; since given the choice of I_{ti} the map is linear, the nonlinearities do not appear in the maps themselves, but in the probabilities of choosing which particular map is applied

at a specific timestep. This allows Lai and Wong to use the notion of soft splits proposed by Jordan and Jacobs (1994), mapping the model to a hierarchical mixture of experts and to use a fast EM (Expectation-Maximization) estimation algorithm. But though the introduction of the random variables I_{ti} looks minor, in fact it changes the asymptotics of the model in important ways. First, it should be noted that the one-step-ahead predictor is the same in the SNN model and in (5), because the expected value of the variables I_{ti} is $F(\mathbf{x}_t; \mathbf{d}_i, \beta_i)$; however, the residuals, and with them the *variance* of the predictor, are different, since, for a given timeset, the variables I_{ti}, $i = 1, \ldots, m$, can assume 2^m distinct values and so introduce a new source of variability beyond the ε_t. Therefore, the $n-$step dynamics of the L^{2}GNN, NCSTAR, and SNN models are quite different, and the estimators differ accordingly. The second difference sets apart the L^{2}GNN model from *both* NCSTAR and SNN models, and is to our mind more fundamental. Given a random choice of the model parameters, if an eigenvalue of the characteristic equation of some of the limiting linear model falls outside the unit circle, the NCSTAR and SNN models will be asymptotically non-stationary with probability strictly greater than 0; particular (i.e., measure zero) choices of parameters have to be made to guarantee asymptotic stationarity in this case. On the contrary, the L^{2}GNN model will remain asymptotically stationary with probability one by imposing some very weak restrictions on the parameter $\mathbf{d}$ (see Theorem 1); particular choices of parameters have to be made to permit the dynamics to diverge. It is thus interesting to notice that although the NCSTAR and SNN models are in some sense "supersets" of the L^{2}GNN model, since each L^{2}GNN map can be written as two maps in (5), an important property which is generic for the L^{2}GNN case (asymptotic stationarity) is not generic for the "more general" models. Furthermore, the stationarity condition presented in Section 3 of Lai and Wong (2001) eliminates the possibility of mixing non-stationary linear models. Asymptotic stationarity of L^{2}GNN model is discussed in Section 4. The core of the idea is that the activation functions of the NCSTAR and SNN models are "large", being "active" in half the space, while the activation functions of the L^{2}GNN model are "small", since they cover a small fraction of any sufficiently large sphere. Thus if the NCSTAR or SNN models are non-stationary, the dynamics can easily escape to infinity; if an L^{2}GNN model is non-stationary, the trajectory has to escape along a direction exactly perpendicular to $\mathbf{d}$, and any deviation will cause the trajectory to "fall off" the activation function and return close to the origin. Both NCSTAR and SNN models could do exactly this by using extra maps; however, the parameters of these extra maps have to be chosen exactly, and a small random perturbation of the model parameters would, with probability one, destroy the property. An important type of dynamical behavior is called "intermittent" dynamics: the system spends a large fraction of the time in a bounded region, but, sporadically, it develops an instability that grows exponentially for some time and then suddenly collapses. Intermittency is a commonly observed behavior in ecology and epidemiology (breakouts), fluid

dynamics (turbulent plumes) and other natural systems. The L^{2}GNN model can fit such dynamics *robustly*, meaning small perturbations of the parameters do not change the behavior; NCSTAR and SNN models can by definition fit that dynamics too, but the fit is sensitive to small perturbations.

3. GEOMETRIC INTERPRETATION

In this section we give a geometric interpretation of a layer of hidden neuron-pairs. Let be $\mathbf{x}_t \in \mathbb{X}$, where $\mathbb{X}$ is a vector space with internal product denoted by $\langle \cdot, \cdot \rangle$. The parameters $\mathbf{d}$, $\beta^{(1)}$ and $\beta^{(2)}$ in (4) define two parallel hyperplanes in $\mathbb{X}$:

$$\mathbb{H}_1 = \left\{ \mathbf{x}_t \in \mathbb{R}^q | \, \langle \mathbf{d}, \mathbf{x}_t \rangle = \beta^{(1)} \right\} \quad \text{and} \quad \mathbb{H}_2 = \left\{ \mathbf{x}_t \in \mathbb{R}^q | \, \langle \mathbf{d}, \mathbf{x}_t \rangle = \beta^{(2)} \right\}. \tag{6}$$

The position of each hyperplane is determined by direction vector $\mathbf{d}$. The scalars $\beta^{(1)}$ and $\beta^{(2)}$ determine the distance of the hyperplanes to the origin of coordinates. As a hyperplane has infinite direction vectors, the restriction $\|\mathbf{d}\| = 1$ reduces this multiplicity, without loss of generality. Thus, the hyperplanes $\mathbb{H}_1$ and $\mathbb{H}_2$ are parallel due to the fact that they have the same direction vector, and divide $\mathbb{X}$ into three different regions: $\mathbb{H}^-$, $\mathbb{H}^0$, $\mathbb{H}^+$ defined as:

$$\begin{aligned} \mathbb{H}^- &= \left\{ \mathbf{x}_t \in \mathbb{R}^q | \, \langle \mathbf{d}, \mathbf{x}_t \rangle < \beta^{(1)} \right\} \\ \mathbb{H}^0 &= \left\{ \mathbf{x}_t \in \mathbb{R}^q | \, \langle \mathbf{d}, \mathbf{x}_t \rangle \geq \beta^{(1)} \text{ and } \langle \mathbf{d}, \mathbf{x}_t \rangle \leq \beta^{(2)} \right\} \\ \mathbb{H}^+ &= \left\{ \mathbf{x}_t \in \mathbb{R}^q | \, \langle \mathbf{d}, \mathbf{x}_t \rangle > \beta^{(2)} \right\} \end{aligned} \tag{7}$$

The region $\mathbb{H}^0$ represents the active state of the neuron pair and regions $\mathbb{H}^-$ and $\mathbb{H}^+$ represent the inactive state. The active or non-active state of the neuron pair is represented by activation-level function $B\left(\mathbf{x}_t; \boldsymbol{\psi}_B\right)$. Parameter γ determines the slope of the activation-level function, characterizing the smoothness of transition from one state to another. Thus, the extreme case $\gamma \to \infty$ represents an abrupt transition between states.

When m neuron-pairs are considered, there are m pairs of hyperplanes. Therefore, m closed $\mathbb{H}^0$-type regions will exist that could intercept one another or not. Thus, $\mathbb{X}$ will be divided into polyhedral regions. If not all hyperplanes are parallel, that is, if $\exists i, j, i \neq j$, such that $\mathbf{d}_i \neq \mathbf{d}_j$ the region formed by the interception of hyperplanes, $\mathbb{H}^0_{ij} = \mathbb{H}^0_i \bigcap \mathbb{H}^0_j$, is non-empty region and represents the region where the neuron-pairs i and j are both active.

One case that worth special mention is when the hyperplanes are parallel to each other, that is $\mathbf{d}_i = \mathbf{d}$, $\forall i$. In that case we would have m parallel regions of the $\mathbb{H}^0$-type. Under condition $\beta^{(2)}_i < \beta^{(1)}_{i+1}$, $\forall i$, the

intersection of these regions is empty. The L^{2}GNN model can thus be interpreted as a piecewise linear model with a smooth transition between regimes. For a review on smooth transition time-series models see van Dijk, Teräsvirta, and Franses (2002).

4. PROBABILISTIC PROPERTIES

Deriving necessary and sufficient conditions for stationarity of nonlinear time-series models is usually not easy and that is also the case of the L^{2}GNN model. One possibility, as the L^{2}GNN model can be interpreted as a functional coefficient autoregressive (FAR) model if $\mathbf{x}_t = [y_{t-1}, \ldots, y_{t-p}]'$, is to apply the results derived in Chen and Tsay (1993) and applied in Lai and Wong (2001). However, the resulting restrictions are extremely restrictive. For example, as ε_t is normally distributed, y_t is geometrically ergodic if all roots of the characteristic equation $\lambda^p - c_1\lambda^{p-1} - \cdots - c_p = 0$ are inside the unit circle, where $c_j = \sum_{i=1}^{m} |a_{ij}|$, $j = 1, \ldots, p$. Fortunately, following a similar rationale as in the case of linear autoregressive (AR) processes, Theorem 1 gives less restrictive sufficient conditions for the asymptotically stationarity of the L^{2}GNN model. It is easy to check that model (4) has at most N limiting linear models of the form $y_t = c_0^{(k)} + c_1^{(k)} y_{t-1} + \cdots + c_p^{(k)} y_{t-p} + \varepsilon_t$, where $N = \sum_{i=1}^{m} \binom{m}{i}$.

THEOREM 1. *The L^{2}GNN model is asymptotically stationary if one of the following restrictions is satisfied:*

(1) *The roots of $\lambda^p - c_1^{(k)}\lambda^{p-1} - \cdots - c_p^{(k)} = 0$, $k = 1, \ldots, N$, are inside the unit circle.*

(2) *There is a $k \in \{1, 2, \ldots, N\}$ such that at least one root of $\lambda^p - c_1^{(k)}\lambda^{p-1} - \cdots - c_p^{(k)} = 0$ is outside the unit circle and $d_{ij} \neq 0$, $i = 1, \ldots, m$, $j = 1, \ldots, p$.*

(3) *There is a $k \in \{1, 2, \ldots, N\}$ such that at least one root of $\lambda^p - c_1^{(k)}\lambda^{p-1} - \cdots - c_p^{(k)} = 0$ is equal to one, the others are inside the unit circle, and $\mathbf{d}_i$, $i = 1, \ldots, m$ is not orthogonal to the eigenvectors of the transition matrix*

$$\mathbf{A}^{(k)} = \begin{bmatrix} c_1^{(k)} & c_2^{(k)} & c_3^{(k)} & \cdots & c_{p-1}^{(k)} & c_p^{(k)} \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}. \tag{8}$$

The proof of the theorem is given in the Appendix and is based on the results for linear autoregressive models. The intuition behind the above result is that when y_t grows in absolute value, the functions $B\left(\mathbf{x}_t; \boldsymbol{\psi}_{B_i}\right) \to 0$, $i = 1, \ldots, m$, and thus y_t is driven back to zero. Condition 1 is trivial and implies that

all the limiting AR models are asymptotically stationary. Condition 2 considers the case where there are explosive regimes. Finally, Condition 3 is related to the unit-root case.

REMARK 1. *When $p = 1$, the L^{2}GNN model is asymptotically stationary independent of the conditions on the autoregressive parameters.*

The following examples show the behavior of some simulated L^{2}GNN models. Examples 1 and 2 show two stationary L^{2}GNN models that are combinations of explosive linear autoregressive models. To illustrate the dependency on the elements of vector $\mathbf{d}_i$, $i = 1, \ldots, m$, Example 3 shows a model where $\mathbf{d}_2 = [1, 0]'$. Example 4 considers the case with unit-roots.

EXAMPLE 1. *1000 observations of the following L^{2}GNN model:*

$$\begin{aligned} y_t = &(-0.5 - 1.5y_{t-1}) \times \left[\frac{1}{1 + \exp\left(10\left(y_{t-1} + 6\right)\right)} - \frac{1}{1 + \exp\left(10\left(y_{t-1} - 1\right)\right)} \right] + \\ &(-0.5 - 1.2y_{t-1}) \times \left[\frac{1}{1 + \exp\left(10\left(y_{t-1} + 2\right)\right)} - \frac{1}{1 + \exp\left(10\left(y_{t-1} - 2\right)\right)} \right] + \varepsilon_t, \end{aligned} \tag{9}$$

where $\varepsilon_t \sim$ NID$(0, 1)$. Figure 3 shows the generated time-series, the activation-level functions, the auto-correlogram of series, and the histogram of the data. Model (9) is a mixture of two explosive autoregressive processes. Either when only one of the activation-level functions are active or when both of them equal one, the autoregressive model driving the series is explosive. However, as can be observed, the series is stationary. The distribution of the data is highly asymmetrical and there is also some evidence of bimodality. When iterating the skeleton of model (9) and making $t \to \infty$ the process has, in the limit, three stable points: 0.0052, 1.0140, and 2.6567.

