Local Global Neural Networks: A New Approach for Nonlinear Time Series Modeling

Mayte SUÁREZ-FARIÑAS, Carlos E. PEDREIRA, and Marcelo C. MEDEIROS

We propose the local-global neural networks model within the context of time series models. This formulation encompasses some already existing nonlinear models and also admits the mixture of experts approach. We emphasize 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 and nonstationary linear models and is able to describe "intermittent" dynamics; the system spends a large fraction of time in a bounded region, but sporadically 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 procedure is illustrated with two real time series.

KEY WORDS: Model building; Model identifiability; Neural network; Nonlinear model; Parameter estimation; Sunspot number; Time series.

1. INTRODUCTION

The past few years have witnessed a vast development of nonlinear time series techniques (Tong 1990; Granger and Teräsvirta 1993). Among these, 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; Fan and Yao 2003.) Another class of models, flexible functional forms, offers an alternative that also leave the functional form of the relationship partially unspecified. Although these models do contain parameters (often a large number of them) the parameters are not globally identified. Identification, if achieved, is local at best, with no restrictions imposed on the parameters. Usually the parameters are not interpretable, as they often are in parametric models.

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, and epidemiology. 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* model. The idea behind this model, proposed by Jacobs, Jordan, Nowlan, and Hinton (1991), is to "divide and conquer." 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 second, the ideas developed by Jacobs (1990) using a similar modular architecture but a different cost function. Jordan and Jacobs (1994) generalized the foregoing ideas by proposing the so-called "hierarchical mixture-of-experts" model. Both the mixture-of-experts and the hierarchical mixture-of-experts models 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) provided an application to financial time series forecasting. Good applications of hierarchical mixtures-of-experts models in time series have been given by Huerta, Jiang, and Tanner (2001, 2003). Carvalho and Tanner (2002a,b) proposed the mixture of generalized linear time series models and derived several asymptotic results. It is also worth mentioning the mixture autoregressive (AR) model proposed by Wong and Li (2000) and its generalization developed by Wong and Li (2001).

In this article we propose a new model, based on ANNs and partly inspired by the ideas from the mixture-of-experts literature, termed the *local global neural network* (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 constitutes a combination of several pairs, each composed of an approximation function and 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, modeling is approached by the specialization of neurons in each sector 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 subsets 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 emphasize the linear expert case. The model is then called the *linear local global neural network* (L^{2}GNN) model. Here we

Mayte Suárez-Fariñas is Postdoctoral Fellow, Center for Studies in Physics and Biology, The Rockefeller University, New York, NY (E-mail: *mayte@babel.rockefeller.edu*). Carlos E. Pedreira is Associate Professor, Department of Electrical Engineering (E-mail: *carlosp@centroin.com.br*), and Marcelo C. Medeiros is Assistant Professor, Department of Economics (E-mail: *mcm@econ.puc-rio.br*), Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil. This work is based on the first author's doctoral thesis at the Department of Electrical Engineering, Pontifical Catholic University, Rio de Janeiro. Financial support from the CNPq is gratefully acknowledged. The authors thank Marcelo O. Magnasco, Maurício Romero Sicre, Juan Pablo Torres-Martínez, and Alvaro Veiga for valuable discussions and two anonymous referees, an associate editor, and the editor for helpful comments.

© 2004 American Statistical Association
Journal of the American Statistical Association
December 2004, Vol. 99, No. 468, Theory and Methods
DOI 10.1198/016214504000001691

give geometric interpretation of the model and discuss the conditions under which the proposed model is asymptotically stationary. We show that the L^{2}GNN model consists of a mixture of stationary and nonstationary linear models that are able to describe "intermittent" dynamics; the system spends a large fraction of the time in a bounded region, but sporadically develops an instability that grows exponentially for some time and then suddenly collapses. Furthermore, based on the work of Trapletti, Leisch, and Hornik (2000), we extensively discuss the existence, consistency, and asymptotic normality of the parameter estimates. We also carefully consider conditions under which the L^{2}GNN model is identifiable. Identification is essential for consistency and asymptotic normality of the parameter estimates. We develop a model building strategy and estimate the parameters by concentrated maximum likelihood, which dramatically reduces the computational burden. The whole procedure is illustrated with two real time series. Similar proposals are the stochastic neural network (SNN) model developed by Lai and Wong (2001) and the neuro-coefficient smooth transition AR (NCSTAR) model of Medeiros and Veiga (2000a).

The article 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^{2}GNN model. Section 5 considers parameter estimation, and Section 6 presents a model building strategy. Section 7 gives examples with real time series, and Section 8 briefly summarizes our results. A technical Appendix provides the proofs of the main results.

2. MODEL FORMULATION

The LGNN model describes a stochastic process $y_t \in \mathbb{R}$ through the following nonlinear model:

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

where $\mathbf{x}_t \in \mathbb{R}^q$ represents a vector of lagged values of y_t and/or some exogenous variables and $\{\varepsilon_t\}$ is sequence of independently and identically distributed random variables with mean 0 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

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

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

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

where

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

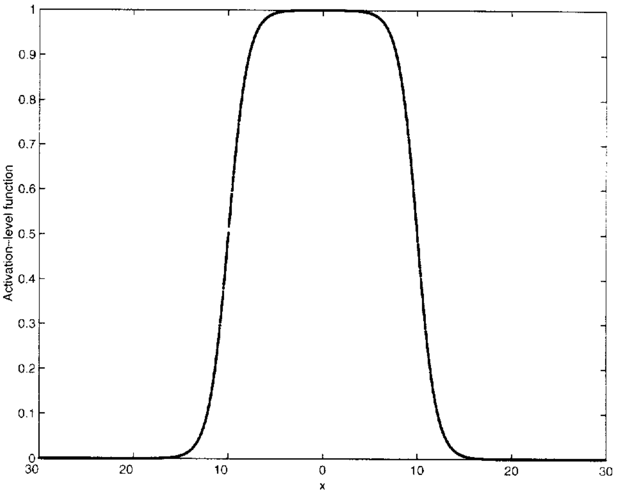

Figure 1. Example of an Activation-Level Function With $x_t \sim$ Unif(−30, 30), $\gamma = 1$, $\mathbf{d} = 1$, $\beta^{(1)} = -10$, and $\beta^{(2)} = 10$.

and $\langle \cdot, \cdot \rangle$ 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 (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 this article, the approximation functions are linear, that is, $L(\mathbf{x}_t; \boldsymbol{\psi}_{L_i}) = \mathbf{a}'_i \mathbf{x}_t + b_i$, with $\mathbf{a}_i = [a_{11}, a_{12}, \ldots, a_{1q}]' \in \mathbb{R}^q$ and $b_i \in \mathbb{R}$. In this case the model is called the L^{2}GNN model, where

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

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

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 by the diagram in Figure 2. Notice that the hidden layer is formed by m pairs of neurons. Each pair of neurons

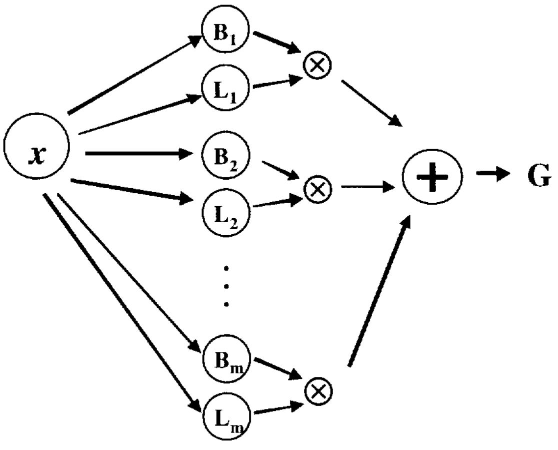

Figure 2. Neural Network Architecture.

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. We address this issue fully in Section 5.2.

