Medeiros, Marcelo C.; Terasvirta, Timo; Rech, Gianluigi

Working Paper

Building Neural Network Models for Time Series: A Statistical Approach

Texto para discussão, No. 461

Provided in Cooperation with:
Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro

Suggested Citation: Medeiros, Marcelo C.; Terasvirta, Timo; Rech, Gianluigi (2002) : Building Neural Network Models for Time Series: A Statistical Approach, Texto para discussão, No. 461, Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Departamento de Economia, Rio de Janeiro

This Version is available at:
https://hdl.handle.net/10419/175948

TEXTO PARA DISCUSSÃO

No. 461

Building neural networks models for time series: a statistical approach

Marcelo C. Medeiros
Timo Teräsvirta
Gianluigi Rech

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

DEPARTAMENTO DE ECONOMIA
PUC-RIO

TEXTO PARA DISCUSSÃO
Nº. 461

BUILDING NEURAL NETWORKS MODELS FOR TIMES SERIES: A STATISTICAL APPROACH

MARCELO C. MEDEIROS
TIMO TERÄSVIRTA
GIANLUIGI RECH

AGOSTO 2002

Building Neural Network Models for Time Series: A Statistical Approach *

Marcelo C. Medeiros
Department of Economics, Catholic University of Rio de Janeiro

Timo Teräsvirta
Department of Economic Statistics, Stockholm School of Economics

Gianluigi Rech
Quantitative Analysis, Electrabel, Louvain-la-Neuve, Belgium

August 21, 2002

Abstract

This paper is concerned with modelling time series by single hidden layer feedforward neural network models. A coherent modelling strategy based on statistical inference is presented. Variable selection is carried out using existing techniques. The problem of selecting the number of hidden units is solved by sequentially applying Lagrange multiplier type tests, with the aim of avoiding the estimation of unidentified models. Misspecification tests are derived for evaluating an estimated neural network model. A small-sample simulation experiment is carried out to show how the proposed modelling strategy works and how the misspecification tests behave in small samples. Two applications to real time series, one univariate and the other multivariate, are considered as well. Sets of one-step-ahead forecasts are constructed and forecast accuracy is compared with that of other nonlinear models applied to the same series.

Keywords: Model misspecification, neural computing, nonlinear forecasting, nonlinear time series, smooth transition autoregression, sunspot series, threshold autoregression, financial prediction.

JEL Classification Codes: C22, C51, C52, C61, G12

Acknowledgments: This research has been supported by the Tore Browaldh's Foundation. The research of the first author has been partially supported by CNPq. The paper is partly based on Chapter 2 of the PhD thesis of the third author. A part of the work was carried out during the visits of the first author to the Department of Economic Statistics, Stockholm School of Economics and the second author to the Department of Economics, PUC-Rio. The hospitality of these departments is gratefully acknowledged. Material from this paper has been presented at the Fifth Brazilian Conference on Neural Networks, Rio de Janeiro, April 2001, the 20$^{\text{th}}$ International Symposium on Forecasting, Dublin, June 2002, and seminars at CORE (Louvain-la-Neuve), Monash University (Clayton, VIC), Swedish School of Economics (Helsinki), University of California, San Diego, Cornell University (Ithaca, NY), Federal Univeristy of Rio de Janeiro, and Catholic University of Rio de Janeiro. We wish to thank the participants of these occasions, Hal White in particular, for helpful comments. Our thanks also go to Chris Chatfield and Dick van Dijk for useful remarks, and Allan Timmermann for the data used in the second empirical example of the paper. The responsibility for any errors or shortcomings in the paper remains ours.

*Address for correspondence: Timo Teräsvirta, Department of Economic Statistics, Stockholm School of Economics, BOX 6501, SE-113 83, Stockholm, Sweden. E-mail: timo.terasvirta@hhs.se.

1 Introduction

Alternatives to linear models in econometric and time series modelling have increased in popularity in recent years. Nonparametric models that do not make assumptions about the parametric form of the functional relationship between the variables to be modelled have become more easily applicable due to computational advances and increased computational power. Another class of models, the flexible functional forms, offers an alternative that in fact also leaves the functional form of the relationship unspecified. While these models do contain parameters, often a large number of them, the parameters are not globally identified or, using the statistical terminology, estimable. Identification or estimability, if achieved, is local at best without additional parameter restrictions. The parameters are not interpretable either 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. Kuan and White (1994) surveyed the use of ANN models in (macro)economics, and several financial applications appeared in a recent special issue of IEEE Transactions on Neural Networks (Abu-Mostafa, Atiya, Magdon-Ismail and White 2001). The use of the ANN model in applied work is generally motivated by a 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; see, for example, Funahashi (1989), Cybenko (1989), Hornik, Stinchombe, and White (1989,1990), White (1990), or Gallant and White (1992). Such an approximator would still contain a finite number of parameters. How to specify such a model, that is, how to find the right combination of parameters and variables, is a central topic in the ANN literature and has been considered in a large number of books such as Bishop (1995), Ripley (1996), Fine (1999), Haykin (1999), or Reed and Marks II (1999), and articles. Many popular specification techniques are "general-to-specific" or "top-down" procedures: the investigator begins with a large model and applies appropriate algorithms to reduce the number of parameters using a predetermined stopping-rule. Such algorithms usually do not rely on statistical inference.

In this paper, we propose a coherent modelling strategy for simple single hidden-layer feedforward ANN time series models. These models discussed here are univariate, but adding exogenous

regressors to them does not pose problems. The difference between our strategy and the general-to-specific approaches is that ours works in the opposite direction, from specific to general. We begin with a small model and expand that according to a set of predetermined rules. The reason for this is that we view our ANN model as a statistical nonlinear model and apply statistical inference to the problem of specifying the model or, as the ANN experts express it, finding the network architecture. We shall argue in the paper that proper statistical inference is not available if we choose to proceed from large models to smaller ones, from general to specific. Our "bottom-up" strategy builds partly on early work by Teräsvirta and Lin (1993). More recently, Anders and Korn (1999) presented a strategy that shares certain features with our procedure. Swanson and White (1995,1997a,1997b) also developed and applied a specific-to-general strategy that deserves mention here. Balkin and Ord (2000) proposed an inference-based method for selecting the number of hidden units in the ANN model. Zapranis and Refenes (1999) developed a computer intensive strategy based on statistics to select the variables and the number of hidden units of the ANN model. Our aim has been to develop a strategy that minimizes the amount of computation required to reach the final specification and, furthermore, contains an in-sample evaluation of the estimated model. We shall consider the differences between our strategy and the other ones mentioned here in later sections of the paper.

The plan of the paper is as follows. Section 2 describes the model and Section 3 discusses geometric and statistical interpretations for it. A model specification strategy, consisting of specification, estimation, and evaluation of the model is described in Section 4. The results concerning a Monte-Carlo experiment are reported in Section 5 and two applications with real data sets are presented in Section 6. Section 7 contains concluding remarks.

2 The Autoregressive Neural Network Model

The AutoRegressive Neural Network (AR-NN) model is defined as

$$y_t = G(\mathbf{x}_t;\boldsymbol{\psi}) + \varepsilon_t = \boldsymbol{\alpha}'\tilde{\mathbf{x}}_t + \sum_{i=1}^{h} \lambda_i F(\tilde{\boldsymbol{\omega}}_i'\mathbf{x}_t - \beta_i) + \varepsilon_t \tag{1}$$

where $G(\mathbf{x}_t;\boldsymbol{\psi})$ is a nonlinear function of the variables $\mathbf{x}_t$ with parameter vector $\boldsymbol{\psi} \in \mathbb{R}^{(q+2)h+q+1}$ defined as $\boldsymbol{\psi} = [\boldsymbol{\alpha}', \lambda_1, \ldots, \lambda_h, \tilde{\boldsymbol{\omega}}_1', \ldots, \tilde{\boldsymbol{\omega}}_h', \beta_1, \ldots, \beta_h]'$. The vector $\tilde{\mathbf{x}}_t \in \mathbb{R}^{q+1}$ is defined as $\tilde{\mathbf{x}}_t = [1, \mathbf{x}_t']'$, where $\mathbf{x}_t \in \mathbb{R}^q$ is a vector of lagged values of y_t and/or some exogenous variables. The function $F(\tilde{\boldsymbol{\omega}}_i'\mathbf{x}_t - \beta_i)$, often called the activation function, is the logistic function

$$F(\tilde{\boldsymbol{\omega}}_i'\mathbf{x}_t - \beta_i) = \left(1 + e^{-(\tilde{\boldsymbol{\omega}}_i'\mathbf{x}_t - \beta_i)}\right)^{-1} \tag{2}$$

where $\tilde{\boldsymbol{\omega}}_i = [\tilde{\omega}_{1i}, \ldots, \tilde{\omega}_{qi}]' \in \mathbb{R}^q$ and $\beta_i \in \mathbb{R}$, and the linear combination of these functions in (1) forms the so-called hidden layer. Model (1) with (2) does not contain lags of ε_t and is therefore called a feedforward NN model. For other choices of the activation function, see Chen, Racine and Swanson (2001). Furthermore, $\{\varepsilon_t\}$ is a sequence of independently normally distributed random variables with zero mean and variance σ^2. The nonlinear function $F(\tilde{\boldsymbol{\omega}}_i'\mathbf{x}_t - \beta_i)$ is usually called a hidden neuron or a hidden unit. The normality assumption enables us to define the log-likelihood function, which is required for the statistical inference we need, but it can be relaxed.

3 Geometric and Statistical Interpretation

The characterization and tasks of a layer of hidden neurons in an AR-NN model are discussed in several textbooks such as Bishop (1995), Haykin (1999), Fine (1999), and Reed and Marks II (1999). It is nevertheless important to review some concepts in order to compare the AR-NN model with other well-known nonlinear time series models.

Consider the output $F(\tilde{\boldsymbol{\omega}}_i'\mathbf{x}_t - \beta_i)$ of a unit of the hidden layer of a neural network as defined in (1) and (2). When $\tilde{\boldsymbol{\omega}}_i'\mathbf{x}_t = \beta_i$, the parameters $\tilde{\boldsymbol{\omega}}_i$ and β_i define a hyperplane in a q-dimensional Euclidean space

$$\mathbb{H} = \{\mathbf{x}_t \in \mathbb{R}^q \,|\, \tilde{\boldsymbol{\omega}}_i'\mathbf{x}_t = \beta_i\}. \tag{3}$$

The direction of $\tilde{\boldsymbol{\omega}}_i$ determines the orientation of the hyperplane and the scalar term $\beta_i/\|\tilde{\boldsymbol{\omega}}_i\|$ the position of the hyperplane in terms of its distance from the origin. A hyperplane induces a partition of the space $\mathbb{R}^q$ into two regions defined by the halfspaces

$$\mathbb{H}^+ = \{\mathbf{x}_t \in \mathbb{R}^q \,|\, \tilde{\boldsymbol{\omega}}_i'\mathbf{x}_t \geq \beta_i\} \tag{4}$$

and

$$\mathbb{H}^{-} = \{\mathbf{x}_t \in \mathbb{R}^q | \tilde{\boldsymbol{\omega}}_i' \mathbf{x}_t < \beta_i\}, \tag{5}$$

associated to the states of the neuron. With h hyperplanes, a q-dimensional space will be split into several polyhedral regions. Each region is defined by the nonempty intersection of the halfspaces (4) and (5) of each hyperplane. Hyperplanes lying parallel to each other constitute a special case. In this situation, $\tilde{\boldsymbol{\omega}}_i \equiv \tilde{\boldsymbol{\omega}}$, $i = 1, \ldots, h$, and equation (1) becomes

$$y_t = \boldsymbol{\alpha}' \tilde{\mathbf{x}}_t + \sum_{i=1}^{h} \lambda_i F(\tilde{\boldsymbol{\omega}}' \mathbf{x}_t - \beta_i) + \varepsilon_t. \tag{6}$$

The input space is thus split in $h + 1$ regions.