EXAMPLE 2. *3000 observations of the following L^{2}GNN model:*

$$\begin{aligned} y_t = &(-0.5 - 2.2y_{t-1} + 2.5y_{t-2}) \times \left[\frac{1}{1 + \exp\left(0.7y_{t-1} - 0.7y_{t-2} + 10\right)} - \right. \\ &\left. \frac{1}{1 + \exp\left(0.7y_{t-1} - 0.7y_{t-2} - 10\right)} \right] + \\ &(0.5 - 1.9y_{t-1} - 1.2y_{t-2}) \times \left[\frac{1}{1 + \exp\left(1.5\left(0.7y_{t-1} - 0.7y_{t-2} + 2\right)\right)} - \right. \\ &\left. \frac{1}{1 + \exp\left(1.5\left(0.7y_{t-1} - 0.7y_{t-2} - 40\right)\right)} \right] + \varepsilon_t, \end{aligned} \tag{10}$$

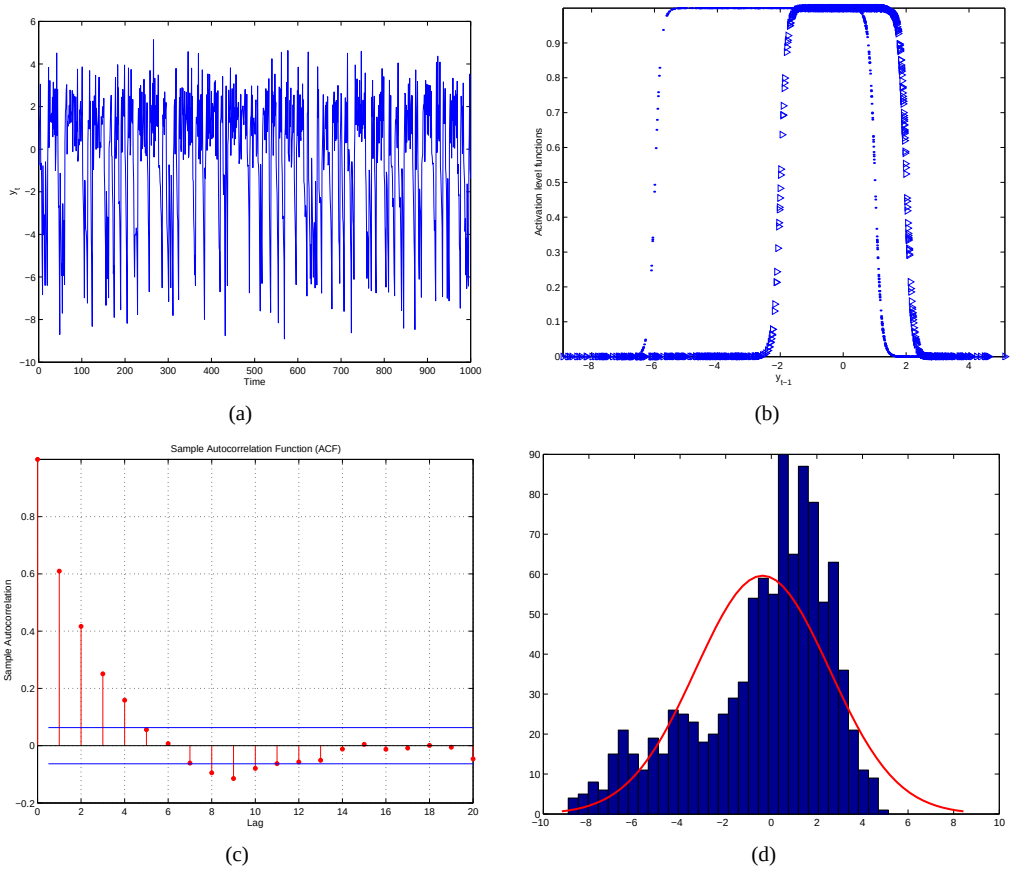


FIGURE 3. Example 1. Panel (a): generated time-series. Panel (b): Scatter plot of the activation-level functions against y_{t-1}. Panel (c): Autocorrelogram of the series. Panel (d): Histogram of the series.

where $\varepsilon_t \sim NID(0,1)$. *Figure 4 shows the generated time-series, the activation-level functions, the autocorrelogram of series, and the histogram of the data. As can be observed , even with explosive regimes, the series is stationary. However, it is strongly not normal and bimodal.*

EXAMPLE 3. *30000 observations of the following* L^2*GNN model:*

$$\begin{aligned} y_t = &(-0.5 - 2.2y_{t-1} + 2.5y_{t-2}) \times \left[\frac{1}{1+\exp\left(0.7y_{t-1} - 0.7y_{t-2} + 10\right)} - \right. \\ &\left. \frac{1}{1+\exp\left(0.7y_{t-1} - 0.7y_{t-2} - 10\right)} \right] + \\ &(0.5 - 1.9y_{t-1} - 1.2y_{t-2}) \times \left[\frac{1}{1+\exp\left(1.5\left(y_{t-1} + \delta y_{t-2} + 2\right)\right)} - \right. \\ &\left. \frac{1}{1+\exp\left(1.5\left(y_{t-1} + \delta y_{t-2} - 40\right)\right)} \right] + \varepsilon_t, \end{aligned} \tag{11}$$

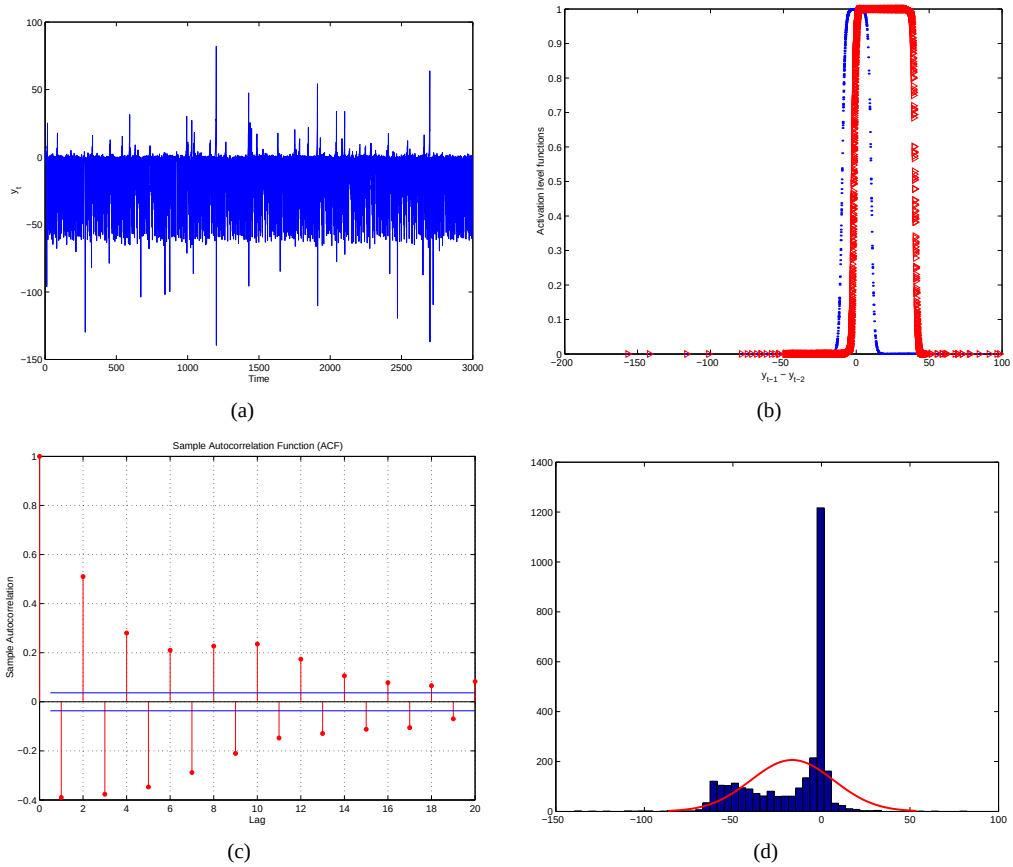
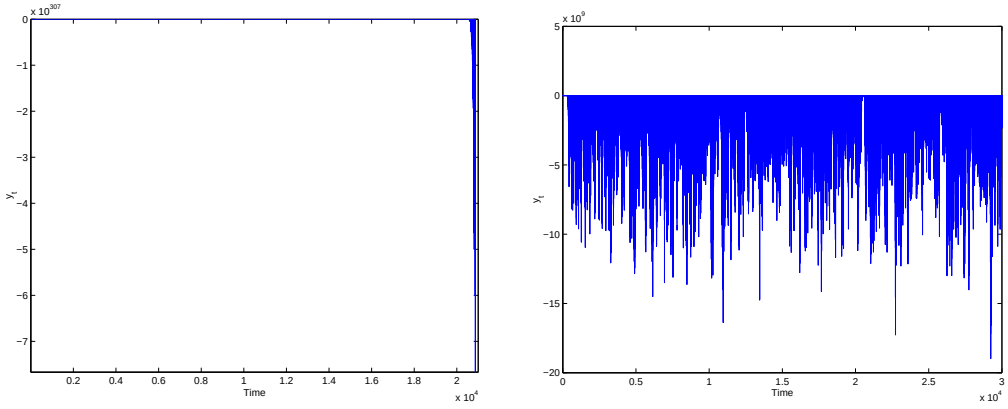


FIGURE 4. Example 2. Panel (a): generated time-series. Panel (b): Scatter plot of the activation-level functions against $y_{t-1} - y_{t-2}$. Panel (c): Autocorrelogram of the series. Panel (d): Histogram of the series.

where $\varepsilon_t \sim NID(0,1)$ *and* $\delta = 0, 10^{-10}$. *Figure 5 shows the generated time-series. As can be observed, the process is explosive when* $\delta = 0$ *but is asymptotically stationary when* $\delta = 10^{-10}$.

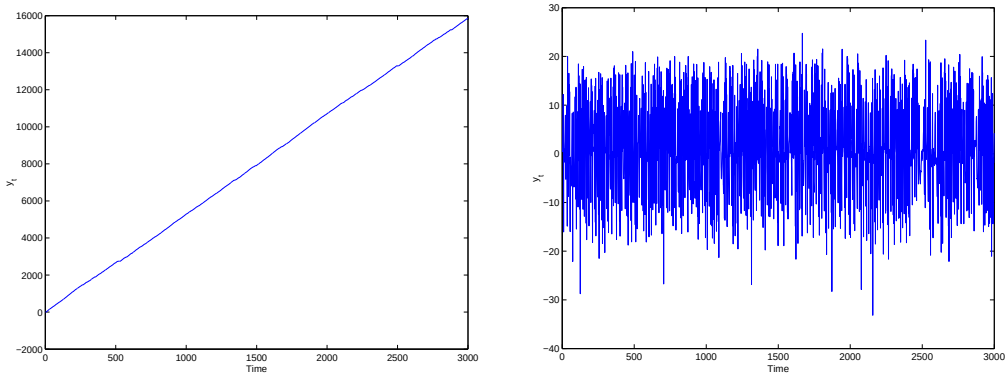
EXAMPLE 4. *30000 observations of the following* L^2*GNN model:*

$$\begin{aligned} y_t = (0.5 + 2y_{t-1} - y_{t-2}) \times \Bigg[& \frac{1}{1+\exp\left(0.7y_{t-1} - 0.7\delta y_{t-2} + 10\right)} - \\ & \frac{1}{1+\exp\left(0.7y_{t-1} - 0.7\delta y_{t-2} - 10\right)} \Bigg] + \\ (0.5 - 0.5y_{t-1} + 0.5y_{t-2}) \times \Bigg[& \frac{1}{1+\exp\left(0.7y_{t-1} - 0.7\delta y_{t-2} - 5\right)} - \\ & \frac{1}{1+\exp\left(0.7y_{t-1} - 0.7\delta y_{t-2} - 15\right)} \Bigg] + \varepsilon_t, \end{aligned} \tag{12}$$



(a) (b)

FIGURE 5. Example 3. Generated time-series. Panel (a): $\delta = 0$. Panel (b): $\delta = 10^{-10}$.



(a) (b)

FIGURE 6. Example 4. Generated time-series. Panel (a): $\delta = 1$. Panel (b): $\delta = -1$.

where $\varepsilon_t \sim$ *NID*$(0, 1)$ *and* $\delta = -1, 1$. *It can be seen that model (12) has three limiting AR regimes. The associated transition matrixes – see Equation (8) – are:*

$$\mathbf{A}^{(1)} = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix}, \ \mathbf{A}^{(2)} = \begin{bmatrix} 1.5 & -0.5 \\ 1 & 0 \end{bmatrix}, \ \mathbf{A}^{(3)} = \begin{bmatrix} -0.5 & 0.5 \\ 1 & 0 \end{bmatrix},$$

with the respective eigenvalues pairs: $(1, 1)$, $(1, 0.5)$, *and* $(-1, 0.5)$. *Figure 6 shows the generated time-series. As can be observed, the process is not stationary when* $\delta = 1$ *but is asymptotically stationary when* $\delta = -1$.