As pointed out in Section 1, 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 although these models are closely related, they have some significant differences. The NCSTAR model can be written as

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

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

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

and ε_t is a Gaussian white noise. The SNN model starts from this same equation [see eq. (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 $F(\mathbf{x}_t; \mathbf{d}_i, \beta_i)$ [eqs. (9a) and (9b) in Lai and Wong 2001]. These differences have two main implications. First, in contrast to the NCSTAR and L^{2}GNN models, the SNN model is a stochastic linear map; because given the choice of I_{ti}, the map is linear, the nonlinearities appear not in the maps themselves, but rather 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 estimation algorithm. But although the introduction of the random variables I_{ti} looks minor, in fact it changes the asymptotics of the model in some 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, because, 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* the NCSTAR and SNN models, and is in our opinion 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, then the NCSTAR and SNN models will be asymptotically nonstationary with probability strictly greater than 0; particular (i.e., measure zero) choices of parameters must be made to guarantee asymptotic stationarity in this case. In contrast, the L^{2}GNN model will remain asymptotically stationary with probability 1 by imposing some very weak restrictions on the parameter $\mathbf{d}$ (see Theorem 1); particular choices of parameters must be made to permit the dynamics to diverge. It is thus interesting to note that although the NCSTAR and SNN models are in some sense "supersets" of the L^{2}GNN model, because each L^{2}GNN map can be written as two maps in (5), an important property that 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 nonstationary linear models. Asymptotic stationarity of the 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, whereas the activation functions of the L^{2}GNN model are "small," because they cover a small fraction of any sufficiently large sphere. Thus if the NCSTAR or SNN models are nonstationary, then the dynamics can easily escape to infinity; if an L^{2}GNN model is nonstationary, then 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 the 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 1, 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 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 that small perturbations of the parameters do not change the behavior; the NCSTAR and SNN models can by definition fit that dynamic also, 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 $\mathbb{X}$,

$$\mathbb{H}_1 = \{\mathbf{x}_t \in \mathbb{R}^q \,|\, \langle \mathbf{d}, \mathbf{x}_t \rangle = \beta^{(1)}\}$$

and

$$\mathbb{H}_2 = \{\mathbf{x}_t \in \mathbb{R}^q \,|\, \langle \mathbf{d}, \mathbf{x}_t \rangle = \beta^{(2)}\}. \qquad (6)$$

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. Because 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 they divide $\mathbb{X}$ into three different regions, $\mathbb{H}^-$, $\mathbb{H}^0$, and $\mathbb{H}^+$, defined as

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

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 nonactive state of the neuron pair is represented by activation-level function $B(\mathbf{x}_t; \boldsymbol{\psi}_B)$. 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 (i.e., if $\exists i, j,\ i \neq j$, such that $\mathbf{d}_i \neq \mathbf{d}_j$), then the region formed by the interception of hyperplanes, $\mathbb{H}^0_{ij} = \mathbb{H}^0_i \cap \mathbb{H}^0_j$, is a nonempty region and represents the region where the neuron pairs i and j are both active.

One case worth special mention is when the hyperplanes are parallel to one another, 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 $\beta_i^{(2)} < \beta_{i+1}^{(1)}\ \forall i$, 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 of 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 this is also true for the L^{2}GNN model. One possibility, because 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 by Chen and Tsay (1993) and applied by 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 AR processes, Theorem 1 gives less-restrictive sufficient conditions for the asymptotic stationarity of the L^{2}GNN model. It is easy to verify 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:

a. 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.
b. 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 $d_{ij} \neq 0$, $i = 1, \ldots, m$, $j = 1, \ldots, p$.
c. 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 1, 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

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

The proof of this theorem, given in the Appendix, is based on the results for linear AR models. The intuition behind the foregoing result is that when y_t grows in absolute value, the functions $B(\mathbf{x}_t; \boldsymbol{\psi}_{B_i}) \to 0$, $i = 1, \ldots, m$, and thus y_t is driven back to 0. Condition a is trivial and implies that all of the limiting AR models are asymptotically stationary. Condition b considers the case where there are explosive regimes. Finally, condition c 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 AR parameters.

The following examples demonstrate 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 AR 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. Consider 1,000 observations of the following L^{2}GNN model:

$$\begin{aligned} y_t &= (-.5 - 1.5y_{t-1}) \\ &\quad \times \left[\frac{1}{1 + \exp(10(y_{t-1} + 6))} - \frac{1}{1 + \exp(10(y_{t-1} - 1))}\right] \\ &\quad + (-.5 - 1.2y_{t-1}) \\ &\quad \times \left[\frac{1}{1 + \exp(10(y_{t-1} + 2))} - \frac{1}{1 + \exp(10(y_{t-1} - 2))}\right] \\ &\quad + \varepsilon_t, \end{aligned} \qquad (9)$$

where $\varepsilon_t \sim \text{NID}(0, 1)$. Figure 3 shows the generated time series, the activation-level functions, the autocorrelogram of series, and the histogram of the data. Model (9) is a mixture of two explosive AR processes. When either only one of the activation-level functions is active or both of them equal 1, the AR 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: .0052, 1.0140, and 2.6567.

Example 2. Consider 3,000 observations of the following L^{2}GNN model:

$$\begin{aligned} y_t &= (-.5 - 2.2y_{t-1} + 2.5y_{t-2}) \\ &\quad \times \left[\frac{1}{1 + \exp(.7y_{t-1} - .7y_{t-2} + 10)} \right. \\ &\quad \left. - \frac{1}{1 + \exp(.7y_{t-1} - .7y_{t-2} - 10)}\right] \\ &\quad + (.5 - 1.9y_{t-1} - 1.2y_{t-2}) \\ &\quad \times \left[\frac{1}{1 + \exp(1.5(.7y_{t-1} - .7y_{t-2} + 2))} \right. \\ &\quad \left. - \frac{1}{1 + \exp(1.5(.7y_{t-1} - .7y_{t-2} - 40))}\right] \\ &\quad + \varepsilon_t, \end{aligned} \qquad (10)$$