Certain special cases of (1) are of interest. When $\mathbf{x}_t = y_{t-d}$ in F, model (1) becomes a multiple logistic smooth transition autoregressive (MLSTAR) model with $h + 1$ regimes in which only the intercept changes according to the regime. The resulting model is expressed as

$$y_t = \boldsymbol{\alpha}' \tilde{\mathbf{x}}_t + \sum_{i=1}^{h} \lambda_i F(\gamma_i (y_{t-d} - c_i)) + \varepsilon_t, \tag{7}$$

where $\gamma_i = \tilde{\omega}_{i1}$ and $c_i = \beta_i / \tilde{\omega}_{i1}$. When $h = 1$, equation (7) defines a special case of an ordinary LSTAR model considered in Teräsvirta (1994). When $\gamma_i \to \infty$, $i = 1, \ldots, h$, model (7) becomes a Self-Exciting Threshold AutoRegressive (SETAR) model with a switching intercept and $h + 1$ regimes. In another important special case $\mathbf{x}_t = t$ in F. In this situation the intercept of a linear model changes smoothly as a function of time. A linear model with h structural shifts in the intercept is obtained as $\gamma_i \to \infty$, $i = 1, \ldots, h$.

An AR-NN model can thus be either interpreted as a semi-parametric approximation to any Borel-measurable function or as an extension of the MLSTAR model where the transition variable can be a linear combination of stochastic variables. We should however stress the fact that model (1) is, in principle, neither globally nor locally identified. Three characteristics of the model imply non-identifiability. The first one is the exchangeability property of the AR-NN model. The value in the likelihood function of the model remains unchanged if we permute the hidden units. This results in $h!$ different models that are indistinguishable from each other and in $h!$ equal local maxima

of the log-likelihood function. The second characteristic is that in (2), $F(x) = 1 - F(-x)$. This yields two observationally equivalent parametrizations for each hidden unit. Finally, the presence of irrelevant hidden units is a problem. If model (1) has hidden units such that $\lambda_i = 0$ for at least one i, the parameters $\tilde{\boldsymbol{\omega}}_i$ and β_i remain unidentified. Conversely, if $\tilde{\boldsymbol{\omega}}_i = \mathbf{0}$ then λ_i and β_i can take any value without the value of the likelihood function being affected.

The first problem is solved by imposing, say, the restrictions $\beta_1 \leq \cdots \leq \beta_h$ or $\lambda_1 \geq \cdots \geq \lambda_h$. The second source of underidentification can be circumvented, for example, by imposing the restrictions $\tilde{\omega}_{1i} > 0$, $i = 1, \ldots, h$. To remedy the third problem, it is necessary to ensure that the model contains no irrelevant hidden units. This difficulty is dealt with by applying statistical inference in model specification; see Section 4. For further discussion of the identifiability of ANN models see, for example, Sussman (1992), Kurková and Kainen (1994), Hwang and Ding (1997), and Anders and Korn (1999).

For estimation purposes it is often useful to reparametrize the logistic function (2) as

$$F\left(\gamma_i\left(\boldsymbol{\omega}_i'\mathbf{x}_t - c_i\right)\right) = \left(1 + e^{-\gamma_i(\boldsymbol{\omega}_i'\mathbf{x}_t - c_i)}\right)^{-1} \tag{8}$$

where $\gamma_i > 0$, $i = 1, \ldots, h$, and $\|\boldsymbol{\omega}_i\| = 1$ with

$$\omega_{i1} = \sqrt{1 - \sum_{j=2}^{q} \omega_{ij}^2} > 0, i = 1, \ldots, h. \tag{9}$$

The parameter vector $\boldsymbol{\psi}$ of model (1) becomes

$$\boldsymbol{\psi} = [\boldsymbol{\alpha}', \lambda_1, \ldots, \lambda_h, \gamma_1, \ldots, \gamma_h, \omega_{12}, \ldots, \omega_{1q}, \ldots, \omega_{h2}, \ldots, \omega_{hq}, c_1, \ldots, c_h]'.$$

In this case the first two identifying restrictions discussed above can be defined as, first, $c_1 \leq \cdots \leq c_h$ or $\lambda_1 \geq \cdots \geq \lambda_h$ and, second, $\gamma_i > 0, i = 1, \ldots, h.$. We shall return to this parameterization in Section 4.3, when maximum likelihood estimation of the parameters is considered.

4 Strategy for Building AR-NN Models

4.1 Three Stages of Model Building

As mentioned in the Introduction, our aim is to construct a coherent strategy for building AR-NN models using statistical inference. The structure or architecture of an AR-NN model has to be determined from the data. We call this stage *specification* of the model, and it involves two sets of decision problems. First, the lags or variables to be included in the model have to be selected. Second, the number of hidden units has to be determined. Choosing the correct number of hidden units is particularly important as selecting too many neurons yields an unidentified model. In this work, the lag structure or the variables included in the model are determined using well-known variable selection techniques. The specification stage of NN modelling also requires *estimation* because we suggest choosing the hidden units sequentially. After estimating a model with h hidden units we shall test it against the one with $h+1$ hidden units and continue until the first acceptance of a null hypothesis. What follows thereafter is *evaluation* of the final estimated model to check if the final model is adequate. NN models are typically only evaluated out-of-sample, but in this paper we suggest in-sample misspecification tests for the purpose. Similar tests are routinely applied in evaluating STAR models (see Eitrheim and Teräsvirta (1996)), and in this work we adapt them to the AR-NN models. All this requires consistency and asymptotic normality for the estimators of parameters of the AR-NN model, conditions for which, based on results in Trapletti, Leisch and Hornik (2000), will be stated below.

We shall begin the discussion of our modelling strategy by considering variable selection. After dealing with that problem we turn to parameter estimation. Finally, after discussing statistical inference for selecting the hidden units and after presenting our in-sample model evaluation tools we outline the modelling strategy as a whole.

4.2 Variable Selection

The first step in our model specification is to choose the variables for the model from a set of potential variables (lags in the pure AR-NN case). Several nonparametric variable selection techniques exist (Tschernig and Yang 2000, Vieu 1995, Tjøstheim and Auestad 1994, Yao and Tong

1994, Auestad and Tjøstheim 1990), but they are computationally very demanding, in particular when the number of observations is not small. In this paper variable selection is carried out by linearizing the model and applying well-known techniques of linear variable selection to this approximation. This keeps computational cost to a minimum. For this purpose we adopt the simple procedure proposed in Rech, Teräsvirta and Tschernig (2001). Their idea is to approximate the stationary nonlinear model by a polynomial of sufficiently high order. Adapted to the present situation, the first step is to approximate function $G(\mathbf{x}_t;\boldsymbol{\psi})$ in (1) by a general k-th order polynomial. By the Stone-Weierstrass theorem, the approximation can be made arbitrarily accurate if some mild conditions, such as the parameter space $\boldsymbol{\psi}$ being compact, are imposed on function $G(\mathbf{x}_t;\boldsymbol{\psi})$. Thus the AR-NN model, itself a universal approximator, is approximated by another function. This yields

$$\begin{aligned} G(\mathbf{x}_t;\boldsymbol{\psi}) = \boldsymbol{\pi}'\tilde{\mathbf{x}}_t + \sum_{j_1=1}^{q}\sum_{j_2=j_1}^{q}\theta_{j_1 j_2}x_{j_1,t}x_{j_2,t} \\ + \cdots + \sum_{j_1=1}^{q}\cdots\sum_{j_k=j_{k-1}}^{q}\theta_{j_1\ldots j_k}x_{j_1,t}\cdots x_{j_k,t} + R(\mathbf{x}_t;\boldsymbol{\psi}), \end{aligned} \tag{10}$$

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

In equation (10), every product of variables involving at least one redundant variable has the coefficient zero. The idea is to sort out the redundant variables by using this property of (10). In order to do that, we first regress y_t on all variables on the right-hand side of equation (10) assuming $R(\mathbf{x}_t;\boldsymbol{\psi})$ and compute the value of a model selection criterion (MSC), AIC or SBIC for example. After doing that, we remove one variable from the original model and regress y_t on all the remaining terms in the corresponding polynomial and again compute the value of the MSC. This procedure is repeated by omitting each variable in turn. We continue by simultaneously omitting two regressors of the original model and proceed in that way until the polynomial is of a function of a single regressor and, finally, just a constant. Having done that, we choose the combination of variables that yields the lowest value of the MSC. This amounts to estimating $\sum_{i=1}^{q}\binom{q}{i}+1$ linear models

by ordinary least squares (OLS). Note that by following this procedure, the variables for the whole ANN model are selected at the same time. Rech et al. (2001) showed that the procedure works well already in small samples when compared to well-known nonparametric techniques. Furthermore, it can be successfully applied even in large samples when nonparametric model selection becomes computationally infeasible.

4.3 Parameter Estimation

As selecting the number of hidden units requires estimation of neural network models, we now turn to this problem. A large number of algorithms for estimating the parameters of a NN model are available in the literature. In this paper we instead estimate the parameters of our AR-NN model by maximum likelihood making use of the assumptions made of ε_t in Section 2. The use of maximum likelihood or quasi maximum likelihood makes it possible to obtain an idea of the uncertainty in the parameter estimates through (asymptotic) standard deviation estimates. This is not possible by using the above-mentioned algorithms. It may be argued that maximum likelihood estimation of neural network models is most likely to lead to convergence problems, and that penalizing the log-likelihood function one way or the other is a necessary precondition for satisfactory results. Two things can be said in favour of maximum likelihood here. First, in this paper model building 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 avoided. Second, the starting-values of the parameter estimates are chosen carefully, and we discuss the details of this in Section 4.3.2.

The AR-NN 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. Using model (1) as our starting-point, we make the following assumptions:

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

(A.2) The roots of the lag polynomial $1 - \sum_{j=1}^{p} \alpha_j z^j$ lie outside the unit circle. This is a sufficient

condition for weak stationarity, see Trapletti et al. (2000).

(A.3) The parameters satisfy the conditions $c_1 \leq \ldots \leq c_h$, $\gamma_i > 0$, $i = 1, ..., h$, and ω_{i1} is defined as in (9) for $i = 1, ..., h$.