5. PARAMETER ESTIMATION

A large number of algorithms for estimating the parameters of models based on neural networks are available in the literature. In this paper we estimate the parameters of our L^{2}GNN model by maximum

likelihood making use of the assumptions made of ε_t in Section 2. The use of maximum likelihood or quasi maximum likelihood makes it possible to obtain an idea of the uncertainty in the parameter estimates through (asymptotic) standard deviation estimates. However, 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 favour of maximum likelihood here. First, we suggest a model building strategy that proceeds from small to large models, so that estimation of unidentified or nearly unidentified models, a major reason for the need to penalize the log-likelihood, is partially avoided. Second, the starting-values of the parameter estimates are chosen carefully, and we discuss the details of this later in this section.

The L^2GNN model is similar to many linear or nonlinear time series models in that the information matrix of the logarithmic likelihood function is block diagonal in such a way that we can concentrate the likelihood and first estimate the parameters of the conditional mean. Thus conditional maximum likelihood is equivalent to nonlinear least squares. Hence the parameter vector $\boldsymbol{\psi}$ of the L^2GNN model defined by (4) is estimated as

$$\widehat{\boldsymbol{\psi}} = \underset{\boldsymbol{\psi}}{\operatorname{argmin}} Q_T(\boldsymbol{\psi}) = \frac{1}{T}\sum_{t=1}^{T}\left[y_t - G\left(\mathbf{x}_t;\boldsymbol{\psi}\right)\right]^2. \tag{13}$$

The least squares estimator (LSE) defined by (13) belongs to the class of M-estimators considered by Pötscher and Prucha (1986). We next discuss the conditions that guarantee the existence, consistency, and asymptotic normality of the LSE. We also state sufficient conditions under which the L^2GNN model is identifiable.

5.1. **Existence of the Estimator.** The proof of existence is based on Lemma 2 of Jennrich (1969), which establishes that under certain conditions of continuity and measurability on the mean square error (MSE) function, the least squares estimator exists. Theorem 2 state the necessary conditions for the existence of the LSE.

THEOREM 2. *The L^2NGG model satisfies the following conditions and the LSE exists.*

(1) *For each $\mathbf{x}_t \in \mathbb{X}$, function $G_{\mathbf{x}}\left(\boldsymbol{\psi}\right) = G\left(\mathbf{x}_t;\boldsymbol{\psi}\right)$ is continuous in compact subset Ψ of the Euclidean space.*

(2) *For each $\boldsymbol{\psi} \in \Psi$, function $G_{\boldsymbol{\psi}}\left(\mathbb{X}\right) = G\left(\mathbf{x}_t;\boldsymbol{\psi}\right)$ is measurable in space $\mathbb{X}$.*

(3) *ε_t are errors independent and identically distributed with mean zero and variance σ^2.*

REMARK 2. *In order to extend the set of approximation functions beyond linear functions, we need to check that conditions (1) and (2) of Theorem 2. Thus, the class of functions $L\left(\mathbf{x}_t;\boldsymbol{\psi}_{L_i}\right)$, $i = 1,\ldots,m$, to*

be considered must be a subset of the continuous functions on compact set Ψ that are also measurable in $\mathbb{X}$.

REMARK 3. *The hypothesis of compactness of the parameter space may seem a little too restrictive. It is presented in Huber (1967) results that only require locally compact spaces, and an extension of this can be applied to obtain similar results in the present case. However, the compactness assumption is convenient for theoretical reasons and is still general enough to be applied whenever the optimization procedure is carried out by a computer.*

5.2. **Identifiability of the Model.** A fundamental problem for statistical inference with nonlinear time series models is the unidentifiability of the model parameters. To guarantee unique identifiability of the mean square error (MSE) function, the sources of uniqueness of the model must be studied. These questions are studied in Sussman (1992), Kurková and Kainen (1994), Hwang and Ding (1997), Trapletti, Leisch, and Hornik (2000), and Medeiros, Teräsvirta, and Rech (2002) in the case of a feedforward neural network model. Here, the main concepts and results will be briefly discussed. In particular, the conditions that guarantee that the proposed model is identifiable and minimal will be established and proven. Before tackling the problem of the identifiability of the model, two related concepts will be discussed: the concept of minimality of the model, established in Sussman (1992) and which Hwang and Ding (1997) called "non-redundancy"; and the concept of reducibility of the model.

DEFINITION 1. *The L^2GNN model is minimal (or non-redundant), if its input-output map cannot be obtained from another model with fewer neuron-pairs.*

One source of unidentifiability comes from the fact that a model may contain irrelevant neuron-pairs. This means that there are cases where the model can then be reduced, eliminating some neuron-pairs without changing the input-output map. Thus, the minimality condition can only hold for irreducible models.

DEFINITION 2. *Define $\boldsymbol{\theta}_{i\ell} = \left[\gamma_i, \mathbf{d}_i', \beta_i^{(\ell)}\right]'$ and let $\varphi\left(\mathbf{x}_t; \boldsymbol{\theta}_{i\ell}\right) = \gamma_i \left(\langle \mathbf{d}_i, \mathbf{x}_t \rangle - \beta_i^{(\ell)}\right)$, $i = 1, \ldots, m$, and $\ell = 1, 2$. The L^2GNN model defined in (4) is reducible if one of the following three conditions holds:*

(1) *One of the pairs $(\mathbf{a}_i, b_i)$ vanishes jointly for some $i = 1, \ldots, m$.*

(2) *$\gamma_i = 0$ for some $i = 1, \ldots, m$.*

(3) *There is at least one pair (i, j), $i \neq j$, $i = 1, \ldots, m$, $j = 1, \ldots, m$, such that $\varphi\left(\mathbf{x}_t; \boldsymbol{\theta}_{i\ell}\right)$ and $\varphi\left(\mathbf{x}_t; \boldsymbol{\theta}_{j\ell}\right)$ are sign-equivalent. That is, $|\varphi\left(\mathbf{x}_t; \boldsymbol{\theta}_{i\ell}\right)| = |\varphi\left(\mathbf{x}_t; \boldsymbol{\theta}_{j\ell}\right)|$, $\forall \mathbf{x}_t \in \mathbb{R}^q$, $t = 1, \ldots, T$.*

DEFINITION 3. *The L^2GNN model is identifiable if there are no two sets of parameters such that the corresponding distributions of the population variable y are identical.*

Four properties of the L^2GNN model cause unidentifiability of the models:

(P.1) The property of interchangeability of the hidden neuron-pairs. The value of the likelihood function of the model does not change if the neuron-pairs in the hidden layer are permuted. This results in $m!$ different models that are indistinct among themselves (related to the input-output map). As a consequence, in the estimation of parameters, we will have $m!$ equal local maxima for the loglikelihood function.

(P.2) The symmetry of the function $B\left(\mathbf{x}_t; \boldsymbol{\psi}_{B_i}\right)$, $i = 1, \ldots, m$. The fact that activation-level function satisfies that

$$B\left(\mathbf{x}_t; \gamma, \mathbf{d}_i, \beta_i^{(1)}, \beta_i^{(2)}\right) = -B\left(\mathbf{x}_t; \gamma, \mathbf{d}_i, \beta_i^{(2)}, \beta_i^{(1)}\right),$$

establishes another indetermination in the model, as we may have 2^m equivalent parameterizations.

(P.3) The fact that $F(-z) = 1 - F(z)$, where $F(z) = [1 + \exp(-z)]^{-1}$ which implies that the activation level function satisfies that

$$B\left(\mathbf{x}_t; \gamma, \mathbf{d}_i, \beta_i^{(1)}, \beta_i^{(2)}\right) = -B\left(\mathbf{x}_t; -\gamma, \mathbf{d}_i, \beta_i^{(2)}, \beta_i^{(1)}\right),$$

or

$$B\left(\mathbf{x}_t; \gamma, \mathbf{d}_i, \beta_i^{(1)}, \beta_i^{(2)}\right) = -B\left(\mathbf{x}_t; \gamma, -\mathbf{d}_i, -\beta_i^{(2)}, -\beta_i^{(1)}\right).$$

(P.4) The presence of irrelevant hidden neuron-pairs. Conditions (1) – (2) in the definition of reducibility give information about the presence of pairs of irrelevant units, which translate into identifiability sources. If the model contains some pair such that $\mathbf{a}_i = 0$ and $b_i = 0$, parameters $\mathbf{d}_i$, $\beta_i^{(1)}$, and $\beta_i^{(2)}$ remain unidentified. On the other hand, if $\gamma_i = 0$, then parameters $\mathbf{a}_i$ and b_i may take on any value without affecting the value of the loglikelihood function. Furthermore, if $\beta_i^{(1)} = \beta_i^{(2)}$, then γ_i, $\mathbf{a}_i$ and b_i remain unidentified.

Properties (P.2)–(P.3) are related to the concept of reducibility. In the same spirit of the results stated in Sussman (1992) and Hwang and Ding (1997) we show that, if the model is irreducible, property (P.1) is is the only form of modifying the parameters without affecting the distribution of y. Hence, establishing restrictions on the parameters of (4) that simultaneously avoid reducibility and any permutation of hidden units, we guarantee the identifiability of the model.

The problem of interchangeability (property (P.1)) can be prevented with the following restriction

(R.1) $\beta_i^{(1)} < \beta_{i+1}^{(1)}$ and $\beta_i^{(2)} < \beta_{i+1}^{(2)}$, $i = 1, \ldots, m$.

Now the consequences due to the symmetry of the activation-level function (property (P.2)) can be resolved if we consider:

(R.2) $\beta_i^{(1)} < \beta_i^{(2)}$, $i = 1, \ldots, m$.

To remove the lack of identification caused by property (P.3) we have to impose two additional restrictions.

(R.3) $\gamma_i > 0$, $i = 1, \ldots, m$.

(R.4) $d_{i1} > 0$, $i = 1, \ldots, m$.

The first one prevents that a simple change of sign in parameter γ leads to problems in the identification of the model. As previously discussed, we saw that condition $\|\mathbf{d}\| = 1$ restricts this multiplicity in the direction vector of the hyperplane. However, there is still some ambivalence arising from the fact that both $\mathbf{d}$, and $-\mathbf{d}$ have the same norm and are orthogonal to the hyperplane. Restriction (R.4) avoids that problem.

Since $\mathbf{d}_i$ is a unit vector, then:

$$d_{i1} = \sqrt{1 - \sum_{j=2}^{q} d_{ij}^2} > 0.$$

The presence of irrelevant hidden neuron-pairs, property (P.4), can be circumvented by applying a "specific-to-general" model building strategy as suggested in Section 6.

Corollaries 2.1 in Sussman (1992) and 2.4 in Hwang and Ding (1997) guarantee that an irreducible model is minimal. The fact that irreducibility and minimality are equivalent implies that there are no mechanisms, other than the ones listed in the definition of irreducibility, that can be used to reduce the number of units without changing the functional input-output relation. Then, restrictions (R.1)–(R.4) guarantee that if irrelevant units do not exist the model is identifiable and minimal.

Before stating the theorem that gives sufficient conditions under which the L^2GNN model is globally identifiable we should make the following assumption.

ASSUMPTION 1. *The parameters $\mathbf{a}_i$ and b_i do not vanish jointly for some $i = 1, \ldots, m$. Furthermore $\gamma_i > 0$, $\forall\, i$ and $\beta_i^{(1)} \neq \beta_i^{(2)}$, $\forall\, i$.*

ASSUMPTION 2. *The covariate vector $\mathbf{x}_t$ has an invariant distribution which has a density everywhere positive in an open ball.*

Assumption 1 guarantees that there are no irrelevant hidden neuron-pairs as described in property (P.4) above and Assumption 2 avoids problems related to multicollinearity.

THEOREM 3. *Under the restrictions:*

(R.1) $\beta_i^{(1)} < \beta_{i+1}^{(1)}$ *and* $\beta_i^{(2)} < \beta_{i+1}^{(2)}$, $i = 1, \ldots, m$;

(R.2) $\beta_i^{(1)} < \beta_i^{(2)}$, $i = 1, \ldots, m$;

(R.3) $\gamma_i > 0$, $i = 1, \ldots, m$;

(R.4) $d_{i1} = \sqrt{1 - \sum_{j=2}^{q} d_{ij}^2} > 0$, $i = 1, \ldots, m$;

and Assumptions 1 and 2 the L^{2}GNN model is globally identifiable.

5.3. **Strong Consistency of Estimators.** In White (1981) and White and Domowitz (1984) the conditions that guarantee the strong convergence of the LSE are established. In the context of stationary time series models, the conditions that assure the (almost certain) convergence are established in White (1994) and Wooldridge (1994). In what follows we state and prove the theorem of consistency of the estimators of the L^{2}GNN model.