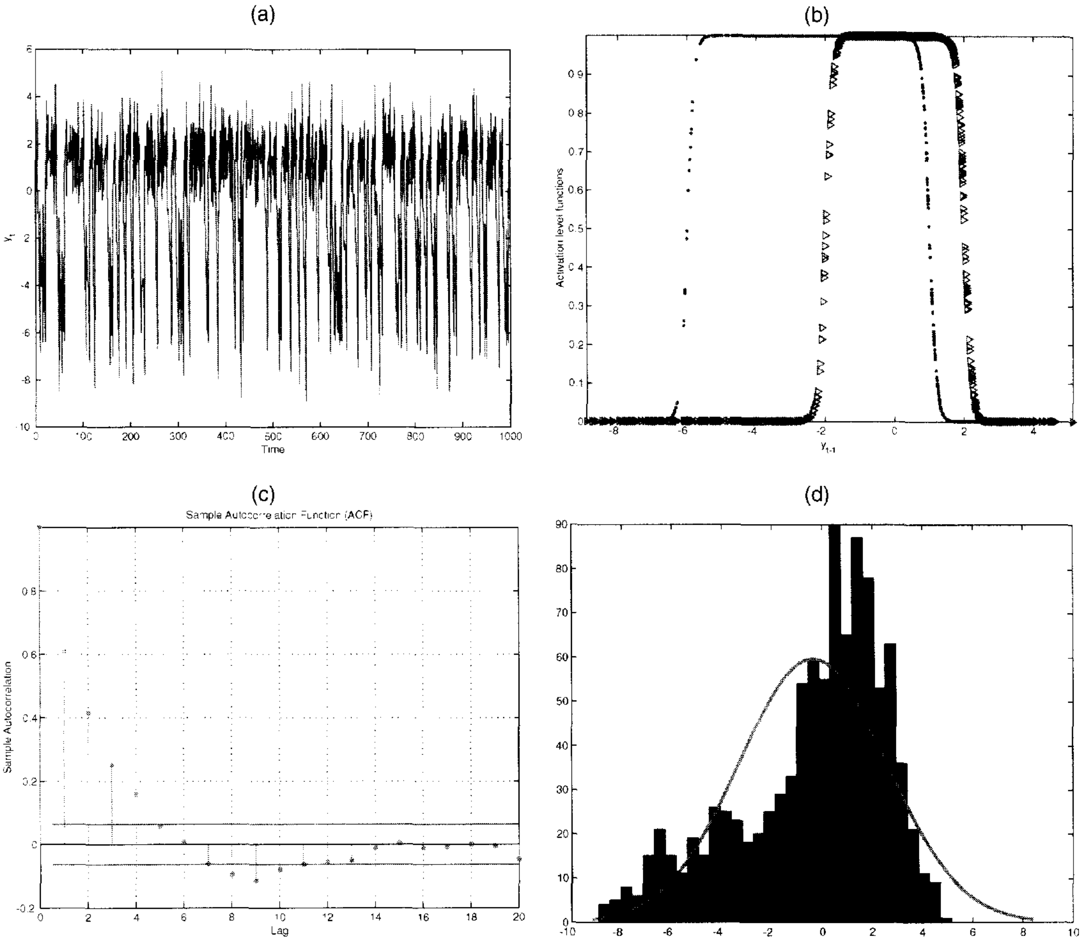

Figure 3. Example 1. (a) Generated time series. (b) Scatterplot of the activation-level functions against y_{t-1}. (c) Autocorrelogram of the series. (d) Histogram of the series.

where $\varepsilon_t \sim \mathrm{NID}(0, 1)$. 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. Consider 3,000 observations of the following L^{2}GNN model:

$$\begin{aligned}
y_t &= (-.5 - 2.2y_{t-1} + 2.5y_{t-2}) \\
&\quad \times \left[\frac{1}{1+\exp(.7y_{t-1} - .7y_{t-2} + 10)}\right. \\
&\quad \left. - \frac{1}{1+\exp(.7y_{t-1} - .7y_{t-2} - 10)}\right] \\
&\quad + (.5 - 1.9y_{t-1} - 1.2y_{t-2}) \\
&\quad \times \left[\frac{1}{1+\exp(1.5(y_{t-1} + \delta y_{t-2} + 2))}\right. \\
&\quad \left. - \frac{1}{1+\exp(1.5(y_{t-1} + \delta y_{t-2} - 40))}\right] \\
&\quad + \varepsilon_t, \qquad (11)
\end{aligned}$$

where $\varepsilon_t \sim \mathrm{NID}(0, 1)$ 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. Consider 3,000 observations of the following L^{2}GNN model:

$$\begin{aligned}
y_t &= (.5 + 2y_{t-1} - y_{t-2}) \\
&\quad \times \left[\frac{1}{1+\exp(.7y_{t-1} - .7\delta y_{t-2} + 10)}\right. \\
&\quad \left. - \frac{1}{1+\exp(.7y_{t-1} - .7\delta y_{t-2} - 10)}\right] \\
&\quad + (.5 - .5y_{t-1} + .5y_{t-2}) \\
&\quad \times \left[\frac{1}{1+\exp(.7y_{t-1} - .7\delta y_{t-2} - 5)}\right. \\
&\quad \left. - \frac{1}{1+\exp(.7y_{t-1} - .7\delta y_{t-2} - 15)}\right] \\
&\quad + \varepsilon_t, \qquad (12)
\end{aligned}$$

where $\varepsilon_t \sim \mathrm{NID}(0, 1)$ and $\delta = -1, 1$. It can be seen that model (12) has three limiting AR regimes. The associated transition matrixes [see eq. (8)] are

$$\mathbf{A}^{(1)} = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix}, \qquad \mathbf{A}^{(2)} = \begin{bmatrix} 1.5 & -.5 \\ 1 & 0 \end{bmatrix},$$

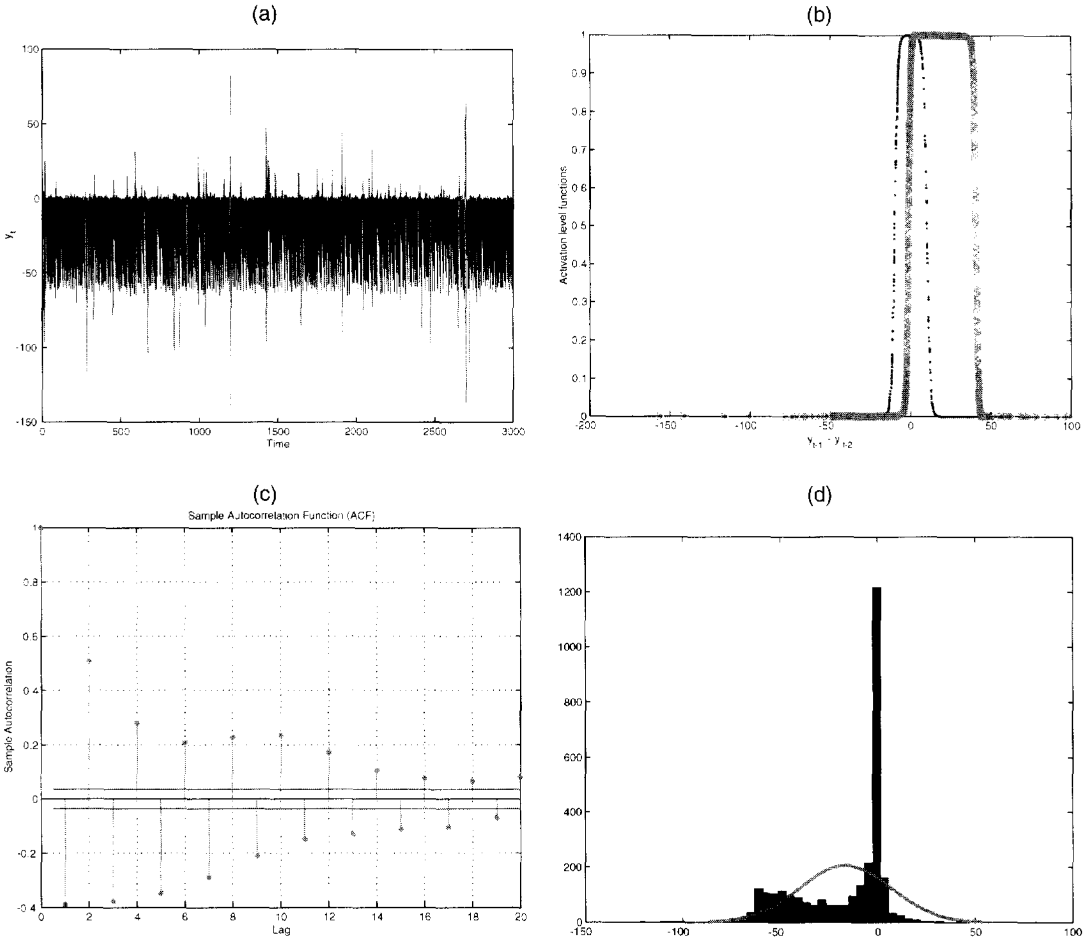

Figure 4. Example 2. (a) Generated time series. (b) Scatterplot of the activation-level functions against $y_{t-1} - y_{t-2}$. (c) Autocorrelogram of the series. (d) Histogram of the series.

and

$$\mathbf{A}^{(3)} = \begin{bmatrix} -.5 & .5 \\ 1 & 0 \end{bmatrix},$$

with the respective eigenvalue pairs (1, 1), (1, .5), and (−1, .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

Numerous algorithms for estimating the parameters of models based on neural networks are available in the literature. In this article we estimate the parameters of our L^{2}GNN model by maximum likelihood, making use of the assumptions made for 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 will most likely lead to convergence problems, and that penalizing the log-likelihood function in one way or another is a necessary precondition for satisfactory results. Two things can be said in favor 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 must be chosen carefully, as we discuss in detail later in this section.