Assumption (A.3) guarantees global identifiability of the model. The maximum likelihood estimator of the parameters of the conditional mean equals

$$\hat{\boldsymbol{\psi}} = \underset{\boldsymbol{\psi}}{\operatorname{argmax}} Q_T(\boldsymbol{\psi}) = -\frac{1}{2} \underset{\boldsymbol{\psi}}{\operatorname{argmin}} \sum_{t=1}^{T} q_t(\boldsymbol{\psi}),$$

where $q_t(\boldsymbol{\psi}) = (y_t - G(\mathbf{x}_t; \boldsymbol{\psi}))^2$. Conditions for consistency and asymptotic normality of $\hat{\boldsymbol{\psi}}$ are stated in the following theorem.

Theorem 1. Under the assumptions (A.1)–(A.3) the maximum likelihood estimator $\hat{\boldsymbol{\psi}}$ is almost surely consistent for $\boldsymbol{\psi}$ and

$$T^{1/2}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \to \mathrm{N}\left(\mathbf{0}, -\operatorname*{plim}_{T\to\infty} \mathbf{A}(\boldsymbol{\psi})^{-1}\right), \tag{11}$$

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

Proof. Assumptions (A.1)-(A.3) together with the normality assumption of the errors satisfy the assumptions of Theorem 4.1 (almost sure consistency) in Trapletti et al. (2000). As every hidden unit in model (1) is a logistic function, assumptions 4.5 and 4.6 of Theorem 4.2 (asymptotic normality) in Trapletti et al. (2000) are satisfied as well, so that result (11) follows.

In this paper, we apply the heteroskedasticity-robust large sample estimator of the covariance matrix of $\hat{\boldsymbol{\psi}}$ (White 1980)

$$\hat{\mathbf{B}}(\hat{\boldsymbol{\psi}}) = \left(\sum_{t=1}^{T} \hat{\mathbf{h}}_t \hat{\mathbf{h}}_t'\right)^{-1} \left(\sum_{t=1}^{T} \hat{\varepsilon}_t^2 \hat{\mathbf{h}}_t \hat{\mathbf{h}}_t'\right) \left(\sum_{t=1}^{T} \hat{\mathbf{h}}_t \hat{\mathbf{h}}_t'\right)^{-1} \tag{12}$$

where $\hat{\mathbf{h}}_t = \left.\frac{\partial q_t(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}$, and $\hat{\varepsilon}_t$ is the residual. In the estimation, the use of algorithms such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) or the Levenberg-Marquardt algorithms is strongly recommended. See, for example, Bertsekas (1995) for details about optimization algorithms or

Fine (1999, Chapter 5) for ones especially applied to the estimation of NN models. Choosing an appropriate line search procedure to select the step-length is another important question. Cubic or quadratic interpolation are usually reasonable choices. All the models in this paper are estimated with the Levenberg-Marquardt algorithm based on a cubic interpolation line search.

4.3.1 Concentrated Maximum Likelihood

In order to reduce the computational burden we can apply concentrated maximum likelihood to estimate $\boldsymbol{\psi}$ as follows. Consider the i^{th} iteration and rewrite model (1) as

$$\mathbf{y} = \mathbf{Z}(\boldsymbol{\phi})\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{13}$$

where $\mathbf{y}' = [y_1, y_2, \ldots, y_T]$, $\boldsymbol{\varepsilon}' = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T]$, $\boldsymbol{\theta}' = [\boldsymbol{\alpha}', \lambda_1, \ldots, \lambda_h]$, and

$$\mathbf{Z}(\boldsymbol{\phi}) = \begin{pmatrix} \tilde{\mathbf{x}}_1' & F(\gamma_1(\boldsymbol{\omega}_1'\mathbf{x}_1 - c_1)) & \ldots & F(\gamma_h(\boldsymbol{\omega}_h'\mathbf{x}_1 - c_h)) \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{x}}_T' & F(\gamma_1(\boldsymbol{\omega}_1'\mathbf{x}_T - c_1)) & \ldots & F(\gamma_h(\boldsymbol{\omega}_h'\mathbf{x}_T - c_h)) \end{pmatrix},$$

with $\boldsymbol{\phi} = [\gamma_1, \ldots, \gamma_h, \boldsymbol{\omega}_1', \ldots, \boldsymbol{\omega}_h', c_1, \ldots, c_h]'$. Assuming $\boldsymbol{\phi}$ fixed (the value is obtained from the previous iteration), the parameter vector $\boldsymbol{\theta}$ can be estimated analytically by

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{Z}(\boldsymbol{\phi})'\mathbf{Z}(\boldsymbol{\phi})\right)^{-1}\mathbf{Z}(\boldsymbol{\phi})'\mathbf{y}. \tag{14}$$

The remaining parameters $\boldsymbol{\phi}$ are estimated conditionally on $\boldsymbol{\theta}$ by applying the Levenberg-Marquardt algorithm, which completes the i^{th} iteration. This form of concentrated maximum likelihood was proposed by Leybourne, Newbold and Vougas (1998) in the context of STAR models. It substantially reduces the dimensionality of the iterative estimation problem, as instead of an inversion of a single large Hessian two smaller matrices are inverted, and a line search is only needed to obtain the i^{th} estimate of $\boldsymbol{\phi}$.

4.3.2 Starting-values

Many iterative optimization algorithms are sensitive to the choice of starting-values, and this is certainly so in the estimation of AR-NN models. Besides, an AR-NN model with h hidden units contains h parameters, $\gamma_i, i = 1, ..., h$, that are not scale-free. Our first task is thus to rescale the input variables in such a way that they have the standard deviation equal to unity. In the univariate AR-NN case, this simply means normalizing y_t. If the model contains exogenous variables, they are normalized separately. This, together with the fact that $\|\omega_h\| = 1$, gives us a basis for discussing the choice of starting-values of $\gamma_i, i = 1, ..., h$. Another advantage of this is that, in the multivariate case normalization generally makes numerical optimization easier as all variables have the same standard deviation. Assume now that we have estimated an AR-NN model with $h - 1$ hidden units and want to estimate one with h units. 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 h neurons is to use the final estimates for the parameters in the first $h-1$ hidden units and the linear unit. The starting-values for the parameters $\gamma_h, \theta_h, \omega_h$ and c_h in the hth hidden unit are obtained in three steps as follows.

1. For $k = 1, \ldots, K$:

   (a) Construct a vector $\mathbf{v}_h^{(k)} = [v_{1h}^{(k)}, \ldots, v_{qh}^{(k)}]'$ such that $v_{1h}^{(k)} \in (0, 1]$ and $v_{jh}^{(k)} \in [-1, 1]$, $j = 2, \ldots, q$. The values for $v_{1h}^{(k)}$ are drawn from a uniform $(0, 1]$ distribution and the ones for $v_{jh}^{(k)}$, $j = 2, \ldots, q$, from a uniform $[-1, 1]$ distribution.

   (b) Define $\boldsymbol{\omega}_h^{(k)} = \mathbf{v}_h^{(k)} \|\mathbf{v}_h^{(k)}\|^{-1}$, which guarantees $\|\boldsymbol{\omega}_h^{(k)}\| = 1$.

   (c) Let $c_h^{(k)} = \text{med}(\boldsymbol{\omega}_h^{(k)'} \mathbf{x})$, where $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$.

2. Define a grid of N positive values $\gamma_h^{(n)}$, $n = 1, \ldots, N$, for the slope parameter. This need not be done randomly. As the changes in γ_h have a small effect of the slope when γ_h is large, only a small number of large values are required, and the grid should be finer at the low end of the halfspace of γ_h-values.

3. For $k = 1, \ldots, K$ and $n = 1, \ldots, N$, estimate $\boldsymbol{\theta}$ using (14) and compute the value of $Q_T(\boldsymbol{\psi})$ for each combination of starting-values. Choose those values of the parameters that minimize

$Q_T(\boldsymbol{\psi})$ (they also maximize the concentrated log-likelihood function) as starting values.

After selecting the initial values of the h^{th} hidden unit we have to reorder the units if necessary in order to ensure that the identifying restrictions discussed in Section 3 are satisfied.

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

4.3.3 Potential Problem in the Estimation of the Slope Parameter

Finding the appropriate starting-values for the parameters is not the only difficult phase of the parameter estimation. It may also be difficult to obtain reasonably accurate estimates for those slope parameters γ_i, $i = 1, \ldots, h$, that are very large. This is the case unless the sample size is also large. To obtain an accurate estimate of a large γ_i it is necessary to have a large number of observations in the neighbourhood of c_i. When the sample size is not very large, there are generally few observations sufficiently close to c_i in the sample, which results in imprecise estimates of the slope parameter. This manifests itself in low absolute t-values for the estimates of γ_i. In such cases, the model builder cannot take a low absolute value of the t-statistic of the parameters of the transition function as evidence for omitting the hidden unit in question. Another reason for not doing so is that the t-value does not have its customary interpretation as a value of an asymptotic t-distributed statistic. This is due to an identification problem to be discussed in the next subsection. For more discussion see, for example, Bates and Watts (1988, p. 87) or Teräsvirta (1994).

4.4 Determining the Number of Hidden Units

The number of hidden units included in an ANN model is usually determined from the data. A popular method for doing that is pruning, in which a model with a large number of hidden units is estimated first, and the size of the model is subsequently reduced by applying an appropriate technique such as cross-validation. Another technique used in this connection is regularization, which may be characterized as penalized maximum likelihood or least squares applied to the estimation of neural network models. For discussion see, for example, Fine (1999, pp. 215–221). Bayesian

regularization, based on selecting a prior distribution for the parameters, may serve as an example. The use of regularization techniques precludes the possibility of applying methods of statistical inference to the estimated model.

As discussed in the Introduction, another possibility is to begin with a small model and sequentially add hidden units to the model, for discussion see, for example, Fine (1999, pp. 232–233), Anders and Korn (1999), or Swanson and White (1995,1997a,b). The decision of adding another hidden neuron is often based on the use of model selection criteria (MSC) or cross-validation. This has the following drawback. Suppose the data have been generated by an AR-NN model with h hidden units. Applying an MSC to decide whether or not another hidden unit should be added to the model requires estimation of a model with $h+1$ hidden neurons. In this situation, however, the larger model is not identified and its parameters cannot be estimated consistently. This is likely to cause numerical problems in maximum likelihood estimation. Besides, even when convergence is achieved, lack of identification causes a severe problem in interpreting the MSC. The NN model with h hidden units is nested in the model with $h+1$ units. A typical MSC comparison of the two models is then equivalent to a likelihood ratio test of h units against $h+1$ ones, see, for example, Teräsvirta and Mellin (1986) for discussion. The choice of MSC determines the (asymptotic) significance level of the test. But then, when the larger model is not identified under the null hypothesis, the likelihood ratio statistic does not have its customary asymptotic χ^2 distribution when the null holds. For more discussion of the general situation of a model only being identified under the alternative hypothesis, see, for example, Davies (1977,1987) or Hansen (1996). In the AR-NN case, this lack of identification shows as an ambiguity in determining the size of the penalty. An AR-NN model with $h+1$ hidden units can be reduced to one with h units by setting $\lambda_{h+1} = 0$, which suggests that the number of degrees of freedom in the penalty term should equal one. On the other hand, the $(h+1)^{\text{th}}$ hidden unit can also be eliminated by setting $\omega_{h+1} = \mathbf{0}$. This in turn suggests that the number of degrees of freedom should be $\dim(\mathbf{x}_t)$. In practice, the most common choice seems to equal $\dim(\mathbf{x}_t) + 2$.