ASSUMPTION 3. *The data generation process (DGP) for the sequence of scalar real valued observations* $\{y_t\}_{t=1}^{T}$ *is a stationary and ergodic L^{2}GNN process with the true parameter vector* $\boldsymbol{\psi}^* \in \Psi$. *The parameter space* Ψ *is a compact subset of* $\mathbb{R}^r$, *where* $r = 2m(2+q)$.

THEOREM 4. *Under Restrictions (R.1)–(R.4) and Assumptions 1 and 3 the least squares estimator is almost surely consistent.*

5.4. **Asymptotic Normality.** The following two conditions are required for the asymptotic normality of the LSE.

ASSUMPTION 4. *The true parameter vector* $\boldsymbol{\psi}^*$ *is interior to* Ψ.

ASSUMPTION 5. *The family of functions*

$$\{x_t\} \bigcup \{B(x_t; \boldsymbol{\psi}_B)\} \bigcup \{\nabla B(x_t; \boldsymbol{\psi}_B))\} \bigcup \{x_t B(x_t; \boldsymbol{\psi}_B)\} \bigcup \{x_t \nabla B(x_t; \boldsymbol{\psi}_B)\},$$

$x_t \in \mathbb{R}$ *and* $\forall t$, *is linearly independent, as long as the functions* $\varphi_i^{(\ell)}(x_t; \boldsymbol{\theta}_{i\ell})$, $i = 1, \ldots, m$, $\ell = 1, 2$, *are not equivalent in sign.*

THEOREM 5. *Under restrictions (R.1)–(R.4) and Assumptions 1–5*

$$\left[\frac{1}{2\sigma^2} \nabla^2 \overline{Q}_T(\boldsymbol{\psi}^*)\right]^{-1/2} \sqrt{T}\left(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*\right) \xrightarrow{d} N(0, \mathbf{I}),$$

where $\nabla^2 \overline{Q}_T(\boldsymbol{\psi}^*) = E\left[\nabla^2 Q_T(\boldsymbol{\psi}^*)\right]$, $\nabla^2 Q_n(\boldsymbol{\psi}^*)$ *is the Hessian matrix of* $Q_T(\boldsymbol{\psi})$ *at* $\boldsymbol{\psi}^*$, *and* σ^2 *is the variance of* ε_t.

5.5. **Concentrated Likelihood.** In order to reduce the computational burden we can apply concentrated maximum likelihood to estimate $\boldsymbol{\psi}$ as follows. Consider the i^{th} iteration of the optimization algorithm and rewrite model (1)–(3) as

$$\mathbf{y} = \mathbf{Z}(\boldsymbol{\psi}_B)\boldsymbol{\psi}_L + \boldsymbol{\varepsilon}, \tag{14}$$

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

$$\mathbf{Z}(\boldsymbol{\psi}_B) = \begin{pmatrix} \mathbf{z}_1' & B(\mathbf{x}_1; \boldsymbol{\psi}_{L_1})\mathbf{z}_1' & \ldots & B(\mathbf{x}_1; \boldsymbol{\psi}_{L_m})\mathbf{z}_1' \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{z}_T' & B(\mathbf{x}_T; \boldsymbol{\psi}_{L_1})\mathbf{z}_T' & \ldots & B(\mathbf{x}_T; \boldsymbol{\psi}_{L_m})\mathbf{z}_T' \end{pmatrix},$$

with $\mathbf{z}_t = [1, \mathbf{x}_t']'$. Assuming $\boldsymbol{\psi}_B$ fixed, the parameter vector $\boldsymbol{\psi}_L$ can be estimated analytically by

$$\widehat{\boldsymbol{\psi}}_L = \left(\mathbf{Z}(\boldsymbol{\psi}_B)'\mathbf{Z}(\boldsymbol{\psi}_B)\right)^{-1}\mathbf{Z}(\boldsymbol{\psi}_B)'\mathbf{y}. \tag{15}$$

The remaining parameters are estimated conditionally on $\boldsymbol{\psi}_L$ by applying the Levenberg-Marquadt algorithm which completes the i^{th} iteration. This form of concentrated maximum likelihood was proposed by Leybourne, Newbold, and Vougas (1998). It reduces the dimensionality of the iterative estimation problem considerably.

5.6. **Starting-values.** Many iterative optimization algorithms are sensitive to the choice of starting-values, and this is certainly so in the estimation of L^2GNN models. Assume now that we have estimated an L^2GNN model model with $m-1$ hidden neuron-pairs and want to estimate one with m neuron-pairs. Our specific-to-general specification strategy has the consequence that this situation frequently occurs in practice. A natural choice of initial values for the estimation of parameters in the model with m neuron-pairs is to use the final estimates for the parameters in the first $m-1$ ones. The starting-values for the parameters in the mth hidden neuron-pair are obtained in steps as follows [1].

(1) For $k = 1, \ldots, K$:
   (a) Construct a vector $\mathbf{v}_m^{(k)} = \left[v_{1m}^{(k)}, \ldots, v_{qm}^{(k)}\right]'$ such that $v_{1m}^{(k)} \in (0, 1]$ and $v_{jm}^{(k)} \in [-1, 1]$, $j = 2, \ldots, q$. The values for $v_{1m}^{(k)}$ are drawn from a uniform $(0, 1]$ distribution and the ones for $v_{jm}^{(k)}$, $j = 2, \ldots, q$, from a uniform $[-1, 1]$ distribution.
   (b) Define $\boldsymbol{d}_m^{(k)} = \mathbf{v}_m^{(k)}\|\mathbf{v}_m^{(k)}\|^{-1}$.
   (c) Compute the projections $\mathbf{p}_m^{(k)} = \left\langle \mathbf{d}_m^{(k)}, \mathbf{x} \right\rangle$, where $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$.

[1] A similar procedure was proposed in Medeiros and Veiga (2000b) and Medeiros, Teräsvirta, and Rech (2002).

(d) Let $c_{1m}^{(k)} = Z_{1/3}\left(\mathbf{p}_m^{(k)}\right)$ and $c_{2m}^{(k)} = Z_{2/3}\left(\mathbf{p}_m^{(k)}\right)$, where Z_α is the α-percentile of the empirical distribution of $\mathbf{p}_m^{(k)}$.

(2) Define a grid of N positive values $\gamma_m^{(n)}$, $n = 1, \ldots, N$, for the slope parameter and estimate $\boldsymbol{\psi}_L$ using equation (15).

(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 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.

6. Model Building

In this section, a specific-to-general specification strategy is developed. From equation (4) two specification problems require special care. The first is variable selection, that is, the correct selection of elements $\mathbf{x}_t$. 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 neuron-pairs. The specification procedure as a whole may be viewed as a sequence consisting of the following steps:

(1) Selecting the elements of $\mathbf{x}_t$.

(2) Determining the number of neuron-pairs.

(3) Evaluation of the estimated model.

The first two steps of the modelling cycle will be discussed in detail. The evaluation is step is beyond the scope of the present paper. However, the results in Medeiros and Veiga (2002), and Medeiros, Teräsvirta, and Rech (2002) can be easily generalized to the case of L^2GNN models.

6.1. Variable Selection.

The first step in our model specification is to choose the variables for the model from a set of potential variables. Several nonparametric variable selection techniques exist (Tcherning and Yang, 2000; Vieu, 1995; Tjøstheim and Auestad, 1994; Yao and Tong, 1994; Auestad and Tjøstheim, 1990), but they are computationally very demanding, in particular when the number of observations is not small. In this paper variable selection is carried out by linearizing the model and applying well-known

techniques of linear variable selection to this approximation. This keeps computational cost to a minimum. For this purpose we adopt the simple procedure proposed in Rech, Teräsvirta, and Tschernig (2001). Their idea is to approximate the stationary nonlinear model by a polynomial of sufficiently high order. Adapted to the present situation, the first step is to approximate function $G(\mathbf{x}_t;\boldsymbol{\psi})$ in (4) by a general k-th order polynomial. By the Stone-Weierstrass theorem, the approximation can be made arbitrarily accurate if some mild conditions, such as the parameter space $\boldsymbol{\psi}$ being compact, are imposed on function $G(\mathbf{x}_t;\boldsymbol{\psi})$. Thus the L^{2}GNN is approximated by another function. This yields

$$\begin{aligned} G(\mathbf{x}_t;\boldsymbol{\psi}) = \boldsymbol{\pi}'\widetilde{\mathbf{x}}_t + \sum_{j_1=1}^{q}\sum_{j_2=j_1}^{q}\theta_{j_1 j_2}x_{j_1,t}x_{j_2,t} \\ + \cdots + \sum_{j_1=1}^{q}\cdots\sum_{j_k=j_{k-1}}^{q}\theta_{j_1\ldots j_k}x_{j_1,t}\cdots x_{j_k,t} + R(\mathbf{x}_t;\boldsymbol{\psi}), \end{aligned} \tag{16}$$

where $\widetilde{\mathbf{x}}_t = [1, \mathbf{x}_t']'$ and $R(\mathbf{x}_t;\boldsymbol{\psi})$ is the approximation error that can be made negligible by choosing k sufficiently high. The $\theta' s$ are parameters, and $\boldsymbol{\pi} \in \mathbb{R}^{q+1}$ is a vector of parameters. The linear form of the approximation is independent of the number of neuron-pairs in (4).

In equation (16), every product of variables involving at least one redundant variable has the coefficient zero. The idea is to sort out the redundant variables by using this property of (16). In order to do that, we first regress y_t on all variables on the right-hand side of equation (16) assuming $R(\mathbf{x}_t;\boldsymbol{\psi}) = 0$ and compute the value of a model selection criterion (MSC), AIC (Akaike, 1974) or SBIC (Schwarz, 1978) for example. After doing that, we remove one variable from the original model and regress y_t on all the remaining terms in the corresponding polynomial and again compute the value of the MSC. This procedure is repeated by omitting each variable in turn. We continue by simultaneously omitting two regressors of the original model and proceed in that way until the polynomial is of a function of a single regressor and, finally, just a constant. Having done that, we choose the combination of variables that yields the lowest value of the MSC. This amounts to estimating $\sum_{i=1}^{q}\binom{q}{i} + 1$ linear models by ordinary least squares (OLS). Note that by following this procedure, the variables for the whole L^{2}GNN model are selected at the same time. Rech, Teräsvirta, and Tschernig (2001) showed that the procedure works well already in small samples when compared to well-known nonparametric techniques. Furthermore, it can be successfully applied even in large samples when nonparametric model selection becomes computationally infeasible.

6.2. **Determining the number of neuron-pairs.** In real applications, the number of neuron-pairs is not known and should be estimated from the data. In the neural network literature, a popular method for selecting the number of neuron is pruning, in which a model with a large number of neurons is estimated

first, and the size of the model is subsequently reduced by applying an appropriate technique such as cross-validation. Another technique used in this connection is regularization, which may be characterized as penalized maximum likelihood or least squares applied to the estimation of neural network models. For discussion see, for example, Fine (1999,pp. 215–221). Bayesian regularization may serve as an example (MacKay, 1992a; MacKay, 1992b).

Another possibility is to use a MSC to determine the number of hidden neuron-pairs. Swanson and White (1995), Swanson and White (1997a), and Swanson and White (1997b) apply the SBIC model selection criterion as follows. They start with a linear model, adding potential variables to it until SBIC indicates that the model cannot be further improved. Then they estimate models with a single hidden neuron and select regressors sequentially to it one by one unless SBIC shows no further improvement. Next, the authors add another hidden unit and proceed by adding variables to it. The selection process is terminated when SBIC indicates that no more hidden units or variables should be added or when a predetermined maximum number of hidden units has been reached. This modelling strategy can be termed fully sequential.

In this paper we adopt a similar strategy as described above. After the variables have been selected with the procedure described before, we start with a model with a single neuron-pair and compute the value of the SBIC. We continue adding neuron-pairs until the SBIC indicates no further improvement. The SBIC is defined as

$$\text{SBIC}(h) = \ln(\widehat{\sigma}^2) + \frac{\ln(T)}{T} \times \left[2m(2+q)\right], \tag{17}$$

where $\widehat{\sigma}^2$ is the estimated residual variance. This means that to choose a model with m neuron-pairs, we need to estimate $m+1$ models.

Another way of determining the number of neuron-pairs is to follow Medeiros and Veiga (2000b) and Medeiros, Teräsvirta, and Rech (2002) and use a sequence of Lagrange Multiplier tests. However, this is beyond the scope of this paper.

7. Numerical Examples

In this section we present numerical results for the L^2GNN model with real time series data. The first example considers only in-sample fitting and the second shows one-step ahead forecasts. The modelling cycle strategy described before was used to select the models.