The L^{2}GNN 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^{2}GNN model defined by (4) is estimated as

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

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^{2}GNN model is identifiable.

5.1 Existence of the Estimator

The proof of existence is based on lemma 2 of Jennrich (1969), which establishes that the LSE exists under certain con-

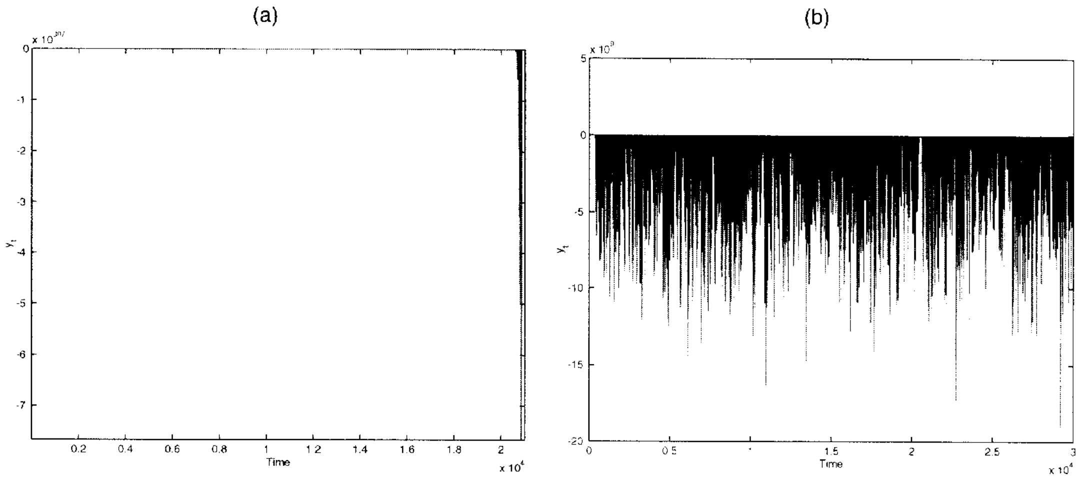

Figure 5. Example 3, Generated Time Series. (a) $\delta = 0$; (b) $\delta = 10^{-10}$.

ditions of continuity and measurability on the mean squared error (MSE) function. Theorem 2 states the necessary conditions for existence of the LSE.

Theorem 2. The L^{2}GNN model satisfies the following conditions, and the LSE exists:

a. For each $\mathbf{x}_t \in \mathbb{X}$, function $G_{\mathbf{x}}(\boldsymbol{\psi}) = G(\mathbf{x}_t; \boldsymbol{\psi})$ is continuous in a compact subset Ψ of the Euclidean space.
b. For each $\boldsymbol{\psi} \in \Psi$, function $G_{\boldsymbol{\psi}}(\mathbb{X}) = G(\mathbf{x}_t; \boldsymbol{\psi})$ is measurable in space $\mathbb{X}$.
c. ε_t are independent and identically distributed errors with mean 0 and variance σ^2.

Remark 2. To extend the set of approximation functions beyond linear functions, we need to verify conditions a and b of Theorem 2. Thus the class of functions $L(\mathbf{x}_t; \boldsymbol{\psi}_{L_i})$, $i = 1, \ldots, m$, 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. Huber (1967) presented results that require only 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 MSE function, the sources of uniqueness of the model must be identified. These questions have been studied by Sussman (1992), Kurková and Kainen (1994), Hwang and Ding (1997), Trapletti et al. (2000), and Medeiros, Teräsvirta, and Rech (2002) in the case of a feedforward neural network model. Here we briefly discuss the main concepts and results. In particular, we establish and prove the conditions guaranteeing that the proposed model is identifiable and minimal. Before tackling the problem of the identifiability of the model, we discuss two related concepts: the concept of minimality of the model, established by Sussman (1992) and "nonredundancy" termed by Hwang and Ding (1997), and the concept of model reducibility.

Definition 1. The L^{2}GNN model is minimal (or nonredundant), if its input–output map cannot be obtained from another model with fewer neuron pairs.

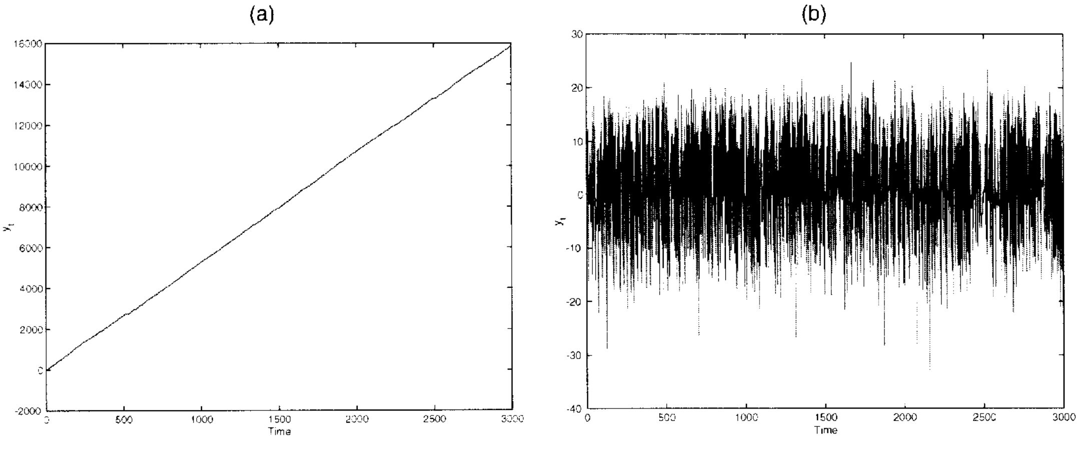

Figure 6. Example 4, Generated Time Series. (a) $\delta = 1$; (b) $\delta = -1$.

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

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

a. One of the pairs $(\mathbf{a}_i, b_i)$ vanishes jointly for some $i = 1, \dots, m$.
b. $\gamma_i = 0$ for some $i = 1, \dots, m$.
c. There is at least one pair (i, j), $i \neq j$, $i = 1, \dots, m$, $j = 1, \dots, m$, such that $\varphi(\mathbf{x}_t; \boldsymbol{\theta}_{i\ell})$ and $\varphi(\mathbf{x}_t; \boldsymbol{\theta}_{j\ell})$ are sign-equivalent. That is, $|\varphi(\mathbf{x}_t; \boldsymbol{\theta}_{i\ell})| = |\varphi(\mathbf{x}_t; \boldsymbol{\theta}_{j\ell})|$ $\forall \mathbf{x}_t \in \mathbb{R}^q$, $t = 1, \dots, T$.

Definition 3. The L^{2}GNN 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^{2}GNN 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 log-likelihood function.

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

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

establishes another indetermination in the model, because 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 the condition

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

or

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

(P.4) The presence of irrelevant hidden neuron pairs. Conditions a and b 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 $b_i = 0$, then 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 log-likelihood function. Furthermore, if $\beta_i^{(1)} = \beta_i^{(2)}$, then γ_i, $\mathbf{a}_i$, and b_i remain unidentified.

Properties (P.2) and (P.3) are related to the concept of reducibility. In the same spirit of the results given by Sussman (1992) and Hwang and Ding (1997), we show that if the model is irreducible, then property (P.1) is the only way to modify the parameters without affecting the distribution of y. Hence, by establishing restrictions on the parameters of (4) that simultaneously avoid reducibility and any permutation of hidden units, we guarantee 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 $\beta_i^{(2)} < \beta_{i+1}^{(2)}$, $i = 1, \dots, m$.

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

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

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

(R.3) $\gamma_i > 0$, $i = 1, \dots, m$.
(R.4) $d_{i1} > 0$, $i = 1, \dots, m$.