We shall also select the hidden units sequentially starting from a small model, in fact from a linear one, but circumvent the identification problem in a way that enables us to control the significance level of the tests in the sequence and thus also the overall significance level of the

procedure. Following Teräsvirta and Lin (1993) we derive a test that is repeated until the first acceptance of the null hypothesis. Assume now that our AR-NN model (1) contains $h+1$ hidden units and write it as follows

$$y_t = \boldsymbol{\alpha}'\tilde{\mathbf{x}}_t + \sum_{i=1}^{h} \lambda_i F(\gamma_i(\boldsymbol{\omega}_i'\mathbf{x}_t - c_i)) + \lambda_{h+1} F(\gamma_{h+1}(\boldsymbol{\omega}_{h+1}'\mathbf{x}_t - c_{h+1})) + \varepsilon_t. \tag{15}$$

Assume further that we have accepted the hypothesis of model (15) containing h hidden units and want to test for the $(h+1)^{\text{th}}$ hidden unit. The appropriate null hypothesis is

$$\mathrm{H}_0 : \gamma_{h+1} = 0, \tag{16}$$

whereas the alternative is $\mathrm{H}_1 : \gamma_{h+1} > 0$. Under (16), the $(h+1)^{\text{th}}$ hidden unit is identically equal to a constant and merges with the intercept in the linear unit.

We assume that under (16) the assumptions of Theorem 1 hold so that the parameters of (15) can be estimated consistently. Model (15) is only identified under the alternative so that, as discussed above, the standard asymptotic inference is not available. This problem is circumvented as in Luukkonen, Saikkonen and Teräsvirta (1988) by expanding the $(h+1)^{\text{th}}$ hidden unit into a Taylor series around the null hypothesis (16). Using a third-order Taylor expansion, rearranging and merging terms results in the following model

$$y_t = \boldsymbol{\pi}'\tilde{\mathbf{x}}_t + \sum_{i=1}^{h} \lambda_i F(\gamma_i(\boldsymbol{\omega}_i'\mathbf{x}_t - c_i)) + \sum_{i=1}^{q}\sum_{j=i}^{q} \theta_{ij} x_{i,t} x_{j,t} + \sum_{i=1}^{q}\sum_{j=i}^{q}\sum_{k=j}^{q} \theta_{ijk} x_{i,t} x_{j,t} x_{k,t} + \varepsilon_t^*, \tag{17}$$

where $\varepsilon_t^* = \varepsilon_t + \lambda_{h+1} R(\mathbf{x}_t)$; $R(\mathbf{x}_t)$ is the remainder. It can be shown that $\theta_{ij} = \gamma_{h+1}^2 \tilde{\theta}_{ij}$, $\tilde{\theta}_{ij} \neq 0$, $i = 1, \ldots, q$; $j = i, \ldots, q$, and $\theta_{ijk} = \gamma_{h+1}^3 \tilde{\theta}_{ijk}$, $\tilde{\theta}_{ijk} \neq 0$, $i = 1, \ldots, q$; $j = i, \ldots, q$, $k = j, \ldots, q$. Thus the null hypothesis $\mathrm{H}_0' : \theta_{ij} = 0$, $i = 1, \ldots, q$; $j = i, \ldots, q$, $\theta_{ijk} = 0$, $i = 1, \ldots, q$; $j = i, \ldots, q$; $k = j, \ldots, q$. Note that under $\mathrm{H}_0 : \varepsilon_t^* = \varepsilon_t$, so that the properties of the error process remain unchanged under the null hypothesis. Finally, it may be pointed out that one may also view (17)

as resulting from a local approximation to the log-likelihood which for observation t takes the form

$$l_t = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left\{ y_t - \boldsymbol{\pi}'\tilde{\mathbf{x}}_t - \sum_{i=1}^{h}\lambda_i F(\gamma_i(\boldsymbol{\omega}_i'\mathbf{x}_t - c_i)) - \sum_{i=1}^{q}\sum_{j=i}^{q}\theta_{ij}x_{i,t}x_{j,t} - \sum_{i=1}^{q}\sum_{j=i}^{q}\sum_{k=j}^{q}\theta_{ijk}x_{i,t}x_{j,t}x_{k,t}\right\}^2. \tag{18}$$

We make the following assumption to accompany the previous assumptions (A.1)–(A.3):

(A.4) $E|x_{t,i}|^\delta < \infty$, $i = 1, \ldots, q$, for some $\delta > 6$. This enables us to state the following well-known result:

Theorem 2. Under $\mathrm{H}_0 : \gamma_{h+1} = 0$ and assumptions (A.1)–(A.4), the LM type statistic

$$LM = \frac{1}{\hat{\sigma}^2}\sum_{t=1}^{T}\hat{\varepsilon}_t\hat{\boldsymbol{\nu}}_t'\left\{\sum_{t=1}^{T}\hat{\boldsymbol{\nu}}_t\hat{\boldsymbol{\nu}}_t' - \sum_{t=1}^{T}\hat{\boldsymbol{\nu}}_t\hat{\mathbf{h}}_t'\left(\sum_{t=1}^{T}\hat{\mathbf{h}}_t\hat{\mathbf{h}}_t'\right)^{-1}\sum_{t=1}^{T}\hat{\mathbf{h}}_t\hat{\boldsymbol{\nu}}_t'\right\}\sum_{t=1}^{T}\hat{\boldsymbol{\nu}}_t\hat{\varepsilon}_t \tag{19}$$

where $\hat{\varepsilon}_t = y_t - G(\mathbf{x}_t; \hat{\boldsymbol{\pi}})$,

$$\begin{aligned}\hat{\mathbf{h}}_t = \left.\frac{\partial G(\mathbf{x}_t;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}'}\right|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} = \Bigg[&\tilde{\mathbf{x}}_t', F(\hat{\gamma}_1(\hat{\boldsymbol{\omega}}_1'\mathbf{x}_t - \hat{c}_1)), \ldots, F(\hat{\gamma}_h(\hat{\boldsymbol{\omega}}_h'\mathbf{x}_t - \hat{c}_h)),\\ &\hat{\lambda}_1\frac{\partial F(\hat{\gamma}_1(\hat{\boldsymbol{\omega}}_1'\mathbf{x}_t - \hat{c}_1))}{\partial\gamma_1}, \ldots, \hat{\lambda}_h\frac{\partial F(\hat{\gamma}_h(\hat{\boldsymbol{\omega}}_h'\mathbf{x}_t - \hat{c}_h))}{\partial\gamma_h},\\ &\hat{\lambda}_1\frac{\partial F(\hat{\gamma}_1(\hat{\boldsymbol{\omega}}_1'\mathbf{x}_t - \hat{c}_1))}{\partial\tilde{\omega}_{12}'}, \ldots, \hat{\lambda}_1\frac{\partial F(\hat{\gamma}_1(\hat{\boldsymbol{\omega}}_1'\mathbf{x}_t - \hat{c}_1))}{\partial\tilde{\omega}_{1q}'}, \ldots,\\ &\hat{\lambda}_h\frac{\partial F(\hat{\gamma}_h(\hat{\boldsymbol{\omega}}_h'\mathbf{x}_t - \hat{c}_h))}{\partial\tilde{\omega}_{h2}'}, \ldots, \hat{\lambda}_h\frac{\partial F(\hat{\gamma}_h(\hat{\boldsymbol{\omega}}_h'\mathbf{x}_t - \hat{c}_h))}{\partial\tilde{\omega}_{hq}'},\\ &\hat{\lambda}_1\frac{\partial F(\hat{\gamma}_1(\hat{\boldsymbol{\omega}}_1'\mathbf{x}_t - \hat{c}_1))}{\partial c_1}, \ldots, \hat{\lambda}_h\frac{\partial F(\hat{\gamma}_h(\hat{\boldsymbol{\omega}}_h'\mathbf{x}_t - \hat{c}_h))}{\partial c_h}\Bigg]'\end{aligned}$$

with

$$\begin{aligned}\frac{\partial F(\hat{\gamma}_i(\hat{\boldsymbol{\omega}}_i'\mathbf{x}_t - \hat{c}_i))}{\partial\gamma_i} &= (\hat{\boldsymbol{\omega}}_i'\mathbf{x}_t - \hat{c}_i)\left[2\cosh\left(\hat{\gamma}_i(\hat{\boldsymbol{\omega}}_i'\mathbf{x}_t - \hat{c}_i)\right)\right]^{-2}, i = 1, \ldots, h\\ \frac{\partial F(\hat{\gamma}_i(\hat{\boldsymbol{\omega}}_i'\mathbf{x}_t - \hat{c}_i))}{\partial\tilde{\omega}_{ij}'} &= \gamma_i\tilde{x}_{j,t}\left[2\cosh\left(\hat{\gamma}_i(\hat{\boldsymbol{\omega}}_i'\mathbf{x}_t - \hat{c}_i)\right)\right]^{-2}, i = 1, \ldots, h, j = 2, \ldots, q\\ \frac{\partial F(\hat{\gamma}_i(\hat{\boldsymbol{\omega}}_i'\mathbf{x}_t - \hat{c}_i))}{\partial c_i} &= -\gamma_i\left[2\cosh\left(\hat{\gamma}_i(\hat{\boldsymbol{\omega}}_i'\mathbf{x}_t - \hat{c}_i)\right)\right]^{-2}, i = 1, \ldots, h\end{aligned}$$

and $\boldsymbol{\nu}_t = \left[x_{1,t}^2, x_{1,t}x_{2,t}, \ldots, x_{i,t}x_{j,t}, \ldots, x_{1,t}^3, \ldots, x_{i,t}x_{j,t}x_{k,t}, \ldots, x_{h,t}^3\right]$, has an asymptotic χ^2 distribution with $m = q(q+1)/2 + q(q+1)(q+2)/6$ degrees of freedom.

The test can also be carried out in stages as follows:

1. Estimate model (1) with h hidden units. If the sample size is small and the model thus difficult to estimate, numerical problems in applying the maximum likelihood algorithm may lead to a solution such that the residual vector is not precisely orthogonal to the gradient matrix of $G(\mathbf{x}_t; \hat{\boldsymbol{\psi}})$. This has an adverse effect on the empirical size of the test. To circumvent this problem, we regress the residuals $\hat{\varepsilon}_t$ on $\hat{\mathbf{h}}_t$ and compute the sum of squared residuals $SSR_0 = \sum_{t=1}^{T} \tilde{\varepsilon}_t^2$. The new residuals $\tilde{\varepsilon}_t$ are orthogonal to $\hat{\mathbf{h}}_t$.

2. Regress $\tilde{\varepsilon}_t$ on $\hat{\mathbf{h}}_t$ and $\hat{\boldsymbol{\nu}}_t$. Compute the sum of squared residuals $SSR_1 = \sum_{t=1}^{T} \hat{v}_t^2$.

3. Compute the χ^2 statistic

$$LM_{\chi^2}^{hn} = T\frac{SSR_0 - SSR_1}{SSR_0}, \tag{20}$$

or the F version of the test

$$LM_F^{hn} = \frac{(SSR_0 - SSR_1)/m}{SSR_1/(T-n-m)}, \tag{21}$$

where $n = (q+2)h + p + 1$. Under H_0, $LM_{\chi^2}^{hn}$ has an asymptotic χ^2 distribution with m degrees of freedom and LM_F^{hn} is approximately F-distributed with m and $T-n-m$ degrees of freedom.