7.1. **The Canadian Lynx series.** The first data set analyzed is the classic 10-based logarithm of the number of Canadian Lynx trapped in the Mackenzie River district of North-west Canada over the period 1821–1934. For further details and a background history see Tong (1990,Chapter 7). Some previous analysis

of this series can be found in Ozaki (1982), Tsay (1989), Teräsvirta (1994), and Xia and Li (1999). We start selecting the variables of the model among the first seven lags of the time series. With the procedure describe in Section 6.1 and using the SBIC, we identified lags 1 and 2 and with the AIC, lags 1,2,3,5,6, and 7. We continue building a L^{2}GNN model with only lags 1 and 2, which is more parsimonious. The final estimated mode has 2 neuron-pairs ($m = 2$), and when compared to a linear AR(2) model, the ratio between the standard deviation of the residuals from the nonlinear model and linear one is $\frac{\widehat{\sigma}}{\widehat{\sigma}_L} = 0.876$.

The estimated residual standard deviation ($\widehat{\sigma} = 0.204$) is smaller than in other models that use only the first two lags as variables. For example, the nonlinear model proposed by (Tong, 1990,p. 410), has a residual standard deviation of 0.222, and the Exponential AutoRegressive (EXPAR) model proposed by (Ozaki, 1982) has $\hat{\sigma}_\varepsilon = 0.208$.

7.2. **The Sunspot Series.** In this example we consider the annual sunspot numbers over the period 1700–1998. The observations for the period 1700–1979 were used to estimate the model and the remaining were used to forecast evaluation. We adopted the same transformation as in Tong (1990), $y_t = 2\left[\sqrt{(1+N_t)}-1\right]$, where N_t is the sunspot number. The series was obtained from the National Geophysical Data Center web page. [2] The sunspot numbers are a heavily modelled nonlinear time series: for a neural network example see Weigend, Huberman, and Rumelhart (1992).

We begin the L^{2}GNN modelling of the series by selecting the relevant lags using the variable selection procedure described in Section 6.1. We use a third-order polynomial approximation to the true model. Applying SBIC, lags 1,2, and 7 are selected whereas AIC yields the lags 1,2,4,5,6,7,8,9, and 10. As in the previous example, we proceed with the lags selected by the SBIC. However, the residuals of the estimated model are strongly autocorrelated. The serial correlation is removed by also including y_{t-3} in the set of selected variables. When building the L^{2}GNN model we select the number of hidden neuron-pairs using the SBIC as described in Section 6.2.

After estimating a model with 3 neuron-pairs, we continue considering the out-of-sample performance of the estimated model. In order to assess the out-of-sample performance of the L^{2}GNN model we compare our one-step-ahead forecasting results with the ones obtained from the two SETAR models, the one reported in Tong (1990,p. 420) and the other in Chen (1995), an artificial neural network (ANN) model with 10 hidden neurons and the first 9 lags as input variables, estimated with Bayesian regularization (MacKay, 1992a; MacKay, 1992b), the Stochastic Neural Network (SNN) model estimated in Lai and Wong (2001), the Neuro-Coefficient STAR (NCSTAR) model of Medeiros and Veiga (2000a), and a linear autoregressive

[2] http://www.ngdc.noaa.gov/stp/SOLAR/SSN/ssn.html

model with lags selected using SBIC. The SETAR model estimated by Chen (1995) 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. The estimated SNN model of Lai and Wong (2001) can be viewed as a form of smooth transition autoregression with multivariate transition variables in the same spirit of the NCSTAR model of Medeiros and Veiga (2000a).

Table 1 shows the results of the one-step-ahead forecasting for the period 1980-1998, with the respective root mean squared error (RMSE) and mean absolute error (MAE). As shown in Table 1, the L^2GNN model has the smallest RMSE and MAE among the alternatives considered in this paper. Over 19 forecasts, the L^2GNN model outperforms the ANN and Tong´s SETAR models in 12 cases, the SETAR model of Chen (1995) in 15 cases, the AR specification in 11 cases, and the SNN and NCSTAR models in 10 cases.

8. CONCLUSIONS

In this paper we have proposed a new nonlinear time-series model based on neural networks. The model is called the Local Global Neural Network and can be interpreted as a mixture of experts model. The case of linear experts is analyzed in detail and its probabilistic and statistical properties were discussed. The proposed model consist of a mixture of stationary or non-stationary linear models and is able to describe "intermittent" dynamics: the system spends a large fraction of the time in a bounded region, but, sporadically, it develops an instability that grows exponentially for some time and then suddenly collapses. Intermittency is a commonly observed behavior in ecology and epidemiology, fluid dynamics and other natural systems. A specific-to-general model building strategy, based on the SBIC, has been suggested to determine the variables and the number of hidden neuron-pairs. When put into test in a real experiment concerning one-step-ahead forecasting, the proposed model outperforms the linear model and other nonlinear specifications considered in this paper, suggesting that the theory developed here is useful and the proposed model thus seems to be a useful tool for practicing time series analysts.

TABLE 1. 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.

		L^2NGG		ANN model		SETAR model (Tong, 1990)		SETAR model (Chen, 1995)		AR model		SNN		NCSTAR	
Year	Observation	Forecast	Error	Forecast	Error	Forecast	Error	Forecast	Error	Forecast	Error	Forecast	Error	Forecast	Error
1980	154.6	149.1	5.5	136.9	17.7	161.0	-6.4	134.3	20.3	159.8	-5.2	157.5	-2.9	132.0	23.0
1981	140.4	131.1	9.3	130.5	9.9	135.7	4.7	125.4	15.0	123.3	17.1	130.5	9.9	134.0	6.4
1982	115.9	101.8	14.1	101.1	14.8	98.2	17.7	99.3	16.6	99.6	16.3	106.3	9.6	94.9	21.0
1983	66.6	81.2	-14.6	88.6	-22.0	76.1	-9.5	85.0	-18.4	78.9	-12.3	77.3	-10.7	77.5	-10.9
1984	45.9	42.7	3.2	45.8	0.1	35.7	10.2	41.3	4.7	33.9	12.0	36.5	9.4	33.6	12.3
1985	17.9	22.4	-4.5	29.5	-11.6	24.3	-6.4	29.8	-11.9	29.3	-11.4	23.5	-5.6	24.5	-6.6
1986	13.4	10.0	3.4	9.5	3.9	10.7	2.7	9.8	3.6	10.7	2.7	8.8	4.6	12.6	0.8
1987	29.4	19.4	10.0	25.2	4.2	20.1	9.3	16.5	12.9	23.0	6.4	26.8	2.6	8.8	20.6
1988	100.2	71.9	28.3	76.8	23.4	54.5	45.7	66.4	33.8	61.2	38.9	68.1	32.1	84.3	16.0
1989	157.6	160.7	-3.1	152.9	4.6	155.8	1.8	121.8	35.8	159.2	-1.6	167.4	-9.8	142.4	15.2
1990	142.6	145.9	-3.3	147.3	-4.7	156.4	-13.8	152.5	-9.9	175.5	-32.9	168.6	-26.0	144.3	-1.7
1991	145.7	118.1	27.5	121.2	24.5	93.3	52.4	123.7	22.0	119.1	26.6	118.6	27.1	127.1	18.6
1992	94.3	101.8	-7.5	114.3	-20.0	110.5	-16.2	115.9	-21.7	118.9	-24.6	110.1	-15.8	105.3	-11.0
1993	54.6	69.3	-14.7	71.0	-16.4	67.9	-13.3	69.2	-14.6	57.9	-3.3	60.8	-6.2	66.5	-11.9
1994	29.9	29.8	0.1	32.9	-3.0	27.0	2.9	35.7	-5.8	29.9	-0.1	27.7	2.2	25.0	4.9
1995	17.5	14.0	3.5	19.2	-1.7	18.4	-0.9	18.9	-1.4	17.6	-0.1	14.3	3.2	19.1	-1.6
1996	8.6	14.8	-6.2	10.2	-1.6	18.1	-9.5	11.6	-3.0	15.7	-7.1	11.7	-3.1	8.3	0.3
1997	21.5	17.2	4.3	21.3	0.2	12.3	9.2	11.8	9.7	16.0	5.5	24.2	-2.7	13.3	8.2
1998	64.3	63.9	0.4	67.6	-3.3	46.7	17.6	58.5	5.8	52.5	11.8	56.2	8.1	66.9	-2.6
RMSE			11.7		13.8		18.7		16.9		16.5		13.3		12.4
MAE			8.6		11.2		13.1		14.0		12.4		10.1		10.2

ACKNOWLEDGMENTS

This work is partly based on the first author´s Ph.D. Thesis at the Department of Electrical Engineering at the Pontifical Catholic University of Rio de Janeiro. The authors would like to thank Marcelo O. Magnasco, Maurício Romero Sicre, Juan Pablo Torres-Martínez, and Alvaro Veiga for valuable discussions, two anonymous referees and an associate editor for helpful comments, and the CNPq for the partial financial support.

APPENDIX A: PROOFS

8.1. **Lemmas.**

LEMMA 1. *If the functions* $\varphi^{(\ell)}(x) = hx - \gamma\beta^{(\ell)}$, $\ell = 1, 2$, $x \in \mathbb{R}$, $h > 0$, $\beta^{(1)} < \beta^{(2)}$ *are not equivalent in sign, the class of functions* $\{B(x;\boldsymbol{\psi}_B)\} \bigcup \{xB(x;\boldsymbol{\psi}_B)\}$, *where*

$$B(x;\boldsymbol{\psi}_B) = -\left\{\left[1+\exp\left(\varphi^{(1)}(x)\right)\right]^{-1} - \left[1+\exp\left(\varphi^{(2)}(x)\right)\right]^{-1}\right\},$$

is linearly independent.

LEMMA 2. *Let* $\{\mathbf{d}_i\}$ *be a family of vectors in* $\mathbb{R}^q$ *such that* $d_{i1} > 0$ *for every* i. *Let* $\mathbf{v}$ *be the unitary vector that, according to Hwang and Ding (1997), exists and satisfies:*

(1) $\langle \mathbf{d}_i, \mathbf{v}\rangle > 0$ *and*

(2) *if* $\mathbf{d}_i \neq \mathbf{d}_j$ *then* $\langle \mathbf{d}_i, \mathbf{v}\rangle \neq \langle \mathbf{d}_j, \mathbf{v}\rangle$.

Thus it follows that there exists a vector base $\mathbf{v}_1, \ldots, \mathbf{v}_q$ *that satisfies the same conditions.*

8.2. **Proofs of Theorems.**

8.2.1. *Proof of Theorem 1.* Write model (4) as

$$\mathbf{Y}_t = \mathbf{a}_{t-1} + \mathbf{A}_{t-1}\mathbf{Y}_{t-1} + \mathbf{e}_t, \tag{18}$$

where

$$\mathbf{Y}_t = \begin{pmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{pmatrix}, \ \mathbf{Y}_{t-1} = \begin{pmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \end{pmatrix}, \ \mathbf{e}_t = \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \ \mathbf{a}_{t-1} = \begin{pmatrix} \sum_{i=1}^{m} b_i B\left(\mathbf{Y}_{t-1}\right) \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\mathbf{A}_{t-1} = \begin{pmatrix} \sum_{i=1}^{m} a_{i1} B_i\left(\mathbf{Y}_{t-1}\right) & \sum_{i=1}^{m} a_{i2} B_i\left(\mathbf{Y}_{t-1}\right) & \cdots & \sum_{i=1}^{m} a_{ip-1} B_i\left(\mathbf{Y}_{t-1}\right) & \sum_{i=1}^{m} a_{ip} B_i\left(\mathbf{Y}_{t-1}\right) \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

and $B_i\left(\mathbf{Y}_{t-1}\right) \equiv B\left(\mathbf{Y}_{t-1}; \boldsymbol{\psi}_{B_i}\right)$.

After recursive substitutions, model (18) can be written as

$$\mathbf{Y}_t = \mathbf{a}_{t-1} + \sum_{i=0}^{t-2} \left[\prod_{j=i+1}^{t-1} \mathbf{A}_j \right] \mathbf{a}_i + \left[\prod_{j=0}^{t-1} \mathbf{A}_j \right] \mathbf{Y}_0 + \sum_{i=1}^{t-1} \left[\prod_{j=i}^{t-1} \mathbf{A}_j \right] \mathbf{e}_i + \mathbf{e}_t. \tag{19}$$

Model (19) will be asymptotically stationary if $\prod_t \mathbf{A}_t \to \mathbf{0}$ as $t \to \infty$. This will be of course the case if Condition 1 in Theorem 1 is satisfied. As

$$B_i\left(\mathbf{Y}_t\right) = -\left\{ \left[1 + \exp\left(\gamma_i \left(\langle \mathbf{d}_i, \mathbf{Y}_t \rangle - \beta_i^{(1)}\right)\right)\right]^{-1} - \left[1 + \exp\left(\gamma_i \left(\langle \mathbf{d}_i, \mathbf{Y}_t \rangle - \beta_i^{(2)}\right)\right)\right]^{-1} \right\},$$

$\prod_t \mathbf{A}_t \to \mathbf{0}$ if $B_i\left(\mathbf{Y}_t\right) \to 0$, $i = 1, \ldots, m$. This will be true if $|\langle \mathbf{d}_i, \mathbf{Y}_t \rangle| \to M$, where $M >> \max\left(\beta_i^{(1)}, \beta_i^{(2)}\right)$. If at least one limiting AR regime is explosive then $|\langle \mathbf{d}_i, \mathbf{Y}_t \rangle| \to \infty$ as far as $d_{ij} \neq 0$ (Condition 2 in Theorem 1). When a given limiting AR regime has unit-roots in order to guarantee that $|\langle \mathbf{d}_i, \mathbf{Y}_t \rangle| \to M$, the vectors $\mathbf{d}_i$ must not be orthogonal to the eigenvectors of the respective transition matrix (Condition 3 in Theorem 1).