The first of these prevents the possibility of a simple change of sign in parameter γ leading to problems in model identification. As discussed previously, 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 this problem.

Because $\mathbf{d}_i$ is a unit vector, we have

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

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.

Corollary 2.1 of Sussman (1992) and corollary 2.4 of 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^{2}GNN model is globally identifiable, we state the following assumptions.

Assumption 1. The parameters $\mathbf{a}_i$ and b_i do not vanish jointly for some $i = 1, \dots, m$. 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 that 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), and Assumption 2 avoids problems related to multicollinearity.

Theorem 3. Under the following restrictions:

(R.1) $\beta_i^{(1)} < \beta_{i+1}^{(1)}$ and $\beta_i^{(2)} < \beta_{i+1}^{(2)}, i = 1, \dots, m$,
(R.2) $\beta_i^{(1)} < \beta_i^{(2)}, i = 1, \dots, m$,
(R.3) $\gamma_i > 0, i = 1, \dots, m$,
(R.4) $d_{i1} = \sqrt{1 - \sum_{j=2}^{q} d_{ij}^2} > 0, i = 1, \dots, m$,

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

5.3 Strong Consistency of Estimators

White (1981) and White and Domowitz (1984) established the conditions that guarantee strong consistency of the LSE. In the context of stationary time series models, the conditions that ensure (almost certain) consistency have been established by 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 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

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

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

Theorem 5. Under restrictions (R.1)–(R.4) and Assumptions 1–5,

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

where $\nabla^2 \overline{Q}_T(\boldsymbol{\psi}^*) = E[\nabla^2 Q_T(\boldsymbol{\psi}^*)]$, $\nabla^2 Q_n(\boldsymbol{\psi}^*)$ 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

To reduce the computational burden, we can apply the concentrated maximum likelihood method to estimate $\boldsymbol{\psi}$ as follows. Consider the ith 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 $\mathbf{y}' = [y_1, y_2, \dots, y_T]$, $\boldsymbol{\varepsilon}' = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T]$, and

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

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 ith iteration. This form of concentrated maximum likelihood, proposed by Leybourne, Newbold, and Vougas (1998), considerably reduces the dimensionality of the iterative estimation problem.

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^{2}GNN models. Assume now that we have estimated an L^{2}GNN 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 the following steps:

1. For $k = 1, \dots, K$:
   a. Construct a vector $\mathbf{v}_m^{(k)} = [v_{1m}^{(k)}, \dots, v_{qm}^{(k)}]'$ such that $v_{1m}^{(k)} \in (0, 1]$ and $v_{jm}^{(k)} \in [-1, 1]$, $j = 2, \dots, q$. The values for $v_{1m}^{(k)}$ are drawn from a uniform $(0, 1]$ distribution, and the values for $v_{jm}^{(k)}$, $j = 2, \dots, q$, are drawn from a uniform $[-1, 1]$ distribution.
   b. Define $\mathbf{d}_m^{(k)} = \mathbf{v}_m^{(k)} \|\mathbf{v}_m^{(k)}\|^{-1}$.
   c. Compute the projections $\mathbf{p}_m^{(k)} = \langle \mathbf{d}_m^{(k)}, \mathbf{x}\rangle$, where $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_T]$.
   d. Let $c_{1m}^{(k)} = Z_{1/3}(\mathbf{p}_m^{(k)})$ and $c_{2m}^{(k)} = Z_{2/3}(\mathbf{p}_m^{(k)})$, where Z_α is the α-percentile of the empirical distribution of $\mathbf{p}_m^{(k)}$.
2. Define a grid of N positive values $\gamma_m^{(n)}$, $n = 1, \dots, N$, for the slope parameter and estimate $\boldsymbol{\psi}_L$ using (15).
3. For $k = 1, \dots, K$ and $n = 1, \dots, N$, 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, to ensure that the identifying restrictions are satisfied. A similar procedure was proposed by Medeiros and Veiga (2000b) and Medeiros et al. (2002).

Typically, $K = 1{,}000$ 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, then we have to select a large K as well.

6. MODEL BUILDING

In this section we develop a specific-to-general specification strategy. From (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 selecting 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. Evaluating the estimated model.

We discuss the first two steps of the modeling cycle in detail. The evaluation step is beyond the scope of this article; however, the results of Medeiros and Veiga (2002) and Medeiros et al. (2002) can be easily generalized to the case of L^{2}GNN models.

6.1 Variable Selection

The first step in our model specification process 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, particularly when the number of observations is not small. Here were carry out variable selection by linearizing the model and applying well-known techniques of linear variable selection to this approximation. This keeps the computational cost to a minimum. For this purpose, we adopt the simple procedure proposed by Rech, Teräsvirta, and Tschernig (2001) 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 kth-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 model is approximated by another function. This yields

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

where $\tilde{\mathbf{x}}_t = [1, \mathbf{x}_t']'$ and $R(\mathbf{x}_t; \boldsymbol{\psi})$ is the approximation error that can be made negligible by choosing k sufficiently high. The θ'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 (16), every product of variables involving at least one redundant variable has the coefficient 0. The idea is to sort out the redundant variables by using this property of (16). To do this, we first regress y_t on all variables on the right side of (16), assuming that $R(\mathbf{x}_t; \boldsymbol{\psi}) = 0$, and compute the value of a model selection criterion (MSC), such as the Akaike information criterion (AIC) (Akaike 1974) or the Bayes information criterion (BIC) (Schwarz 1978). After doing this, we remove one variable from the original model and regress y_t on all of 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 this way until the polynomial is of a function of a single regressor and, finally, just a constant. Having done this, 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. Note that by following this procedure, the variables for the whole L^{2}GNN model are selected at the same time. Rech et al. (2001) showed that this procedure works well already in small samples when compared with 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 neurons 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 (see, e.g., Fine 1999, pp. 215–221). Bayesian regularization may serve as an example (MacKay 1992a,b).

Another possibility is to use a MSC to determine the number of hidden neuron pairs. Swanson and White (1995, 1997a,b) applied the BIC model selection criterion as follows. They started with a linear model, adding potential variables to it until the BIC indicated that the model cannot be further improved. Then they estimated models with a single hidden neuron and selected regressors sequentially to it one by one unless the BIC showed no further improvement. Next, they added another hidden unit, and proceeded by adding variables to it. The selection process is terminated when BIC indicates that no more hidden units or variables should be added or when a predetermined maximum number of hidden units has been reached. This modeling strategy can be termed fully sequential.

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

$$\mathrm{BIC}(h) = \ln(\hat{\sigma}^2) + \frac{\ln(T)}{T} \times [2m(2+q)], \qquad (17)$$

where $\hat{\sigma}^2$ 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 et al. (2002) and use a sequence of Lagrange multiplier tests. However, this approach is beyond the scope of this article.

7. NUMERICAL EXAMPLES

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

7.1 The Canadian Lynx Series

The first dataset analyzed is the classic 10-based logarithm of the number of Canadian lynx trapped in the Mackenzie River district of Northwest Canada over the period 1821–1934. (For further details and a background history, see Tong 1990, chap. 7.) Previous analyses of this series have been given by Ozaki (1982), Tsay (1989), Teräsvirta (1994), and Xia and Li (1999). We start by selecting the variables of the model among the first seven lags of the time series. With the procedure described in Section 6.1 and using the BIC, we identified lags 1 and 2; using the AIC, we identified lags 1, 2, 3, 5, 6, and 7. We continue building an L^{2}GNN model with only lags 1 and 2, which is more parsimonious. The final estimated mode has two neuron pairs ($m = 2$), and when compared with a linear AR(2) model, the ratio between the standard deviation of the residuals from the nonlinear model and linear one is $\frac{\widehat{\sigma}}{\widehat{\sigma}_L} = .876$.