The following cautionary remark is in order. If any $\hat{\gamma}_i$, $i = 1, \ldots, h$, is very large, the gradient matrix becomes near-singular and the test statistic numerically unstable, which distorts the size of the test. The reason is that the vectors corresponding to the partial derivatives with respect to γ_i, $\boldsymbol{\omega}_i$, and c_i, respectively, tend to be almost perfectly linearly correlated. This is due to the fact that the time series of those elements of the gradient resemble dummy variables being constant most of the time and nonconstant simultaneously. The problem may be remedied by omitting these elements from the regression in step 2. This can be done without significantly affecting the value of the test statistic; see Eitrheim and Teräsvirta (1996) for discussion. This situation may also cause problems in obtaining standard deviation estimates for parameter estimates through the outer product matrix (12). The same remedy can be applied: omit the rows and columns

corresponding to the two parameters and invert the reduced matrix. This yields more reliable standard deviation estimates for the remaining parameter estimates. In fact, omitting the rows and columns corresponding to the high values $\hat{\gamma}_i$ should suffice.

Testing zero hidden units against at least one is a special case of the above test. This amounts to testing linearity, and the test statistic is in this case identical to the one derived for testing linearity against the AR-NN model in Teräsvirta, Lin and Granger (1993). A natural alternative to our procedure is the one first suggested in White (1989) and investigated later in Lee, White and Granger (1993). In order to test the null hypothesis of h hidden units, one adds q hidden units to model (1) by randomly selecting the parameters $\tilde{\omega}_{h+j}$, β_{h+j}, $j = 1, \ldots, q$. If q is large (a small value may not render the test powerful enough), a small number of principal components of the q extra units may be used instead. This solves the identification problem as the extra neurons or the transformed neurons are observable, and the null hypothesis $\lambda_{h+1} = \cdots = \lambda_{h+q} = 0$ or its equivalent when the principal component approach is taken, can be tested using standard inference. When $h = 0$, this technique also collapses into a linearity test: see Lee et al. (1993). Simulation results in Teräsvirta et al. (1993) and Anders and Korn (1999) indicate that the polynomial approximation method presented here compares well with White's approach, and it is applied in the rest of this work.

As mentioned in the Introduction, the normality assumption can be relaxed while the consistency and asymptotic normality of the (quasi) maximum likelihood estimators are retained. This is important, for example, in financial applications of the AR-NN model. In financial applications, at least in ones to high-frequency data, such as intradaily, daily or even weekly series, the series typically contain conditional heteroskedasticity. This possibility can be accounted for by robustifying the tests against heteroskedasticity following Wooldridge (1990). A heteroskedasticity-robust version of the LM type test, based on the notion of robustifying statistic (18), can be carried out as follows.

1. As before.

2. Regress $\hat{\boldsymbol{\nu}}_t$ on $\hat{\mathbf{h}}_t$ and compute the residuals $\mathbf{r}_t$.

3. Regress 1 on $\tilde{\varepsilon}_t \mathbf{r}_t$ and compute the sum of squared residuals SSR_1.

4. Compute the value of the test statistic

$$LM_{\chi^2}^r = T - SSR_1. \tag{22}$$

The test statistic has the same asymptotic χ^2 null distribution as before.

It should be noticed that in the case of conditional heteroskedasticity, the maximum likelihood estimates discussed in Section 4.3 are just quasi maximum likelihood estimates. Under regularity conditions they are still consistent and asymptotically normal.

4.5 Evaluation of the Estimated Model

After a model has been estimated it has to be evaluated. We propose two in-sample misspecification tests for this purpose. The first one tests for the instability of the parameters. The second one tests the assumption of no serial correlation in the errors and is an application of the results in Eitrheim and Teräsvirta (1996) and Godfrey (1988, pp. 112–121).

4.5.1 Test of Parameter Constancy

Testing parameter constancy is an important way of checking the adequacy of linear or nonlinear models. Many parameter constancy tests are tests against unspecified alternatives or a single structural break. In this section we present a parametric alternative to parameter constancy which allows the parameters to change smoothly as a function of time under the alternative hypothesis. In the following we assume that the logistic function (8) has constant parameters whereas both $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}_i$, $i = 1, \ldots, h$, may be subject to changes over time. This assumption is made mainly because changes in the parameters of the logistic function are more difficult to detect than changes in the linear parameters.

In order to derive the test, consider a model with time-varying parameters defined as

$$y_t = \tilde{G}(\mathbf{x}_t; \boldsymbol{\psi}, \tilde{\boldsymbol{\psi}}) + \varepsilon_t = \tilde{\boldsymbol{\alpha}}'(t)\tilde{\mathbf{x}}_t + \sum_{i=1}^{h} \left\{ \tilde{\lambda}_i(t) F(\gamma_i(\boldsymbol{\omega}_i'\mathbf{x}_t - c_i)) \right\} + \varepsilon_t, \tag{23}$$

where

$$\tilde{\boldsymbol{\alpha}}(t) = \boldsymbol{\alpha} + \breve{\boldsymbol{\alpha}} F\left(\zeta(t-\eta)\right), \tag{24}$$

and

$$\tilde{\lambda}_i(t) = \lambda_i + \breve{\lambda}_i F\left(\zeta(t-\eta)\right), i = 1, \ldots, h. \tag{25}$$

The function F in (24) and (25) is defined as in (2), and $\zeta > 0$. The parameter vector $\boldsymbol{\psi}$ is defined as before, and $\tilde{\boldsymbol{\psi}} = [\breve{\boldsymbol{\alpha}}, , \breve{\lambda}_1, \ldots, \breve{\lambda}_h, \zeta, \eta]'$. The parameter ζ controls the smoothness of the monotonic change in the autoregressive parameters. When $\zeta \to \infty$, equations (23)–(25) represent a model with a single structural break at $t = \eta$. Combining (24) and (25) with (23), we have the following model

$$y_t = \left\{\boldsymbol{\alpha}' + \breve{\boldsymbol{\alpha}}' F\left(\zeta(t-\eta)\right)\right\} \tilde{\mathbf{x}}_t + \sum_{i=1}^{h} \left\{\lambda_i + \breve{\lambda}_i F\left(\zeta(t-\eta)\right)\right\} F(\gamma_i(\boldsymbol{\omega}_i' \mathbf{x}_t - c_i)) + \varepsilon_t. \tag{26}$$

The null hypothesis hypothesis of parameter constancy is

$$\mathrm{H}_0 : \zeta = 0. \tag{27}$$

Note that model (26) is only identified under the alternative $\zeta > 0$. As is obvious from Section 4.4, a consequence of this complication is that the standard asymptotic distribution theory for the likelihood ratio or other classical test statistics for testing (27) is not available. To remedy this problem we expand $F\left(\zeta(t-\eta)\right)$ into a first-order Taylor expansion around $\zeta = 0$. This yields

$$t_1 = \frac{1}{4}\zeta(t-\eta) + R(t;\zeta,\eta), \tag{28}$$

where $R(t;\zeta,\eta)$ is the remainder. Replacing $F\left(\zeta(t-\eta)\right)$ in (26) by (28) and reparameterizing we obtain

$$y_t = \left(\boldsymbol{\theta}_0' + \boldsymbol{\mu}_0' t\right) \tilde{\mathbf{x}}_t + \sum_{i=1}^{h} \left(\theta_i + \mu_i t\right) F(\gamma_i(\boldsymbol{\omega}_i' \mathbf{x}_t - c_i)) + \varepsilon_t^*, \tag{29}$$

where $\boldsymbol{\theta}_0 = \boldsymbol{\alpha} - \breve{\boldsymbol{\alpha}}\zeta\eta/4$, $\boldsymbol{\mu}_0 = \breve{\boldsymbol{\alpha}}\zeta/4$, $\theta_i = \lambda_i - \breve{\lambda}_i\zeta\eta/4$, $\mu_i = \breve{\lambda}_i\zeta/4$, $i = 1, \ldots, h$, and $\varepsilon_t^* =$

$\varepsilon_t + R(t;\zeta,\eta)$. The null hypothesis becomes

$$\mathrm{H}_0 : \boldsymbol{\mu}_0 = \mathbf{0}, \mu_1 = \cdots = \mu_h = 0. \tag{30}$$

Under H_0, $R(t;\zeta,\eta) = 0$ and $\varepsilon_t^* = \varepsilon_t$, so that standard asymptotic distribution theory is available. The local approximation to the t^{th} term in the normal log likelihood function in a neighbourhood of H_0 for observation t, ignoring $R(t;\zeta,\eta)$, is

$$\begin{aligned} l_t = &-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 \\ &-\frac{1}{2\sigma^2}\left\{ y_t - \left(\boldsymbol{\theta}_0' + \boldsymbol{\mu}_0' t\right)\tilde{\mathbf{x}}_t - \sum_{i=1}^{h}\left(\theta_i + \mu_i t\right) F(\gamma_i(\boldsymbol{\omega}_i'\mathbf{x}_t - c_i)) \right\}^2 . \end{aligned} \tag{31}$$

The consistent estimators of the relevant partial derivatives of the log likelihood under the null hypothesis are

$$\left.\frac{\partial \hat{l}_t}{\partial \boldsymbol{\theta}_0'}\right|_{\mathrm{H}_0} = \frac{1}{\hat{\sigma}^2}\hat{\varepsilon}_t\tilde{\mathbf{x}}_t \tag{32}$$

$$\left.\frac{\partial \hat{l}_t}{\partial \boldsymbol{\mu}_0'}\right|_{\mathrm{H}_0} = \frac{1}{\hat{\sigma}^2}\hat{\varepsilon}_t t\tilde{\mathbf{x}}_t \tag{33}$$

$$\left.\frac{\partial \hat{l}_t}{\partial \theta_i}\right|_{\mathrm{H}_0} = \frac{1}{\hat{\sigma}^2}\hat{\varepsilon}_t F(\hat{\gamma}_i(\hat{\boldsymbol{\omega}}_i'\mathbf{x}_t - \hat{c}_i)) \tag{34}$$

$$\left.\frac{\partial \hat{l}_t}{\partial \mu_i'}\right|_{\mathrm{H}_0} = \frac{1}{\hat{\sigma}^2}\hat{\varepsilon}_t t F(\hat{\gamma}_i(\hat{\boldsymbol{\omega}}_i'\mathbf{x}_t - \hat{c}_i)) \tag{35}$$

$$\left.\frac{\partial \hat{l}_t}{\partial \gamma_i'}\right|_{\mathrm{H}_0} = \frac{1}{\hat{\sigma}^2}\hat{\varepsilon}_t\hat{\theta}_i\frac{\partial F(\hat{\gamma}_i(\hat{\boldsymbol{\omega}}_i'\mathbf{x}_t - \hat{c}_i))}{\partial \gamma_i'} \tag{36}$$