Q.E.D

8.2.2. *Proof of Theorem 2.* Lemma 2 of Jennrich (1969) shows that the conditions (1)–(3) in Theorem 2 are enough to guarantee the existence (and measurability) of the LSE. In order to apply this result to the L^2GNN model we have to check if the above conditions are satisfied by the model.

Condition (3) in Theorem 2 was already assumed when defining the model. It is easy to prove in our case that $G\left(\mathbf{x}_t; \boldsymbol{\psi}\right)$ is continuous in the parameter vector $\boldsymbol{\psi}$. This follows from the fact that $B\left(\mathbf{x}_t; \boldsymbol{\psi}_{B_i}\right)$

and $L_i(\mathbf{x}_t;\boldsymbol{\psi}_L)$, $i=1,\ldots,m$, depend continuously on $\boldsymbol{\psi}_B$ and $\boldsymbol{\psi}_L$ for each value of $\mathbf{x}_t$. Similarly, we can see that $G(\mathbf{x}_t,\boldsymbol{\psi})$ is continuous in $\mathbf{x}_t$, and therefore measurable, for each fixed value of the parameter vector $\boldsymbol{\psi}$. Thus (1) and (2) are satisfied.

Q.E.D

8.2.3. *Proof of Theorem 3.* Suppose that $\widetilde{\boldsymbol{\psi}}=\left[\widetilde{\boldsymbol{\psi}}_L',\widetilde{\boldsymbol{\psi}}_B'\right]'$ is another vector of parameters such that

$$\sum_{i=1}^{m}\left(\mathbf{a}_i'\mathbf{x}_t+b_i\right)B\left(\mathbf{x}_t;\boldsymbol{\psi}_{B_i}\right)=\sum_{i=1}^{m}\left(\widetilde{\mathbf{a}}_i'\mathbf{x}_t+\widetilde{b}_i\right)B\left(\mathbf{x}_t;\widetilde{\boldsymbol{\psi}}_{B_i}\right). \tag{20}$$

In order to show global identifiability of the L^{2}GNN model, we need to prove that, under Assumption 1 and restrictions (R.1)–(R.4), (20) is satisfied if, and only if, $\mathbf{a}_i=\tilde{\mathbf{a}}_i$, $b_i=\tilde{b}_i$, and $\boldsymbol{\psi}_B=\widetilde{\boldsymbol{\psi}}_B$, $i=1,\ldots,m$, $\forall\mathbf{x}_t\in\mathbb{R}^q$.

Equation (20) can be rewritten as

$$\sum_{j=1}^{2m}\left(\mathbf{c}_j'\mathbf{x}_t+e_j\right)B\left(\mathbf{x}_t,\breve{\boldsymbol{\psi}}_{B_j}\right)=0, \tag{21}$$

where $B\left(\mathbf{x}_t,\breve{\boldsymbol{\psi}}_{B_j}\right)=B\left(\mathbf{x}_t,\boldsymbol{\psi}_{B_j}\right)$ for $j=1,\ldots,m$, $B\left(\mathbf{x}_t,\breve{\boldsymbol{\psi}}_{B_j}\right)=B\left(\mathbf{x}_t,\widetilde{\boldsymbol{\psi}}_{B_{j-m}}\right)$, for $j=m+1,\ldots,2m$, $\mathbf{c}_j=\mathbf{a}_j$, for $j=1,\ldots,m$, $\mathbf{c}_j=-\widetilde{\mathbf{a}}_{j-m}$, for $j=m+1,\ldots,2m$, $e_j=b_j$, for $j=1,\ldots,m$, and $e_j=-\widetilde{b}_{j-m}$, for $j=m+1,\ldots,2m$.

To relate this problem to Lemma 1, we reduce the dimension of $\mathbf{x}_t$ to one. Following Hwang and Ding (1997), let $\mathbf{v}$ be the unit vector such that for distinct $\mathbf{d}_i$s, the projections over $\mathbf{v}$ are likewise different. Since the set $\{\mathbf{d}_1,\ldots,\mathbf{d}_m\}$ has a finite number of points, $\gamma_i>0$ (restriction (R.3)), and $d_{i1}>0$ (restriction (R.4)), $i=1,\ldots,m$, it is possible to construct a vector $\mathbf{v}$ such that the projection $h_i=\gamma_i\langle\mathbf{d}_i,\mathbf{v}\rangle$ is positive. Replacing $\mathbf{x}_t$ in (21) by $x_t\mathbf{v}$, $x_t\in\mathbb{R}$, leads to

$$\sum_{j=1}^{2m}\left(\overline{c}_jx_t+e_j\right)B\left(x_t\mathbf{v},\breve{\boldsymbol{\psi}}_{B_j}\right)=0, \tag{22}$$

where $\overline{c}_j=\langle\mathbf{c}_j,\mathbf{v}\rangle$.

For simplicity of notation let $\varphi_j^{(\ell)}=\varphi\left(\mathbf{x}_t;\boldsymbol{\theta}_{j\ell}\right)$, $j=1,\ldots,2m$. Lemma 1 imply that if $\varphi_{j_1}^{(\ell)}$ and $\varphi_{j2}^{(\ell)}$ are not sign-equivalent, $j_1\in\{1,\ldots,2m\}$, $j_2\in\{1,\ldots,2m\}$, (22) holds if, and only if, $\overline{c}_j$ and e_j vanish jointly for every $j\in\{1,\ldots,2m\}$. However, the condition $\overline{c}_j$, $j=1,\ldots,2m$, does not imply that $\mathbf{c}_j=\mathbf{0}$. Lemma 2 shows in fact that vector $\mathbf{v}$ is not unique and that there exists vectors $\mathbf{v}_1,\ldots,\mathbf{v}_q$ that satisfy the same conditions as $\mathbf{v}$ and form a basis of $\mathbb{R}^q$. Then the inner product $\langle\mathbf{c}_i,\mathbf{v}_j\rangle=0$, $\forall j$, implying that $\mathbf{c}_i=\mathbf{0}$. However, Assumption 1 precludes that possibility. Hence, $\varphi_{j_1}^{(\ell)}$ and $\varphi_{j2}^{(\ell)}$ must be

sign-equivalent. But restrictions (R.2)–(R.4) avoid that two functions $\varphi_{j_1}^{(\ell)}$ and $\varphi_{j_2}^{(\ell)}$ coming from the same model being sign-equivalent. Consequently, $\exists\, j_1 \in \{1,\ldots,m\}$ and $j_2 \in \{m+1,\ldots,2m\}$ such that $\varphi_{j_1}^{(\ell)}$ and $\varphi_{j_2}^{(\ell)}$, $\ell = 1,2$ are sign-equivalent. Under restrictions (R.2)–(R.4) the only possibility is that the hidden neuron-pairs are permuted. Restriction (R.1) excludes that possibility. Hence, the only case where (20) holds is when $\mathbf{a}_i = \tilde{\mathbf{a}}_i$, $b_i = \tilde{b}_i$, and $\boldsymbol{\psi}_B = \widetilde{\boldsymbol{\psi}}_B$, $i = 1,\ldots,m$, $\forall \mathbf{x}_t \in \mathbb{R}^q$.

Q.E.D

8.2.4. *Proof of Theorem 4.* For the proof of this theorem we use Theorem 3.5 of White (1994), 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 $q(\mathbf{x}_t;\boldsymbol{\psi}) = [y_t - G(\mathbf{x}_t;\boldsymbol{\psi})]^2$. Assumption 3.1a states that for each $\boldsymbol{\psi} \in \Psi$, $-\mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi}))$ exists and is finite, $t = 1,\ldots,T$. Under the conditions of Theorem 3 and the fact that ε_t is a zero mean normally distributed random variable with finite variance, hence k-integrable, Assumption 3.1a follows.

Assumption 3.1b states that $-\mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi}))$ is continuous in Ψ, $t = 1,\ldots,T$. Let $\boldsymbol{\psi} \to \boldsymbol{\psi}^*$, since for any t, $G(\mathbf{x}_t;\boldsymbol{\psi})$ is continuous on Ψ, then $q(\mathbf{x}_t;\boldsymbol{\psi}) \to q(\mathbf{x}_t;\boldsymbol{\psi}^*)$, $\forall t$ (pointwise convergence). From the continuity of $G(\mathbf{x}_t,\boldsymbol{\psi})$ on the compact set Ψ, we have uniform continuity and we obtain that $q(\mathbf{x}_t;\boldsymbol{\psi})$ is dominated by an integrable function dF. Then, by Lebesgue´s Dominated Convergence Theorem, we get $\int q(\mathbf{x}_t;\boldsymbol{\psi})\, dF \to \int q(\mathbf{x}_t;\boldsymbol{\psi}^*)\, dF$, and $\mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi}))$ is continuous.

Assumption 3.1c states that $-\mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi}))$ obeys the strong (weak) Uniform Law of Large Numbers (ULLN). Lemma A2 of Pötscher and Prucha (1986) guarantees that $\mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi}))$ obeys the strong law of large numbers. The set of hypothesis (b) of this lemma is satisfied:

(1) we are working with a strictly stationary and ergodic process;

(2) from the continuity of $\mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi}))$ and from the compactness of Ψ we have that $\inf \mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi})) = \mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi}^*))$ for $\boldsymbol{\psi}^* \in \Psi$, and with Assumption 3.1a we may guarantee that $\mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi}^*))$ exists and is finite, getting that $\inf \mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi})) > -\infty$.

Assumption 3.2 is related to the unique identifiability of $\boldsymbol{\psi}^*$. In Theorem 3, we have showed that under Assumption 1 and with the restrictions (R.1)–(R.4) imposed, the L^2GNN is globally identifiable.

Q.E.D

8.2.5. *Proof of Theorem 5.* We use Theorem 6.4 of White (1994) and check its assumptions.

Assumptions 2.1, 2.3, and 3.1 follow from the proof of Theorem 4 (consistency).

Assumptions 3.2' and 3.6 follow from the fact that $G(\mathbf{x}_t;\boldsymbol{\psi})$ 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 $\mathrm{E}\left(\nabla Q_n(\boldsymbol{\psi})\right) < \infty$ and $\mathrm{E}\left(\nabla^2 Q_n(\boldsymbol{\psi})\right) < \infty$, $\forall n$. The expected gradient and the expected Hessian of $Q_n(\boldsymbol{\psi})$ are given by

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

and

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

respectively.

Assumptions 3.7a and 3.8a follow considering the normality condition on ε_t, the properties of the function $G\left(\mathbf{x}_t;\boldsymbol{\psi}\right)$, and the fact that $\nabla G\left(\mathbf{x}_t;\boldsymbol{\psi}\right)$ and $\nabla^2 G\left(\mathbf{x}_t;\boldsymbol{\psi}\right)$ contains at most terms of order $x_{i,t}x_{j,t}$, $i = 1,\ldots,q$, $i = 1,\ldots,q$. Following the same reasoning used in the proof of Assumptions 3.1a in Theorem 4, Assumptions 3.7a and 3.8a hold.

Assumption 3.8b: Under Assumption 4, the fact that the function $G\left(\mathbf{x}_t;\boldsymbol{\psi}\right)$ is continuous, and dominated convergence, Assumption 3.8b follows.

Assumption 3.8c: The proof of Theorem 4 and the ULLN from Pötscher and Prucha (1986) yields the result.