The estimated residual standard deviation ($\widehat{\sigma} = .204$) is smaller than that 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 .222, and the exponential AR (EXPAR) model proposed by Ozaki (1982) has $\hat{\sigma}_\varepsilon = .208$.

7.2 The Sunspot Series

In this example we consider the annual sunspot numbers over the period 1700–1998. We used the observations for the period 1700–1979 to estimate the model and used the remaining observations to forecast evaluation. We adopted the same transformation as used by Tong (1990), $y_t = 2[\sqrt{(1+N_t)} - 1]$, where N_t is the sunspot number. The series was obtained from the National Geophysical Data Center web page (*http://www.ngdc.noaa.gov/stp/SOLAR/SSN/ssn.html*). The sunspot numbers are a heavily modeled nonlinear time series; a neural network example was provided by Weigend, Huberman, and Rumelhart (1992).

We begin L^{2}GNN modeling 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 the BIC, lags 1, 2, and 7 are selected, whereas the AIC yields lags 1, 2, 4, 5, 6, 7, 8, 9, and 10. As in the previous example, we proceed with the lags selected by the BIC; however, the residuals of the estimated model are strongly autocorrelated. We remove the serial correlation 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 BIC, as described in Section 6.2.

After estimating a model with three neuron pairs, we continue considering the out-of-sample performance of the estimated model. 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 by Tong (1990, p. 420) and the other reported by Chen (1995); an ANN model with 10 hidden neurons and the first 9 lags as input variables, estimated with Bayesian regularization (MacKay 1992a,b); the SNN model estimated in Lai and Wong (2001); the NCSTAR model of Medeiros and Veiga (2000a), and a linear AR model with lags selected using the BIC. The threshold variable is a nonlinear function of lagged values of the time series in the SETAR model estimated by Chen (1995), 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 AR with multivariate transition variables in the same spirit as 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 the table, the L^{2}GNN model has the smallest RMSE and MAE among the alternatives considered herein. Over 19 forecasts, the L^{2}GNN 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 article we have proposed a new nonlinear time series model based on neural networks. This model, the LGNN, can be interpreted as a mixture-of-experts model. We analyzed the case of linear experts in detail and discussed its probabilistic and statistical properties. The proposed model consists of a mixture of stationary and nonstationary linear models and is able to describe "intermittent" dynamics; the system spends a large fraction of the time in a bounded region, but sporadically 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 BIC, has been suggested to determine the variables and the number of hidden neuron pairs. When put to the test in a real experiment involving one-step-ahead forecasting, the proposed model outperforms the linear model and other nonlinear specifications considered in this article, suggesting that the theory developed here is useful. Thus the proposed model appears to be a useful tool for practicing time series analysts.

APPENDIX: PROOFS

A.1 Lemmas

Lemma A.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, then the class of functions $\{B(x; \boldsymbol{\psi}_B)\} \cup \{x B(x; \boldsymbol{\psi}_B)\}$, where

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

is linearly independent.

Lemma A.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 the following:

a. $\langle \mathbf{d}_i, \mathbf{v} \rangle > 0$.
b. 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 these same conditions.

Table 1. One-Step-Ahead Forecasts, Their RMSEs, and MAEs 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	.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	.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	.1	32.9	−3.0	27.0	2.9	35.7	−5.8	29.9	−.1	27.7	2.2	25.0	4.9
1995	17.5	14.0	3.5	19.2	−1.7	18.4	−.9	18.9	−1.4	17.6	−.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	.3
1997	21.5	17.2	4.3	21.3	.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	.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

A.2 Proofs of Theorems

A.2.1 Proof of Theorem 1. Write model (4) as

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

where

$$\mathbf{Y}_t = \begin{pmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{pmatrix}, \qquad \mathbf{Y}_{t-1} = \begin{pmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \end{pmatrix},$$

$$\mathbf{e}_t = \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \mathbf{a}_{t-1} = \begin{pmatrix} \sum_{i=1}^m b_i B(\mathbf{Y}_{t-1}) \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

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

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

After recursive substitutions, model (A.1) can be written as

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

Model (A.2) 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 a in Theorem 1 is satisfied. As

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

$\prod_t \mathbf{A}_t \to \mathbf{0}$ if $B_i(\mathbf{Y}_t) \to 0$, $i = 1, \ldots, m$. This will be true if $|\langle \mathbf{d}_i, \mathbf{Y}_t\rangle| \to M$, where $M \gg \max(\beta_i^{(1)}, \beta_i^{(2)})$. 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 b in Theorem 1). When a given limiting AR regime has unit roots, 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 c in Theorem 1).

A.2.2 Proof of Theorem 2. Lemma 2 of Jennrich (1969) shows that conditions a–c in Theorem 2 are sufficient to guarantee the existence (and measurability) of the LSE. To apply this result to the $\mathrm{L^2GNN}$ model, we need to check whether these conditions are satisfied by the model.

Condition c of Theorem 2 was already assumed when defining the model. It is easy to prove in our case that $G(\mathbf{x}_t; \boldsymbol{\psi})$ is continuous in the parameter vector $\boldsymbol{\psi}$. This follows from the fact that $B(\mathbf{x}_t; \boldsymbol{\psi}_{B_i})$ 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 thus is measurable, for each fixed value of the parameter vector $\boldsymbol{\psi}$. Thus conditions a and b are satisfied.

A.2.3 Proof of Theorem 3. Suppose that $\widetilde{\boldsymbol{\psi}} = [\widetilde{\boldsymbol{\psi}}'_L, \widetilde{\boldsymbol{\psi}}'_B]'$ is another vector of parameters such that

$$\sum_{i=1}^m (\mathbf{a}'_i \mathbf{x}_t + b_i) B(\mathbf{x}_t; \boldsymbol{\psi}_{B_i}) = \sum_{i=1}^m (\widetilde{\mathbf{a}}'_i \mathbf{x}_t + \widetilde{b}_i) B(\mathbf{x}_t; \widetilde{\boldsymbol{\psi}}_{B_i}). \tag{A.3}$$

To show global identifiability of the $\mathrm{L^2GNN}$ model, we need to prove that, under Assumption 1 and restrictions (R.1)–(R.4), (A.3) is satisfied if and only if $\mathbf{a}_i = \widetilde{\mathbf{a}}_i$, $b_i = \widetilde{b}_i$, and $\boldsymbol{\psi}_B = \widetilde{\boldsymbol{\psi}}_B$, $i = 1, \ldots, m$, $\forall \mathbf{x}_t \in \mathbb{R}^q$.

Equation (A.3) can be rewritten as

$$\sum_{j=1}^{2m}(\mathbf{c}'_j\mathbf{x}_t + e_j)B(\mathbf{x}_t, \breve{\boldsymbol{\psi}}_{B_j}) = 0, \qquad (A.4)$$

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

To relate this problem to Lemma A.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. Because the set $\{\mathbf{d}_1, \dots, \mathbf{d}_m\}$ has a finite number of points, $\gamma_i > 0$ [restriction (R.3)], and $d_{i1} > 0$ [restriction (R.4)], $i = 1, \dots, m$, 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 (A.4) by $x_t\mathbf{v}$, $x_t \in \mathbb{R}$, leads to

$$\sum_{j=1}^{2m}(\overline{c}_j x_t + e_j)B(x_t\mathbf{v}, \breve{\boldsymbol{\psi}}_{B_j}) = 0, \qquad (A.5)$$

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

For simplicity of notation, let $\varphi_j^{(\ell)} = \varphi(\mathbf{x}_t; \boldsymbol{\theta}_{j\ell})$, $j = 1, \dots, 2m$. Lemma A.1 implies that if $\varphi_{j_1}^{(\ell)}$ and $\varphi_{j2}^{(\ell)}$ are not sign-equivalent, then $j_1 \in \{1. \dots, 2m\}$, $j_2 \in \{1, \dots, 2m\}$, and (A.5) holds if and only if $\overline{c}_j$ and e_j vanish jointly for every $j \in \{1, \dots, 2m\}$. However, the condition $\overline{c}_j$, $j = 1, \dots, 2m$, does not imply that $\mathbf{c}_j = \mathbf{0}$. Lemma A.2 in fact shows that vector $\mathbf{v}$ is not unique and that there exist vectors $\mathbf{v}_1, \dots, \mathbf{v}_q$ that satisfy the same conditions as $\mathbf{v}$ and form a basis for $\mathbb{R}^q$. 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 the possibility that two functions $\varphi_{j_1}^{(\ell)}$ and $\varphi_{j_2}^{(\ell)}$ coming from the same model are sign-equivalent. Consequently, $\exists\, j_1 \in \{1, \dots, m\}$ and $j_2 \in \{m+1, \dots, 2m\}$ 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 (A.3) holds is when $\mathbf{a}_i = \widetilde{\mathbf{a}}_i$, $b_i = \widetilde{b}_i$, and $\boldsymbol{\psi}_B = \widetilde{\boldsymbol{\psi}}_B$, $i = 1, \dots, m$, $\forall\,\mathbf{x}_t \in \mathbb{R}^q$.