$$\left.\frac{\partial \hat{l}_t}{\partial \boldsymbol{\omega}_i'}\right|_{\mathrm{H}_0} = \frac{1}{\hat{\sigma}^2}\hat{\varepsilon}_t\hat{\theta}_i\frac{\partial F(\hat{\gamma}_i(\hat{\boldsymbol{\omega}}_i'\mathbf{x}_t - \hat{c}_i))}{\partial \boldsymbol{\omega}_i'} \tag{37}$$

$$\left.\frac{\partial \hat{l}_t}{\partial c_i}\right|_{\mathrm{H}_0} = \frac{1}{\hat{\sigma}^2}\hat{\varepsilon}_t\hat{\theta}_i\frac{\partial F(\hat{\gamma}_i(\hat{\boldsymbol{\omega}}_i'\mathbf{x}_t - \hat{c}_i))}{\partial c_i} \tag{38}$$

where $i = 1, \ldots, h$, $\hat{\sigma}^2 = (1/T)\sum_{t=1}^{T}\hat{\varepsilon}_t^2$, and $\hat{\varepsilon}_t = y_t - G(\mathbf{x}_t;\hat{\boldsymbol{\psi}}) = y_t - \hat{\boldsymbol{\alpha}}'\tilde{\mathbf{x}}_t - \sum_{i=1}^{h}\hat{\lambda}_i F(\hat{\gamma}_i(\hat{\boldsymbol{\omega}}_i'\mathbf{x}_t - \hat{c}_i))$ are

the residuals estimated under the null hypothesis. The LM statistic is (19) with $\hat{\mathbf{h}}_t = \frac{\partial G(\mathbf{x}_t;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}'}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}$ and $\hat{\boldsymbol{\nu}}_t = \left[t\tilde{\mathbf{x}}_t', tF(\hat{\gamma}_1(\hat{\boldsymbol{\omega}}_1'\mathbf{x}_t - \hat{c}_1)), \ldots, tF(\hat{\gamma}_h(\hat{\boldsymbol{\omega}}_h'\mathbf{x}_t - \hat{c}_h))\right]'$. Testing hypothesis of just a subset of coefficients being constant is also possible in this framework. Under H_0, statistic (19) has an asymptotic χ^2 distribution with $q + h + 1$ degrees of freedom. This is the case even if $\hat{\boldsymbol{\nu}}_t$ contains elements dominated by a deterministic trend, see Lin and Teräsvirta (1994) for details.

The test can be carried out in stages as before. The only differences are the new definition of $\hat{\boldsymbol{\nu}}_t$ at stage 2 and the degrees of freedom in the χ^2 or F test. As before, use of F-version of the test is recommended.

The alternative to parameter constancy may be made more general simply by defining

$$F(\zeta(t-c)) = \left(1 + \exp\left\{-\zeta \prod_{k=1}^{K}(t - \eta_k)\right\}\right)^{-1}, \zeta > 0 \tag{39}$$

with $K \geq 1$. Transition function (39) with $K \geq 2$ allows the parameters to change nonmonotonically over time, in contrast to the case $K = 1$. Consequently, in the test statistic (19), $\hat{\boldsymbol{\nu}}_t = [\hat{\boldsymbol{\nu}}_{1t}', \ldots, \hat{\boldsymbol{\nu}}_{Kt}']'$ where

$$\hat{\boldsymbol{\nu}}_{kt}' = \left[t^k\mathbf{x}_t', t^kF(\hat{\gamma}_1(\hat{\boldsymbol{\omega}}_1'\mathbf{x}_t - \hat{c}_1)), \ldots, t^kF(\hat{\gamma}_h(\hat{\boldsymbol{\omega}}_h'\mathbf{x}_t - \hat{c}_h)\right], \quad k = 1, \ldots, K$$

and the number of degrees of freedom in the test statistic is adjusted accordingly. In this paper we report results for $K = 1, 2, 3$, and call the corresponding test statistics $LM_K, K = 1, 2, 3$, respectively. If the error process is heteroskedastic, a robust version of the test, immediately obvious from Section 4.4, has to be employed instead of the standard test. When the model is assumed not to contain any hidden units: $\lambda_i(t) \equiv 0$, $i = 1, \ldots, h$, the test collapses into the parameter constancy test in Lin and Teräsvirta (1994).

4.5.2 Test of Serial Independence

Assume that the errors in equation (1) follow an r^{th} order autoregressive process defined as

$$\varepsilon_t = \boldsymbol{\pi}'\boldsymbol{\nu}_t + u_t, \tag{40}$$

where $\boldsymbol{\pi}' = [\pi_1, \ldots, \pi_r]$ is a parameter vector, $\boldsymbol{\nu}_t' = [\varepsilon_{t-1}, \ldots, \varepsilon_{t-r}]$, and $u_t \sim \text{NID}(0, \sigma^2)$. We assume that under H_0, assumptions (A.1)–(A.3) hold. Consider the null hypothesis $\text{H}_0 : \boldsymbol{\pi} = \mathbf{0}$ whereas $\text{H}_1 : \boldsymbol{\pi} \neq \mathbf{0}$. The conditional normal log-likelihood of (1) with (40) for observation t, given the fixed starting values has the form

$$\begin{aligned} l_t = &-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 \\ &-\frac{1}{2\sigma^2}\left\{ y_t - \sum_{j=1}^{r}\pi_j y_{t-j} - G(\mathbf{x}_t;\boldsymbol{\psi}) + \sum_{j=1}^{r}\pi_j G(\mathbf{x}_{t-j};\boldsymbol{\psi}) \right\}^2 . \end{aligned} \tag{41}$$

The first partial derivatives of the normal log-likelihood for observation t with respect to $\boldsymbol{\pi}$ and $\boldsymbol{\psi}$ are

$$\begin{aligned} \frac{\partial l_t}{\partial \pi_j} &= \left(\frac{u_t}{\sigma^2}\right)\left\{ y_{t-j} - G(\mathbf{x}_{t-j};\boldsymbol{\psi}) \right\}, j = 1, \ldots, r \\ \frac{\partial l_t}{\partial \boldsymbol{\psi}} &= -\left(\frac{u_t}{\sigma^2}\right)\left\{ \frac{\partial G(\mathbf{x}_t;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} - \sum_{j=1}^{r}\pi_j \frac{\partial G(\mathbf{x}_{t-j};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \right\}. \end{aligned} \tag{42}$$

Under the null hypothesis, the consistent estimators of the score are

$$\sum_{t=1}^{T}\frac{\partial \hat{l}_t}{\partial \boldsymbol{\pi}}\bigg|_{\text{H}_0} = \frac{1}{\hat{\sigma}^2}\sum_{t=1}^{T}\hat{\varepsilon}_t\hat{\boldsymbol{\nu}}_t \quad \text{and} \quad \sum_{t=1}^{T}\frac{\partial \hat{l}_t}{\partial \boldsymbol{\psi}}\bigg|_{\text{H}_0} = -\frac{1}{\hat{\sigma}^2}\sum_{t=1}^{T}\hat{\varepsilon}_t\hat{\mathbf{h}}_t,$$

where $\hat{\boldsymbol{\nu}}_t' = [\hat{\varepsilon}_{t-1}, \ldots, \hat{\varepsilon}_{t-r}]$, $\hat{\varepsilon}_{t-j} = y_{t-j} - G(\mathbf{x}_{t-j};\hat{\boldsymbol{\psi}})$, $j = 1, \ldots, r$, $\hat{\mathbf{h}}_t = \frac{\partial G(\mathbf{x}_t;\hat{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}}$, and $\hat{\sigma}^2 = (1/T)\sum_{t=1}^{T}\hat{\varepsilon}_t^2$. The LM statistic is (19) with $\hat{\mathbf{h}}_t$ and $\hat{\boldsymbol{\nu}}_t$ defined as above, and it has an asymptotic χ^2 distribution with r degrees of freedom under the null hypothesis. For details, see Godfrey (1988, pp. 112–121).

The test can be performed in three stages as shown before. It may be pointed out that the Ljung-Box test or its asymptotically equivalent counterpart, the Box-Pierce test, both recommended for use in connection with NN models by Zapranis and Refenes (1999), are not available. Their asymptotic null distribution is unknown when the estimated model is an AR-NN model.

4.6 Modelling strategy

At this point we are ready to combine the above statistical ingredients into a coherent modelling strategy. We first define the potential variables (lags) and select a subset of them applying the variable selection technique considered in Section 4.2. After selecting the variables we select the number of hidden units sequentially. We begin testing linearity against a single hidden unit as described in Section 4.4 at significance level α. The model under the null hypothesis is simply a linear AR(p) model. If the null hypothesis is not rejected, the AR model is accepted. In case of a rejection, an AR-NN model with a single unit is estimated and tested against a model with two hidden units at the significance level $\alpha\varrho$, $0 < \varrho < 1$. Another rejection leads to estimating a model with two hidden units and testing it against a model with three hidden neurons at the significance level $\alpha\varrho^2$. The sequence is terminated at the first acceptance of the null hypothesis. The significance level is reduced at each step of the sequence and converges to zero. In the applications in Section 6, we use $\varrho = 1/2$. This way we avoid excessively large models and control the overall significance level of the procedure. An upper bound for the overall significance level α^* may be obtained using the Bonferroni bound. For example, if $\alpha = 0.1$, and $\varrho = 1/2$ then $\alpha^* \leq 0.187$. Note that if we instead of our LM type test apply a model selection criterion such as AIC or SBIC to this sequence, we in fact use the same significance level at each step. Besides, the upper bound that can be worked out in the linear case, see, for example, Teräsvirta and Mellin (1986), remains unknown due to the identification problem mentioned above.

In following the above path we have indeed assumed that all hidden neurons contain the variables that are originally selected to the AR-NN model. Another variant of the strategy is the one in which the variables in each hidden unit are chosen individually from the set of originally selected variables. In the present context this may be done, for example, by considering the estimated parameter vector $\hat{\boldsymbol{\omega}}_h$ of the most recently added hidden neuron, removing the variables whose coefficients have the lowest t-values and re-estimating the model. Anders and Korn (1999) recommended this alternative. It has the drawback that the computational burden may become high as frequent estimation of neural network models may be involved. Because of this we suggest another technique that combines sequential testing for hidden units and variable selection. Consider equation (15). Instead of just testing a single null hypothesis as is done within (15), we can do the following. First test the null

hypothesis involving all variables. Then remove one variable from the extra unit under test and test the model with h hidden units against this reduced alternative. Remove each variable in turn and carry out the test. Continue by removing two variables at a time. Finally, test the model with h neurons against the alternatives in which the $(h+1)^{\text{th}}$ unit only contains a single variable and an intercept. Find the combination of variables for which the p-value of the test is minimized. If this p-value is lower than a prescribed value, "significance level", add the $(h+1)^{\text{th}}$ unit with the corresponding variables to the model. Otherwise accept the AR-NN model with h hidden units and stop. This way of selecting the variables for each hidden unit is analogous to the variable selection technique discussed in Section 4.2.

Compared to our first strategy, this one adds to the flexibility and on the average leads to more parsimonious models than the other one. On the other hand, as every choice of hidden unit involves a possibly large number of tests, we do not control the significance level of the overall hidden unit test. We do that, albeit conditionally on the variables selected, if the set of input variables is determined once and for all before choosing the number of hidden units.