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

Assumption 6.1: Using Theorem 2.4 from White and Domowitz (1984) we can show that the sequence $2\boldsymbol{\xi}'\nabla G\left(\mathbf{x}_t;\boldsymbol{\psi}^*\right)\varepsilon_t$ 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 White and Domowitz (1984) hold because ε_t is NID. Assumption A(ii) holds with $V = 4\sigma^2\boldsymbol{\xi}'\mathrm{E}\left(\nabla G\left(\mathbf{x}_t;\boldsymbol{\psi}^*\right)\nabla' G\left(\mathbf{x}_t;\boldsymbol{\psi}^*\right)\right)$. Furthermore, since any measurable transformation of mixing processes is itself mixing (see Lemma 2.1 in White and Domowitz (1984)), $2\boldsymbol{\xi}'\nabla G\left(\mathbf{x}_t;\boldsymbol{\psi}^*\right)\varepsilon_t$ is a strong mixing sequence and obeys the CLT. By using the Cramér-Wold device $\nabla Q\left(\mathbf{x}_t;\boldsymbol{\psi}\right)$ also obeys the CLT with covariance matrix $B_n^* = 4\sigma^2\mathrm{E}\left(\nabla G\left(\mathbf{x}_t;\boldsymbol{\psi}^*\right)\nabla' G\left(\mathbf{x}_t;\boldsymbol{\psi}^*\right)\right) = 2\sigma^2 A_n^*$ which is $O(1)$ and non-singular.

Q.E.D

8.3. **Proofs of Lemmas.**

8.3.1. *Proof of Lemma 1.* In order to make the proof clearer let $\varphi_i^{(\ell)}(x) = \left(h_i x - \gamma_i \beta_i^{(\ell)}\right)$, where $h_i = \gamma_i\left\langle \mathbf{d}_i, \mathbf{v}\right\rangle$, and write $B\left(x;\boldsymbol{\psi}_{B_i}\right)$ as $B\left(\varphi_i^{(1)}(x), \varphi_i^{(2)}(x)\right)$. Let n be a positive integer. We should prove

that if there are scalars λ_i, ω_i, $\gamma_i > 0$, $h_i > 0$, and $\beta_i^{(1)} < \beta_i^{(2)}$, $i = 1, \ldots, n$, with $\left(h_i, \gamma_i, \beta_i^{(1)}, \beta_i^{(2)}\right) \neq \left(h_j, \gamma_j, \beta_j^{(1)}, \beta_j^{(2)}\right)$ for $i \neq j$ (due to their not being equivalent in sign) such that $\forall x \in \mathbb{R}$ we have:

$$\sum_{i=1}^{n} (\lambda_i + \omega_i x)\, B\left(\varphi^{(1)}(x), \varphi^{(2)}(x)\right) = 0, \tag{23}$$

then $\lambda_i = \omega_i = 0$, $i = 1, \ldots, n$.

Considering that $B\left(\varphi_i^{(1)}(x), \varphi_i^{(2)}(x)\right) = F\left(-\varphi_i^{(1)}(x)\right) - F\left(-\varphi_i^{(2)}(x)\right)$, where $F(\cdot)$ is the logistic function, (23) is equivalent to:

$$\sum_{i=1}^{n} (\lambda_i + \omega_i x) \left[F\left(-\varphi_i^{(1)}(x)\right) - F\left(-\varphi_i^{(2)}(x)\right)\right] = 0. \tag{24}$$

Developing the Taylor series of $F\left(-\varphi_i^{(\ell)}(x)\right)$, $\ell = 1, 2$, we have:

$$F\left(-\varphi_i^{\ell}(x)\right) = \sum_{k=1}^{\infty} (-1)^k\, e^{-k\gamma_i \beta_i^{(\ell)}} e^{k h_i x}. \tag{25}$$

The series converges absolutely when $e^{-k\gamma_i\beta_i^{(\ell)}} < 1$, that is for $x < \left(\frac{\gamma_i \beta_i^{(\ell)}}{h_i}\right)$. Therefore, there exist M small enough such that (25) converges for every $x \in (-\infty, M)$. Substituting (25) in (24) and writing $C_i^{(\ell)} = \gamma_i \beta_i^{(\ell)}$ we obtain:

$$\sum_{i=1}^{n} \left\{ (\lambda_i + \omega_i x) \sum_{k=1}^{\infty} (-1)^k \left[e^{-C_i^{(1)}} - e^{-C_i^{(2)} k}\right] e^{k h_i x} \right\} = 0. \tag{26}$$

Notice that due to the fact that γ_i is positive, then $C_i^{(1)} < C_i^{(2)}$. Denoting $W_i^{(\ell)} = -e^{-C_i^{(\ell)}}$, $\ell = 1, 2$, we have that $W_i^{(1)} < W_i^{(2)}$ and substituting in (26):

$$\sum_{i=1}^{n} \left\{ (\lambda_i + \omega_i x) \sum_{k=1}^{\infty} (-1)^k \left[\left(W_i^{(1)}\right)^k - \left(W_i^{(2)}\right)^k\right] e^{k h_i x} \right\} = 0.$$

This series can be written (as it is absolutely convergent) as:

$$\sum_{k=1}^{\infty} \alpha_k^* e^{h_k^* x} + \alpha_k^{**} x e^{h_k^* x} = 0, \tag{27}$$

where $h_1^* < h_2^* < \cdots < h_\infty^*$, and each h_i^* is an integer multiple of some h_j. However, we can prove that $\alpha_k^* = \alpha_k^{**} = 0$.

Dividing (27) by $x e^{h_1^* x}$, we obtain

$$\sum_{k=1}^{\infty} \left\{ \alpha_k^* e^{x(h_k^* - h_1^*)} + \alpha_k^{**} \frac{e^{x(h_k^* - h_1^*)}}{x} \right\} = 0 \tag{28}$$

assuming the limit in (28) as $x \to \infty$ and considering that $h_k^* - h_1^* > 0$, for $k \neq 1$, we conclude that $\alpha_1^* = 0$. Considering the expression (27) with $\alpha_1^* = 0$ and dividing by $e^{h_1^*}$ we obtain

$$\alpha_1^{**} + \sum_{k=2}^{\infty} \left(\alpha_k^* + x\alpha_k^{**}\right) e^{x(h_k^* - h_1^*)} = 0.$$

Now, taking the limit when $x \to -\infty$, the terms in the sum go to zero and we obtain $\alpha_1^{**} = 0$. Repeating this procedure we will thus obtain that $\alpha_k^* = \alpha_k^{**} = 0$.

There is still left to prove that starting from $\alpha_k^* = \alpha_k^{**} = 0$ it follows that $\lambda_i = \omega_i = 0$. The expressions for λ_i and ω_i in terms of α_k^* and α_k^{**} are similar, so we will present only the proof for α_k^*.

Let $J = \{j \in \{1, \ldots, m\} : h_j = h_1\}$. We should prove that $\lambda_j = \omega_j = 0$, $\forall j \in J$. For each $s \in \mathbb{N}$, there exist k_s, such that $h_{k_s}^* = sh_1$. Also there exists an integer $N > 0$ such that for every ℓ and $i \geq 2$, $(1 + N\ell)h_1$ is not an integer multiple of h_i. Denote $\theta_i = \frac{h_1}{h_i}$. As $0 < h_1 < h_i$, θ_i is a non-integer number less than 1. So, we have to prove that there are a sequence K_n such as for all $i \geq 2$, $K_n\theta_i$ is not an integer. Let $J_Z = \{j \in J | \exists r \text{ integer, such that } r\theta_j \in \mathbb{Z}\}$. Select $K = \prod_{j \in J_Z} r_j$. Then, the sequence $K_n = (1 + nK)$ satisfies the desired statement. If $i \in J_Z$, then $K_n\theta_i = \theta_i + n \prod_{j \in J_Z, j \neq i} (r_j)\, r_i\theta_i$, where $r_i\theta_i$, $\prod_j r_j$ and n are all integer numbers and θ_i is a non-integer, so $K_n\theta_i$ cannot be an integer number. Otherwise, if $i \notin J_Z$, then there are no integer number such as as $K_n\theta_i$ would be an integer. As K_n is an integer number, then $K_n\theta_i$ is not an integer.

For each k_s it is satisfied that $\alpha_{k_s}^* = 0$, in particular for $s = (1 + N\ell)$ we have:

$$\alpha_{k_s}^* = \sum_{j \in J} \lambda_j \left[\left(W_j^{(1)}\right)^s - \left(W_j^{(2)}\right)^s \right] = 0,$$

that is

$$\sum_{j \in J} \lambda_j \left(W_j^{(1)}\right)^s = \sum_{j \in J} \lambda_j \left(W_j^{(2)}\right)^s. \tag{29}$$

If $j \in J$ then $h_j = h_{i_0}$ and due to the definition of the h_i's this can only happen if $\forall j \in J$, $d_j = d_{i_0}$, then it follows that $d_j = d_{i_0}$ and $\gamma_j = \gamma_{i_0}$. Considering that $\left(h_i, \gamma_i, \beta_i^{(1)}, \beta_i^{(2)}\right) \neq \left(h_j, \gamma_j, \beta_j^{(1)}, \beta_j^{(2)}\right)$ it follows that $\beta_i^{(1)} \neq \beta_j^{(1)}$, $\beta_i^{(2)} \neq \beta_j^{(2)}$ we have then that obtaining that $\forall j, j' \in J$, $j \neq j'$: $W_j^{(\ell)} \neq W_{j'}^{(\ell)}$; and considering that $\beta_j^{(1)} < \beta_j^{(2)}$, it follows that $W_j^{(1)} < W_j^{(2)}$, $\forall j \in J$.

Let n_J be the cardinal of J and $\phi : \{1, \ldots, n_J\} \to J$ a reordering of J such that $W_{\phi(1)}^{(1)} < W_{\phi(2)}^{(1)} < \cdots < W_{\phi(n_J)}^{(1)}$ and $W_{\phi(1)}^{(2)} < W_{\phi(2)}^{(2)} < \cdots < W_{\phi(n_J)}^{(2)}$. Dividing (29) by $W_{\phi(n_J)}^{(2)}$ and passing to the limit as

$k \to \infty$ we have

$$\lim_{k\to\infty}\left(\sum_{j=1}^{n_J}\left(\frac{W^{(1)}_{\phi(j)}}{W^{(2)}_{\phi(n_J)}}\right)^k\right) = a_{\phi(n_J)} + \lim_{k\to\infty}\left(\sum_{j=1}^{n_J-1}\left(\frac{W^{(2)}_{\phi(j)}}{W^{(2)}_{\phi(n_J)}}\right)^k\right)$$

and from this we obtain $a_{\phi(n_J)} = 0$. Repeating this procedure, we obtain $a_{\phi(n_J-1)} = \cdots = a_{\phi(1)} = 0$. Considering $i = 2, \ldots, m$ and with the corresponding set J that defines group J and following an identical line of reasoning, we arrive at the conclusion that $\lambda_i = 0$, $i = 1, \ldots, m$. Similarly , we obtain $\omega_i = 0$, $i = 1, \ldots, m$.

Q.E.D

8.3.2. *Proof of Lemma 2.* Let $\mathbf{v}_0$ be a unitary vector such that for different $\mathbf{d}_i$s, the projections on $\mathbf{v}_0$, $b_i = \langle \mathbf{d}_i, \mathbf{v}_0 \rangle$ are also different and positive. We should find a vector base $\mathbf{v}_1, \ldots, \mathbf{v}_q$ such that these vectors satisfy the same conditions as $\mathbf{v}_0$. Let $\mathbf{v}_0$ be given, let us define the $\mathbf{v}_j$s as:

$$\mathbf{v}_1 = \mathbf{v}_0,\ \mathbf{v}_2 = \mathbf{v}_0 - \delta_2\mathbf{e}_2,\ \mathbf{v}_3 = \mathbf{v}_0 - \delta_3\mathbf{e}_3, \ldots, \mathbf{v}_q = \mathbf{v}_0 - \delta_q\mathbf{e}_q, \tag{30}$$

where $\mathbf{e}_j$ is the canonical vector with 1 in position j and zero otherwise and δ_j is small enough. We should prove (1) that they satisfy the conditions of Lemma 2 and (2) that they form a vector base of the space. For every j, the projection of the $\mathbf{d}_i$s on $\mathbf{v}_i$ is $b_i = \langle \mathbf{d}_i, \mathbf{v}_j \rangle = \langle \mathbf{d}_i, \mathbf{v}_0 \rangle + \delta_j d_{ij}$, where the first terms in the sums are always positive and different when the $\mathbf{d}_i$s are different. Therefore, we can choose δ_j small enough such that $b_i = \langle \mathbf{d}_i, \mathbf{v}_j \rangle$ remains positive and different for different $\mathbf{d}_i$s. To show that the q vectors already defined form a vector base it is enough to show that they are linearly independent. Let us consider an arbitrary linear combination of these vectors equal to zero:

$$\sum_{j=1}^{q}\alpha_j\mathbf{v}_j = 0 \Rightarrow \alpha_1\mathbf{v}_0 + \sum_{j=2}^{q}\alpha_j\left(\mathbf{v}_0 - \delta_j\mathbf{e}_j\right) = 0 \Rightarrow \mathbf{v}_0\sum_{j=1}^{q}\alpha_j - \sum_{j=2}^{q}\alpha_j\delta_j\mathbf{e}_j = 0. \tag{31}$$

From this it follows that:

$$\mathbf{v}_0\sum_{j=1}^{q}\alpha_j = \sum_{j=2}^{q}\alpha_j\delta_j\mathbf{e}_j. \tag{32}$$

Writing the previous equality for the first component of each vector and taking in to consideration that the left member contains sums of the canonical vectors from 2 to q, we have that:

$$\left(\mathbf{v}_0\sum_{j=1}^{q}\alpha_j\right)_1 = \left(\sum_{j=2}^{q}\alpha_j\delta_j\mathbf{e}_j\right)_1 = 0, \tag{33}$$

since $v_{01}\sum_{j=1}^{q}\alpha_j = 0$ and $v_{01} \neq 0$. Writing (33) for the component k, $k = 2, 3, \ldots, q$, we have that:

$$0 = \left(\sum_{j=2}^{q}\alpha_j\delta_j\mathbf{e}_j\right)_k = \alpha_k + \delta_k \Rightarrow \alpha_k = 0,\ k = 2, \ldots, q. \tag{34}$$

Considering that $\sum_{j=1}^{q}\alpha - j = 0$, it follows that $\alpha_1 = 0$. Therefore, all the α_js are zero and the $\{\mathbf{v}_j\}$ are linearly independent, forming a base of $\mathbb{R}^q$.