A.2.4 Proof of Theorem 4. For the proof of this theorem, we draw on 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, $t = 1, \dots, T$. Under the conditions of Theorem 3 and the fact that ε_t is a mean 0, normally distributed random variable with finite variance, and hence k-integrable, Assumption 3.1a follows.

Assumption 3.1b states that $-\mathrm{E}(q(\mathbf{x}_t; \boldsymbol{\psi}))$ is continuous in Ψ, $t = 1, \dots, T$. Let $\boldsymbol{\psi} \to \boldsymbol{\psi}^*$, because 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 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 $\boldsymbol{\psi}^*$. Theorem 3 demonstrates that under Assumption 1 and with the restrictions (R.1)–(R.4) imposed, the L^{2}GNN is globally identifiable.

A.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.

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

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

and

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

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

Assumption 3.8b: Under Assumption 4, the fact that the function $G(\mathbf{x}_t; \boldsymbol{\psi})$ 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 $A_n^* \equiv \mathrm{E}(\nabla^2 Q(\psi^*)) = 2\mathrm{E}(\nabla G(\mathbf{x}_t; \boldsymbol{\psi}^*) \times \nabla' G(\mathbf{x}_t; \boldsymbol{\psi}*))$ is $O(1)$ in our setup. Assumption 5, the properties of function $G(\mathbf{x}_t; \boldsymbol{\psi})$, and the unique identification of $\boldsymbol{\psi}$ imply the nonsingularity of $\mathrm{E}(\nabla G(\mathbf{x}_t; \boldsymbol{\psi}^*)\nabla' G(\mathbf{x}_t; \boldsymbol{\psi}^*))$.

Assumption 6.1: Using theorem 2.4 of White and Domowitz (1984) we can show that the sequence $2\boldsymbol{\xi}'\nabla G(\mathbf{x}_t; \boldsymbol{\psi}^*)\varepsilon_t$ obeys the central limit theorem (CLT) for some $(r \times 1)$ vector $\boldsymbol{\xi}$, such that $\boldsymbol{\xi}'\boldsymbol{\xi} = 1$. Assumptions A(i) and A(iii) of White and Domowitz (1984) hold because ε_t is NID. Assumption A(ii) holds with $V = 4\sigma^2\boldsymbol{\xi}'\mathrm{E}(\nabla G(\mathbf{x}_t; \boldsymbol{\psi}^*)\nabla' G(\mathbf{x}_t; \boldsymbol{\psi}^*))$. Furthermore, because any measurable transformation of mixing processes is itself mixing (see lemma 2.1 in White and Domowitz 1984), $2\boldsymbol{\xi}'\nabla G(\mathbf{x}_t; \boldsymbol{\psi}^*)\varepsilon_t$ is a strong mixing sequence and obeys the CLT. Using the Cramér–Wold device, $\nabla Q(\mathbf{x}_t; \boldsymbol{\psi})$ also obeys the CLT with covariance matrix $B_n^* = 4\sigma^2\mathrm{E}(\nabla G(\mathbf{x}_t; \boldsymbol{\psi}^*)\nabla' G(\mathbf{x}_t; \boldsymbol{\psi}^*)) = 2\sigma^2 A_n^*$, which is $O(1)$ and nonsingular.

A.3 Proofs of Lemmas

A.3.1 Proof of Lemma A.1. To make the proof clearer, let $\varphi_i^{(\ell)}(x) = (h_i x - \gamma_i\beta_i^{(\ell)})$, where $h_i = \gamma_i\langle\mathbf{d}_i, \mathbf{v}\rangle$, and write $B(x; \boldsymbol{\psi}_{B_i})$ as $B(\varphi_i^{(1)}(x), \varphi_i^{(2)}(x))$. 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 $\beta_i^{(1)} < \beta_i^{(2)}$, $i = 1, \dots, n$, with $(h_i, \gamma_i, \beta_i^{(1)}, \beta_i^{(2)}) \neq (h_j, \gamma_j, \beta_j^{(1)}, \beta_j^{(2)})$ for $i \neq j$ (due to their not being equivalent in sign) such that $\forall x \in \mathbb{R}$, we have

$$\sum_{i=1}^{n}(\lambda_i + \omega_i x)B(\varphi^{(1)}(x), \varphi^{(2)}(x)) = 0, \qquad (A.6)$$

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

Considering that $B(\varphi_i^{(1)}(x), \varphi_i^{(2)}(x)) = F(-\varphi_i^{(1)}(x)) - F(-\varphi_i^{(2)}(x))$, where $F(\cdot)$ is the logistic function, (A.6) is equivalent to

$$\sum_{i=1}^{n} (\lambda_i + \omega_i x)\left[F\left(-\varphi_i^{(1)}(x)\right) - F\left(-\varphi_i^{(2)}(x)\right)\right] = 0. \qquad (A.7)$$

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

$$F\left(-\varphi_i^{\ell}(x)\right) = \sum_{k=1}^{\infty} (-1)^k e^{-k\gamma_i \beta_i^{(\ell)}} e^{kh_i x}. \qquad (A.8)$$

The series converges absolutely when $e^{-k\gamma_i\beta_i^{(\ell)}} < 1$, that is, for $x < (\frac{\gamma_i \beta_i^{(\ell)}}{h_i})$. Therefore, there exist M small enough such that (A.8) converges for every $x \in (-\infty, M)$. Substituting (A.8) in (A.7) and writing $C_i^{(\ell)} = \gamma_i \beta_i^{(\ell)}$, we obtain

$$\sum_{i=1}^{n} \left\{ (\lambda_i + \omega_i x) \sum_{k=1}^{\infty} (-1)^k \left[e^{-C_i^{(1)}} - e^{-C_i^{(2)} k}\right] e^{kh_i x} \right\} = 0. \qquad (A.9)$$

Notice that because γ_i is positive, $C_i^{(1)} < C_i^{(2)}$. Denoting $W_i^{(\ell)} = -e^{-C_i^{(\ell)}}$, $\ell = 1, 2$, we have that $W_i^{(1)} < W_i^{(2)}$ and, substituting in (A.9),

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

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

$$\sum_{k=1}^{\infty} \alpha_k^* e^{h_k^* x} + \alpha_k^{**} x e^{h_k^* x} = 0, \qquad (A.10)$$

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 (A.10) by $xe^{h_1^* x}$, we obtain

$$\sum_{k=1}^{\infty} \left\{ \alpha_k^* e^{x(h_k^* - h_1^*)} + \alpha_k^{**} \frac{e^{x(h_k^* - h_1^*)}}{x} \right\} = 0, \qquad (A.11)$$

and, assuming the limit in (A.11) 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 (A.10) with $\alpha_1^* = 0$ and dividing by $e^{h_1^*}$, we obtain