Evaluation following the estimation of the final model is carried out by subjecting the model to misspecification tests discussed in Section 4.5. If the model does not pass the tests, the model builder has to reconsider the specification. For example, important lags (or, in the more general case, exogenous variables), may be missing from the model. If parameter constancy is rejected, model (23) may be estimated and used for forecasting, but reconsidering the whole specification may often be a more sensible option of the two.

Another way of evaluating the model is out-of-sample forecasting. As AR-NN models are most often constructed for forecasting purposes, this is important. This part of the model evaluation is carried out by saving the last observations in the series for forecasting and comparing the forecast results with those from at least one benchmark model. Note, however, that the results are dependent on the observations contained in that particular prediction period and may not allow very general conclusions about the properties of the estimated model. They are also conditional on the structure of the model remaining unchanged over the forecasting period, which may not necessarily be the case. For more discussion about this, see Clements and Hendry (1999, Chapter 2).

4.7 Discussion and comparisons

It is useful to compare our modelling strategy with other bottom-up approaches available in the literature. Swanson and White (1995), Swanson and White (1997a), and Swanson and White (1997b) apply the SBIC model selection criterion (Schwarz 1978) as follows. They start with a linear model, adding potential variables to it until SBIC indicates that the model cannot be further improved. Then they estimate models with a single hidden unit and select regressors sequentially to it one by one unless SBIC shows no further improvement. Next Swanson and White add another hidden unit and proceed by adding variables to it. The selection process is terminated when SBIC indicates that no more hidden units or variables should be added or when a predetermined maximum number of hidden units has been reached. This modelling strategy can be termed fully sequential.

Anders and Korn (1999) essentially adopt the procedure of Teräsvirta and Lin (1993) described in Section 4.4 for selecting the number of hidden units. After estimating the largest model they suggest proceeding from general-to-specific by sequentially removing those variables from hidden units whose parameter estimates have the lowest (t-test) p-values. Note that this presupposes parameterizing the hidden units as in (2), not as in (8) and (9).

Balkin and Ord (2000) select the ordered variables (lags) sequentially using a linear model and a forward stepwise regression procedure. If the F-test statistic of adding another lag obtains a value exceeding 2, this lag is added to the set of input variables. The number of variables selected also serves as a maximum number of hidden units. The authors suggest estimating all models from the one with a single hidden unit up to the one with the maximum number of neurons. The final choice is made using the Generalized Cross-Validation Criterion of Golub, Heath and Wahba (1979). The model for which the value of this model selection criterion is minimized is selected.

Refenes and Zapranis (1999) (see also Zapranis and Refenes (1999)) propose adding hidden units into the model sequentially (there is a flow chart in the paper indicating this). The number of units, however, is selected only after adding all units up to a predetermined maximum number, so that the procedure is not genuinely sequential. The choice is made by applying the Network Information Criterion (Murata, Yoshizawa and Amari 1994) that can be traced back to Stone (1977). The model is then pruned by removing redundant variables from the neurons and re-estimating the model.

Unlike the others, Refenes and Zapranis (1999) underline the importance of misspecification testing which also forms an integral part of our modelling procedure. They suggest, for example, that the hypothesis of no error autocorrelation should be tested, by the Ljung-Box or the asymptotically equivalent Box-Pierce test. Unfortunately, these tests do not have their customary asymptotic null distribution when the estimated model is an AR-NN model instead of a linear autoregressive one.

Of these strategies, the Swanson and White one is computationally the most intensive one, as the number of steps involving an estimation of a NN model is large. Our procedure is in this respect the least demanding. The difference between our scheme and the Anders and Korn one is that in our strategy, variable selection does not require estimation of NN models because it is wholly based on LM type tests (the model is only estimated under the null hypothesis). Furthermore, there is a possibility of omitting certain potential variables before even estimating neural network models.

Like ours, the Swanson and White strategy is truly sequential: the modeller proceeds by considering nested models. The difference lies in how to compare two nested models in the sequence. Swanson and White apply SBIC whereas Anders and Korn and we use LM type tests. The problems with the former technique have been discussed in Section 4.4. The problem of estimating unidentified models is still more acute in the approaches of Balkin and Ord and Refenes and Zapranis. Because these procedures require the estimation of NN models up to one containing a predetermined maximum number of hidden units, several estimated models may thus be unidentified. The problem is even more serious if statistical inference is applied in subsequent pruning as the selected model may also be unidentified. The probability of this happening is smaller in the Anders and Korn case, in particular when the sequence of hidden unit tests has gradually decreasing significance levels.

5 Monte-Carlo Study

In this section we report results from two Monte Carlo experiments. The purpose of the first one is to illustrate some features of the NN model selection strategy described in Section 4.6 and compare it with the alternative in which model selection is carried out using an appropriate model selection criterion. In the second experiment, the performance of the misspecification tests of Section 4.5 is

considered. In both experiments we make use of the following model

$$
\begin{aligned}
y_t = {} & 0.10 + [0.75 - 0.2\pi \times F^*(0.05(t-90))]\, y_{t-1} - 0.05 y_{t-4} \\
& + [0.80 - 0.5\pi \times F^*(0.05(t-90))]\, F(2.24(0.45 y_{t-1} - 0.89 y_{t-4} + 0.09)) \\
& + [-0.70 + 2.0\pi \times F(0.05(t-90))]\, F(1.12(0.44 y_{t-1} + 0.89 y_{t-4} + 0.35)) + \varepsilon_t, \\
\varepsilon_t = {} & \kappa \varepsilon_{t-1} + u_t, \quad u_t \sim \mathrm{NID}(0, \sigma^2)
\end{aligned}
\tag{43}
$$

where F^* is defined as in 43) with $K = 1$ or 2. However, in the first experiment, $\pi = \kappa = 0$. The number of observations in the first experiment is either 200 or 1000, in the second one we report results for 100 observations. In every replication, the first 500 observations are discarded to eliminate the initialization effects. The number of replications equals 500.

5.1 Architecture Selection

Results from simulating the modelling strategy can be found in Table 1. The table also contains results on choosing the number of hidden units using SBIC. This model selection criterion was chosen for the experiment because Swanson and White (1995, 1997a,b) applied it to this problem. In this case it is assumed that the model contains the correct variables. This is done in order to obtain an idea of the behaviour of SBIC free from the effects of an incorrectly selected model.

Different results can be obtained by varying the error variance, the size of the coefficients of hidden units or "connection strengths" and, in the case of our strategy, the significance levels. In this experiment, the significance level is halved at every step, but other choices are of course possible. It seems, at least in the present experiment, that selecting the variables is easier than choosing the right number of hidden units. In small samples, there is a strong tendency to choose a linear model but, as can be expected, nonlinearity becomes more apparent with an increasing sample size. The larger initial significance level ($\alpha = 0.10$) naturally leads to larger models on the average than the smaller one ($\alpha = 0.05$). Overfitting is relatively rare but the results suggest, again not unexpectedly, that the initial significance level should be lowered when the number of observations increases. Finally, improving the signal-to-noise ratio improves the performance of our strategy.

The results of the hidden unit selection by SBIC show that the empirical significance level implied by it is, at least in this experiment, very low for both $T = 200$ and $T = 1000$, although it changes with the sample size. Compared to our approach, the linear model is still chosen relatively often for $T = 1000$ whereas the correct model with two hidden units is not selected at all. A drawback of SBIC is that the significance level is not known to the model builder, whereas in our strategy it can be controlled at will.

5.2 Parameter Constancy Tests

The parameter constancy test statistic is simulated for $K = 1, 2$ in (39). The results of size simulations are presented using size discrepancy plots introduced in Davidson and MacKinnon (1998). They can be found in Figure 1. The results are based on realizations with 100 observations and two error standard deviations, $\sigma = 0.125$ and $\sigma = 0.25$. For the latter one, the test is seen to be conservative, whereas this tendency disappears at customary levels of significance (up to 0.10) when the signal-to-noise ratio is increased (the error standard deviation decreased). Panels (d)-(f) show that the size is somewhat sensitive to error autocorrelation.

The results of power simulations are not shown here. The reason is that they are what can be expected: the test has reasonable power when smooth parameter change is present in model (43). The results are available from the authors upon request.

5.3 Serial Independence Test

The test of no error autocorrelation is simulated using model (43) with $\pi = 0, 1$ and $\sigma = 0.125, 0.25$. The maximum lags in the alternative equal 1,2 and 4. The size discrepancy plots appear in Figure 2. Again, the test is somewhat conservative for $\sigma = 0.25$ and less so for $\sigma = 0.125$. Not unexpectedly, panels (c) and (d) show that the test has power against time-varying parameters ($\pi = 1$ in (43)). The results of power simulations do not offer any surprises: the power increases with parameter κ. They are thus omitted to save space but are available upon request.

Table 1: Outcomes of the experiments of selecting the number of hidden units using the test sequence starting at significance levels $\alpha = 0.05$ and 0.10 and sample sizes 200 and 1000 based on 500 replications of model (43) for $\pi = \kappa = 0$ and three different values for σ and the same using SBIC.

$\sigma = 1$

$\alpha = 0.05, \varrho = 1/2$

	200 observations				1000 observations			
	Correct variables	Too many variables	Too few variables	SBIC	Correct variables	Too many variables	Too few variables	SBIC
$\hat{h} = 0$	405	6	10	483	114	0	0	394
$\hat{h} = 1$	73	2	1	17	363	0	0	106
$\hat{h} = 2$	3	0	0	0	3	0	0	0
$\hat{h} > 2$	0	0	0	0	0	0	0	0

$\alpha = 0.10, \varrho = 1/2$

	200 observations				1000 observations			
	Correct variables	Too many variables	Too few variables	SBIC	Correct variables	Too many variables	Too few variables	SBIC
$\hat{h} = 0$	335	7	0	483	68	0	0	394
$\hat{h} = 1$	122	5	0	17	387	0	0	106
$\hat{h} = 2$	4	0	0	0	33	0	0	0
$\hat{h} > 2$	1	0	0	0	4	0	0	0

$\sigma = 0.5$

$\alpha = 0.05, \varrho = 1/2$

	200 observations				1000 observations			
	Correct variables	Too many variables	Too few variables	SBIC	Correct variables	Too many variables	Too few variables	SBIC
$\hat{h} = 0$	365	2	0	475	18	0	0	256
$\hat{h} = 1$	127	1	0	25	440	0	0	244
$\hat{h} = 2$	5	0	0	0	38	0	0	0
$\hat{h} > 2$	0	0	0	0	4	0	0	0

$\alpha = 0.10, \varrho = 1/2$

	200 observations				1000 observations			
	Correct variables	Too many variables	Too few variables	SBIC	Correct variables	Too many variables	Too few variables	SBIC
$\hat{h} = 0$	282	0	0	475	4	0	0	256
$\hat{h} = 1$	205	1	0	25	438	0	0	244
$\hat{h} = 2$	11	0	0	0	54	0	0	0
$\hat{h} > 2$	1	0	0	0	4	0	0	0