Q.E.D

REFERENCES

AKAIKE, H. (1974): "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

AUESTAD, B., AND D. TJØSTHEIM (1990): "Identification of Nonlinear Time Series: First Order Characterization and Order Determination," *Biometrika*, 77, 669–687.

CARVALHO, A. X., AND M. A. TANNER (2002a): "Mixtures-of-Experts of Generalized Time Series: Consistency of the Maximum Likelihood Estimator," Technical report, University of British Columbia and Northwestern University.

——— (2002b): "Mixtures-of-Experts of Generalized Time Series: Asymptotic Normality and Model Specification," Technical report, University of British Columbia and Northwestern University.

CHEN, R. (1995): "Threshold Variable Selection in Open-Loop Threshold Autoregressive Models," *Journal of Time Series Analysis*, 16, 461–481.

CHEN, R., AND R. S. TSAY (1993): "Functional Coefficient AutoRegressive Models," *Journal of the American Statistical Association*, 88, 298–308.

CYBENKO, G. (1989): "Approximation by Superposition of Sigmoidal Functions," *Mathematics of Control, Signals, and Systems*, 2, 303–314.

FAN, J., AND Q. YAO (2003): *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag, New York, NY.

FINE, T. L. (1999): *Feedforward Neural Network Methodology*. Springer, New York.

FUNAHASHI, K. (1989): "On the Approximate Realization of Continuous Mappings by Neural Networks," *Neural Networks*, 2, 183–192.

GALLANT, A. R., AND H. WHITE (1992): "On Learning the Derivatives of an Unknown Mapping with Multilayer Feedforward Networks," *Neural Networks*, 5, 129–138.

GRANGER, C. W. J., AND T. TERÄSVIRTA (1993): *Modelling Nonlinear Economic Relationships*. Oxford University Press, Oxford.

HEILER, S. (1999): "A Survey on Nonparametric Time Series Analysis," Economics Working Papers at WUSTL 9904005, Washington University.

HORNIK, K., M. STINCHOMBE, AND H. WHITE (1989): "Multi-Layer Feedforward Networks are Universal Approximators," *Neural Networks*, 2, 359–366.

——— (1990): "Universal Approximation of an Unknown Mapping and its Derivatives Using Multi-Layer Feedforward Networks," *Neural Networks*, 3, 551–560.

HÄRDLE, W. (1990): *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.

HÄRDLE, W., H. LÜTKEPOHL, AND R. CHEN (1997): "A Review of Nonparametric Time Series Analysis," *International Statistical Review*, 65, 49–72.

HUBER, P. J. (1967): "The Behavior of Maximum Likelihood Estimates Under Non Standard Conditions," in *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*, pp. 221–223.

HUERTA, G., W. JIANG, AND M. TANNER (2001): "Mixtures of Time Series Models," *Journal of Computational and Graphical Statistics*, 10, 82–89.

——— (2003): "Time Series Modeling via Hierachical Mixtures," *Statistica Sinica*, 13, 1097–1118.

HWANG, J. T. G., AND A. A. DING (1997): "Prediction Intervals for Artificial Neural Networks," *Journal of the American Statistical Association*, 92, 109–125.

JACOBS, R. A. (1990): "Task Decomposition Through Computation in a Modular Connectionist Architecture," Ph.d. thesis, University of Massachusetts.

JACOBS, R. A., M. I. JORDAN, S. J. NOWLAN, AND G. E. HINTON (1991): "Adaptive Mixtures of Local Experts," *Neural Computation*, 3, 79–87.

JENNRICH, R. I. (1969): "Asymptotic Properties of Non-linear Least Squares Estimators," *The Annals of Mathematical Statistics*, 40, 633–643.

JORDAN, M. I., AND R. A. JACOBS (1994): "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, 6, 181–214.

KURKOVÁ, V., AND P. C. KAINEN (1994): "Functionally Equivalent Feedforward Neural Networks," *Neural Computation*, 6, 543–558.

LAI, T. L., AND S. P.-S. WONG (2001): "Stochastic Neural Networks with Applications to Nonlinear Time Series," *Journal of the American Statistical Association*, 96, 968–981.

LEYBOURNE, S., P. NEWBOLD, AND D. VOUGAS (1998): "Unit Roots and Smooth Transitions," *Journal of Time Series Analysis*, 19, 83–97.

MACKAY, D. J. C. (1992a): "Bayesian Interpolation," *Neural Computation*, 4, 415–447.

——— (1992b): "A practical Bayesian framework for Backpropagation Networks," *Neural Computation*, 4, 448–472.

MEDEIROS, M. C., T. TERÄSVIRTA, AND G. RECH (2002): "Building Neural Network Models for Time Series: A Statistical Approach," Working Paper Series in Economics and Finance 508, Stockholm School of Economics.

MEDEIROS, M. C., AND A. VEIGA (2000a): "A Hybrid Linear-Neural Model for Time Series Forecasting," *IEEE Transactions on Neural Networks*, 11, 1402–1412.

——— (2000b): "A Flexible Coefficient Smooth Transition Time Series Model," Working Paper Series in Economics and Finance 361, Stockholm School of Economics.

——— (2002): "Diagnostic Checking in a Flexible Nonlinear Time Series Model," *Journal of Time Series Analysis*, 24, 461–482.

NOWLAN, S. J. (1990): "Maximum Likelihood Competitive Learning," in *Advances in Neural Information Processing Systems*, vol. 2, pp. 574–582. Morgan Kaufmann.

OZAKI, T. (1982): "The Statistical Analysis of Perturbed Limit Cycle Process Using Nonlinear Time Series Models," *Journal of Time Series Analysis*, 3, 29–41.

PEDREIRA, C. E., L. C. PEDROZA, AND M. FARIÑAS (2001): "Local-Global Neural Networks for Interpolation," in *ICANNGA – Praga*, pp. 55–58.

PÖTSCHER, B. M., AND I. R. PRUCHA (1986): "A Class of Partially Adaptive One-step M-Estimators for the Non-linear Regression Model with Dependent Observations," *Journal of Econometrics*, 32, 219–251.

RECH, G., T. TERÄSVIRTA, AND R. TSCHERNIG (2001): "A Simple Variable Selection Technique for Nonlinear Models," *Communications in Statistics, Theory and Methods*, 30, 1227–1241.

SCHWARZ, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.

SUSSMAN, H. J. (1992): "Uniqueness of the Weights for Minimal Feedforward Nets with a Given Input-Output Map," *Neural Networks*, 5, 589–593.

SWANSON, N. R., AND H. WHITE (1995): "A Model Selection Approach to assesssing the information in the term structure using linear models and artificial neural networks," *Journal of Business and Economic Statistics*, 13, 265–275.

——— (1997a): "Forecasting Economic Time Series Using Flexible Versus Fixed Specification and Linear Versus Nonlinear Econometric Models," *International Journal of Forecasting*, 13, 439–461.

——— (1997b): "A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks," *Review of Economic and Statistics*, 79, 540–550.

TCHERNING, R., AND L. YANG (2000): "Nonparametric Lag Selection for Time Series," *Journal of Time Series Analysis*, 21, 457–487.

TERÄSVIRTA, T. (1994): "Specification, Estimation, and Evaluation of Smooth Transition Autoregressive Models," *Journal of the American Statistical Association*, 89(425), 208–218.

TJØSTHEIM, D., AND B. AUESTAD (1994): "Nonparametric Identification of Nonlinear Time Series – Selecting Significant Lags," *Journal of the American Statistical Association*, 89(428), 1410–1419.

TONG, H. (1990): *Non-linear Time Series: A Dynamical Systems Approach*, vol. 6 of *Oxford Statistical Science Series*. Oxford University Press, Oxford.

TRAPLETTI, A., F. LEISCH, AND K. HORNIK (2000): "Stationary and Integrated Autoregressive Neural Network Processes," *Neural Computation*, 12, 2427–2450.

TSAY, R. (1989): "Testing and Modeling Threshold Autoregressive Processes," *Journal of the American Statistical Association*, 84, 431–452.

VAN DIJK, D., T. TERÄSVIRTA, AND P. H. FRANSES (2002): "Smooth Transition Autoregressive Models - A Survey of Recent Developments," *Econometric Reviews*, 21, 1–47.

VIEU, P. (1995): "Order Choice in Nonlinear Autoregressive Models," *Statistics*, 26, 307–328.

WEIGEND, A., B. HUBERMAN, AND D. RUMELHART (1992): "Predicting Sunspots and Exchange Rates with Connectionist Networks," in *Nonlinear Modeling and Forecasting*, ed. by M. Casdagli, and S. Eubank. Addison-Wesley.

WEIGEND, A. S., M. MANGEAS, AND A. N. SRIVASTAVA (1995): "Nonlinear gated experts for time series: Discovering regimes and avoiding overfitting," *International Journal of Neural Systems*, 6, 373–399.

WHITE, H. (1981): "Consequences and Detection of Misspecified Nonlinear Regression Models," *Journal of the American Statistical Association*, 76, 419–433.

——— (1990): "Connectionist Nonparametric Regression: Multilayer Feedforward Networks Can Learn Arbitrary Mappings," *Neural Networks*, 3, 535–550.

——— (1994): *Estimation, Inference and Specification Analysis*. Cambridge University Press, New York, NY.

WHITE, H., AND I. DOMOWITZ (1984): "Nonlinear Regression with Dependent Observations," *Econometrica*, 52, 143–162.

WONG, C. S., AND W. K. LI (2000): "On a Mixture Autoregressive Model," *Journal of the Royal Statistical Society, Series B*, 62, 95–115.

——— (2001): "On a Mixture Autoregressive Conditional Heterocedastic Model," *Journal of the American Statistical Association*, 96, 982–995.

WOOLDRIDGE, J. M. (1994): "Estimation and Inference for Dependent Process," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. 4, pp. 2639–2738. Elsevier Science.

XIA, Y., AND W. K. LI (1999): "On Single-Index Coefficient Regression Models," *Journal of the American Statistical Association*, 94, 1275–1285.

YAO, Q., AND H. TONG (1994): "On Subset Selection in Non-Parametric Stochastic Regression," *Statistica Sinica*, 4, 51–70.

(M. S. Fariñas) CENTER FOR STUDIES IN PHYSICS AND BIOLOGY, THE ROCKEFELLER UNIVERSITY, NEW YORK, NY, USA.

E-mail address: mayte@babel.rockefeller.edu

(C. E. Pedreira) DEPARTMENT OF ELECTRICAL ENGINEERING, PONTIFICAL CATHOLIC UNIVERSITY OF RIO DE JANEIRO, RIO DE JANEIRO, RJ, BRAZIL.

E-mail address: pedreira@ele.puc-rio.br

(M. C. Medeiros – Corresponding author) DEPARTMENT OF ECONOMICS, PONTIFICAL CATHOLIC UNIVERSITY OF RIO DE JANEIRO, RIO DE JANEIRO, RJ, BRAZIL.

E-mail address: mcm@econ.puc-rio.br

Departamento de Economia PUC-Rio
Pontifícia Universidade Católica do Rio de Janeiro
Rua Marques de Sâo Vicente 225 - Rio de Janeiro 22453-900, RJ
Tel.(21) 31141078 Fax (21) 31141084
www.econ.puc-rio.br
flavia@econ.puc-rio.br