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

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

There remains 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 present the proof only 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 noninteger number smaller than 1. So we must prove that there is a sequence K_n such that for all $i \geq 2$, $K_n\theta_i$ is not an integer. Let $J_Z = \{j \in J \mid \exists r$ 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 noninteger, so $K_n\theta_i$ cannot be an integer number. Otherwise, if $i \notin J_Z$, then there are no integer numbers such that $K_n\theta_i$ would be an integer. Because K_n is an integer number, $K_n\theta_i$ 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,

$$\sum_{j \in J} \lambda_j \left(W_j^{(1)}\right)^s = \sum_{j \in J} \lambda_j \left(W_j^{(2)}\right)^s. \qquad (A.12)$$

If $j \in J$, then $h_j = h_{i_0}$, and due to the definition of the h_i's, this can happen only if $\forall j \in J$, $d_j = d_{i_0}$. It then follows that $d_j = d_{i_0}$ and $\gamma_j = \gamma_{i_0}$. Considering that $(h_i, \gamma_i, \beta_i^{(1)}, \beta_i^{(2)}) \neq (h_j, \gamma_j, \beta_j^{(1)}, \beta_j^{(2)})$, it follows that $\beta_i^{(1)} \neq \beta_j^{(1)}$ and $\beta_i^{(2)} \neq \beta_j^{(2)}$. We then have 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 let $\phi : \{1, \ldots, n_J\} \to J$ represent 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 (A.12) by $W_{\phi(n_J)}^{(2)}$ and passing, to the limit as $k \to \infty$, we have

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

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 $\omega_i = 0$, $i = 1, \ldots, m$.

A.3.2 Proof of Lemma A.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; then define the $\mathbf{v}_j$'s as

$$\begin{aligned} &\mathbf{v}_1 = \mathbf{v}_0, \qquad \mathbf{v}_2 = \mathbf{v}_0 - \delta_2 \mathbf{e}_2, \\ &\mathbf{v}_3 = \mathbf{v}_0 - \delta_3 \mathbf{e}_3, \ldots, \qquad \mathbf{v}_q = \mathbf{v}_0 - \delta_q \mathbf{e}_q, \end{aligned} \qquad (A.13)$$

where $\mathbf{e}_j$ is the canonical vector with 1 in position j and 0 otherwise and δ_j is small enough. We should prove that these vectors satisfy the conditions of Lemma A.2 and (2) also 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 0,

$$\begin{aligned} \sum_{j=1}^{q} \alpha_j \mathbf{v}_j = 0 \quad &\Longrightarrow \quad \alpha_1 \mathbf{v}_0 + \sum_{j=2}^{q} \alpha_j (\mathbf{v}_0 - \delta_j \mathbf{e}_j) = 0 \\ &\Longrightarrow \quad \mathbf{v}_0 \sum_{j=1}^{q} \alpha_j - \sum_{j=2}^{q} \alpha_j \delta_j \mathbf{e}_j = 0. \end{aligned} \qquad (A.14)$$

From this, it follows that

$$\mathbf{v}_0 \sum_{j=1}^{q} \alpha_j = \sum_{j=2}^{q} \alpha_j \delta_j \mathbf{e}_j. \qquad (A.15)$$

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

$$\left(\mathbf{v}_0 \sum_{j=1}^{q} \alpha_j\right)_1 = \left(\sum_{j=2}^{q} \alpha_j \delta_j \mathbf{e}_j\right)_1 = 0, \qquad \text{(A.16)}$$

because $v_{01} \sum_{j=1}^{q} \alpha_j = 0$ and $v_{01} \neq 0$. Writing (A.16) for the component k, $k = 2, 3, \ldots, q$, we have that

$$0 = \left(\sum_{j=2}^{q} \alpha_j \delta_j \mathbf{e}_j\right)_k = \alpha_k + \delta_k$$
$$\Longrightarrow \quad \alpha_k = 0, \qquad k = 2, \ldots, q. \qquad \text{(A.17)}$$

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

[Received December 2002. Revised March 2004.]

REFERENCES

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

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

Carvalho, A. X., and Tanner, M. A. (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 Tsay, R. S. (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 Yao, Q. (2003), *Nonlinear Time Series: Nonparametric and Parametric Methods*, New York: Springer-Verlag.

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

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

Gallant, A. R., and White, H. (1992), "On Learning the Derivatives of an Unknown Mapping With Multilayer Feedforward Networks," *Neural Networks*, 5, 129–138.

Granger, C. W. J., and Teräsvirta, T. (1993), *Modelling Nonlinear Economic Relationships*, Oxford, U.K.: Oxford University Press.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge, U.K.: Cambridge University Press.

Härdle, W., Lütkepohl, H., and Chen, R. (1997), "A Review of Nonparametric Time Series Analysis," *International Statistical Review*, 65, 49–72.

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

Hornik, K., Stinchombe, M., and White, H. (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.

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*, Berkeley, CA: California University Press, pp. 221–223.

Huerta, G., Jiang, W., and Tanner, M. (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 Ding, A. A. (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," unpublished doctoral thesis, University of Massachusetts.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (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 Jacobs, R. A. (1994), "Hierarchical Mixtures of Experts and the EM Algorithm," *Neural Computation*, 6, 181–214.

Kurková, V., and Kainen, P. C. (1994), "Functionally Equivalent Feedforward Neural Networks," *Neural Computation*, 6, 543–558.

Lai, T. L., and Wong, S. P.-S. (2001), "Stochastic Neural Networks With Applications to Nonlinear Time Series," *Journal of the American Statistical Association*, 96, 968–981.

Leybourne, S., Newbold, P., and Vougas, D. (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., Teräsvirta, T., and Rech, G. (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 Veiga, A. (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," *IEEE Transactions on Neural Networks*, forthcoming.

——— (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, New York: Morgan Kaufmann, pp. 574–582.

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., Pedroza, L. C., and Fariñas, M. (2001), "Local-Global Neural Networks for Interpolation," in *Proceedings of the 5th International Conference on Artificial Neural Networks and Genetic Algorithms, Prague, April 2001*, pp. 55–58.

Pötscher, B. M., and Prucha, I. R. (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., Teräsvirta, T., and Tschernig, R. (2001), "A Simple Variable Selection Technique for Nonlinear Models," *Communications in Statistics, Part A—Theory and Methods*, 30, 1227–1241.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The 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 White, H. (1995), "A Model Selection Approach to Assesssing the Information in the Term Structure Using Linear Models and Artificial Neural Networks," *Journal of Business & 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 Yang, L. (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, 208–218.

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

Tong, H. (1990), *Non-Linear Time Series: A Dynamical Systems Approach*, Oxford, U.K.: Oxford University Press.

Trapletti, A., Leisch, F., and Hornik, K. (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., Teräsvirta, T., and Franses, P. H. (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., Huberman, B., and Rumelhart, D. (1992), "Predicting Sunspots and Exchange Rates With Connectionist Networks," in *Nonlinear Modeling and Forecasting*, eds. M. Casdagli and S. Eubank, Reading, MA: Addison-Wesley, pp. 395–432.

Weigend, A. S., Mangeas, M., and Srivastava, A. N. (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*, New York: Cambridge University Press.

White, H., and Domowitz, I. (1984), "Nonlinear Regression With Dependent Observations," *Econometrica*, 52, 143–162.

Wong, C. S., and Li, W. K. (2000), "On a Mixture Autoregressive Model," *Journal of the Royal Statistical Society*, Ser. 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 4*, eds. R. F. Engle and D. L. McFadden, Amsterdam: Elsevier Science, pp. 2639–2738.

Xia, Y., and Li, W. K. (1999), "On Single-Index Coefficient Regression Models," *Journal of the American Statistical Association*, 94, 1275–1285.

Yao, Q., and Tong, H. (1994), "On Subset Selection in Non-Parametric Stochastic Regression," *Statistica Sinica*, 4, 51–70.