$\sigma = 0.125$

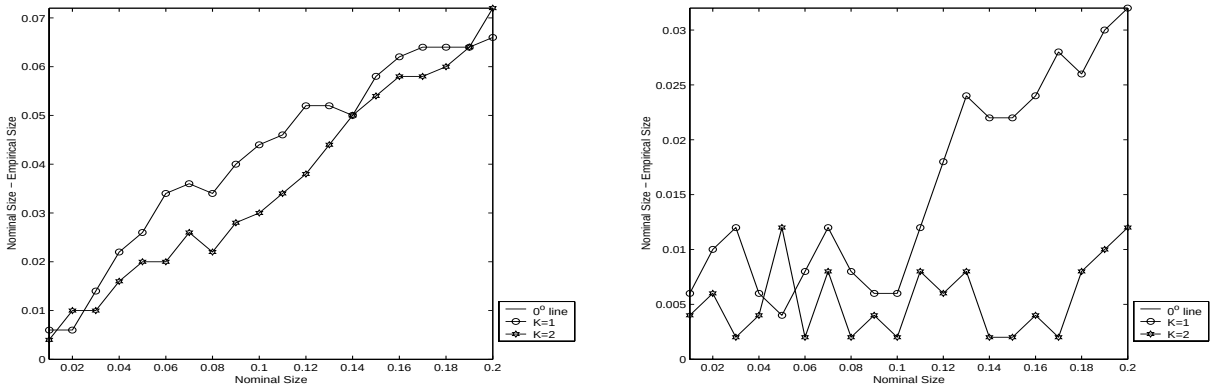
$\alpha = 0.05, \varrho = 1/2$

	200 observations				1000 observations			
	Correct variables	Too many variables	Too few variables	SBIC	Correct variables	Too many variables	Too few variables	SBIC
$\hat{h} = 0$	116	0	0	423	0	0	0	4
$\hat{h} = 1$	360	0	0	77	304	0	0	495
$\hat{h} = 2$	23	0	0	0	177	0	0	1
$\hat{h} > 2$	1	0	0	0	19	0	0	0

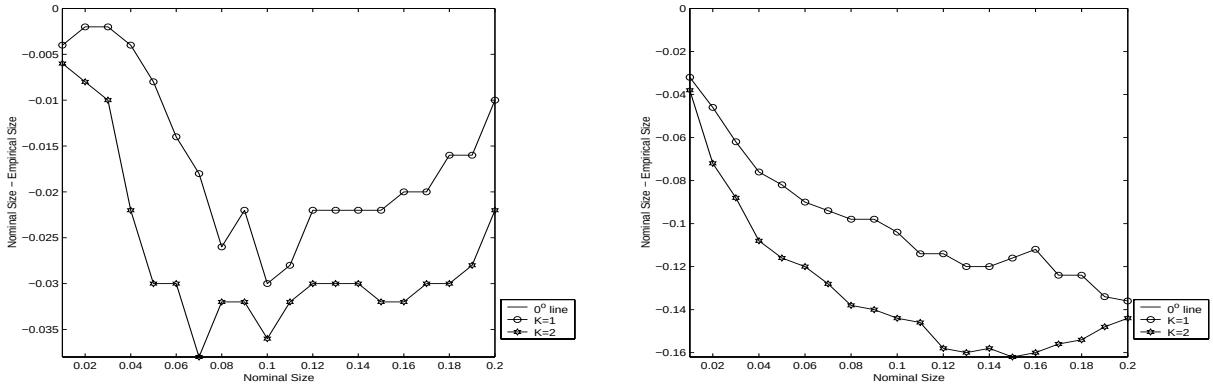
$\alpha = 0.10, \varrho = 1/2$

	200 observations				1000 observations			
	Correct variables	Too many variables	Too few variables	SBIC	Correct variables	Too many variables	Too few variables	SBIC
$\hat{h} = 0$	86	0	0	423	0	0	0	4
$\hat{h} = 1$	382	0	0	77	262	0	0	495
$\hat{h} = 2$	30	0	0	0	205	0	0	1
$\hat{h} > 2$	2	0	0	0	33	0	0	0

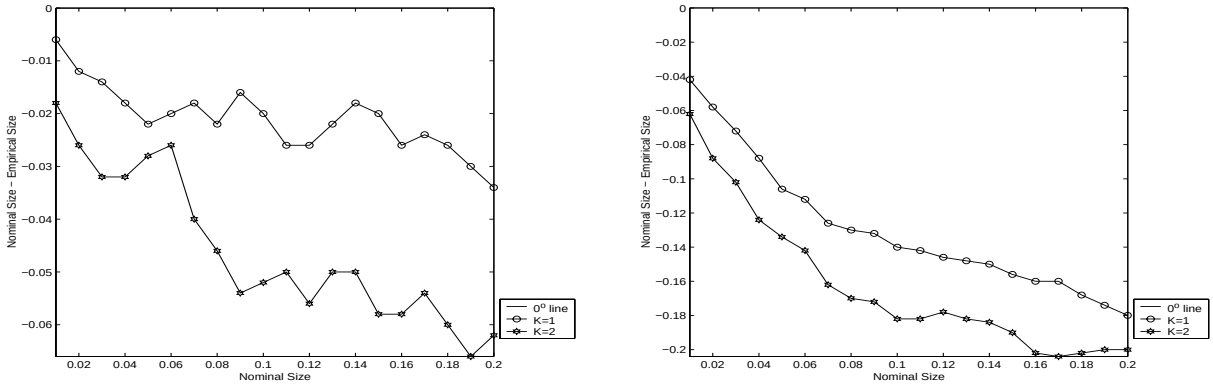
Notes: (a) The cases where the number of variables is correct but the combination is not the correct one appear under the heading "Too few variables". (b) The results concerning model selection using SBIC do not depend on the value of α.



(a) (b)



(c) (d)



(e) (f)

Figure 1: Size discrepancy curves of the simulated parameter constancy test. Panel (a): $\pi = 0$, $\kappa = 0$, and $\sigma = 0.25$. Panel (b): $\pi = 0$, $\kappa = 0$, and $\sigma = 0.125$. Panel (c): $\pi = 0$, $\kappa = 0.2$, and $\sigma = 0.25$. Panel (d): $\pi = 0$, $\kappa = 0.4$, and $\sigma = 0.25$. Panel (e): $\pi = 0$, $\kappa = 0.2$, and $\sigma = 0.125$. Panel (f): $\pi = 0$, $\kappa = 0.4$, and $\sigma = 0.125$.

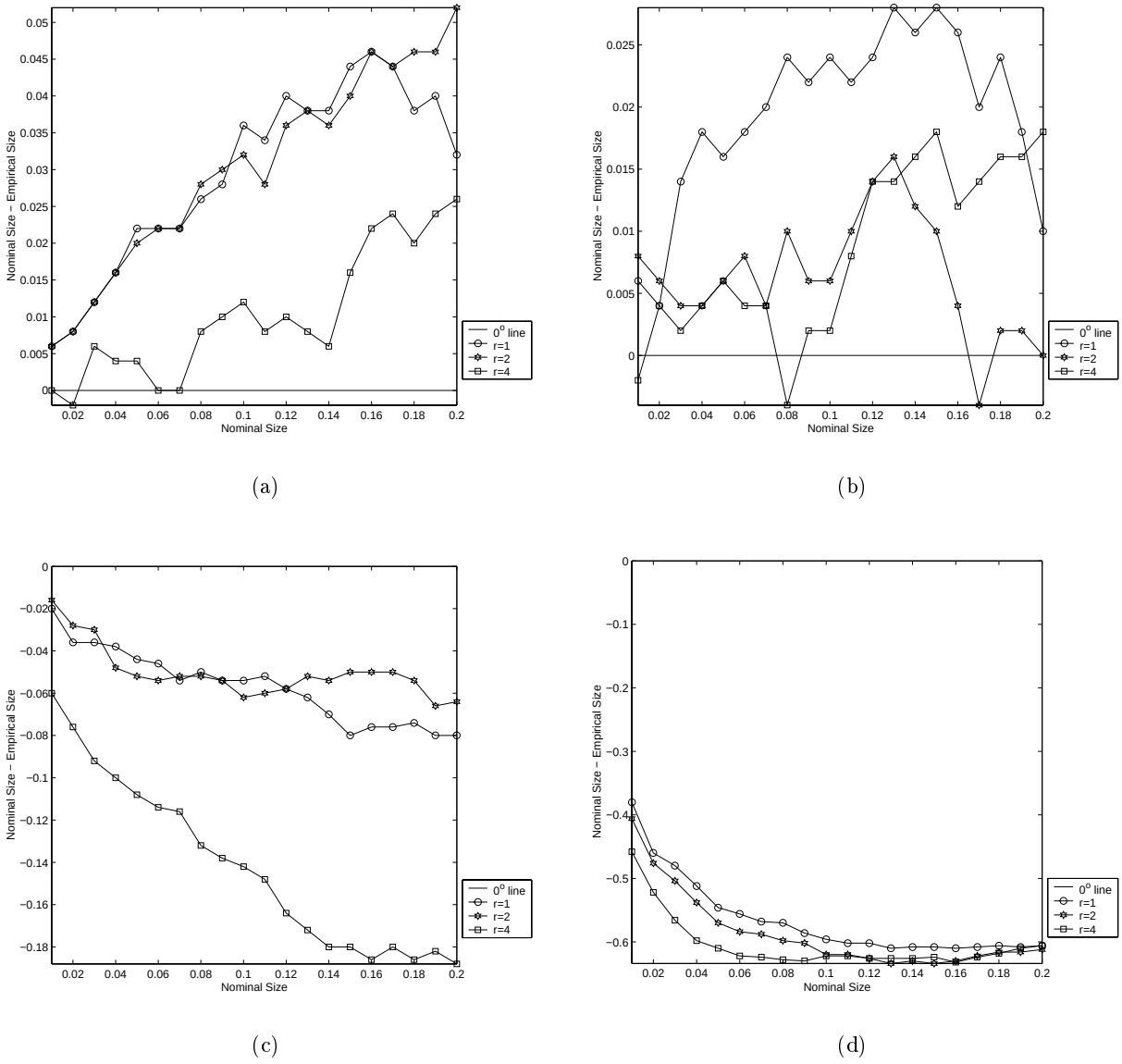


Figure 2: Size discrepancy curves of the no error autocorrelation test. Panel (a): $\pi = 0$, $\kappa = 0$, and $\sigma = 0.25$. Panel (b): $\pi = 0$, $\kappa = 0$, and $\sigma = 0.125$. Panel (c): $\pi = 1$, $\kappa = 0$, and $\sigma = 0.25$. Panel (d): $\pi = 2$, $\kappa = 0$, and $\sigma = 0.125$ in (43)

6 Case Studies

6.1 Example 1: Annual Sunspot Numbers, 1700–2000

In this section we illustrate our modelling strategy by two empirical examples. In the first example we build an AR-ANN model for the annual sunspot numbers over the period 1700–1979 and forecast with the estimated model up until the year 2001. The series, consisting of the years 1700–2001, was obtained from the National Geophysical Data Center web page. [1] The sunspot numbers are a heavily modelled nonlinear time series: for a neural network example see Weigend, Huberman and Rumelhart (1992). In this work we adopt the square-root transformation of Ghaddar and Tong (1981) and Tong (1990, p. 420). The transformed observations have the form $y_t = 2\left[\sqrt{(1+N_t)} - 1\right]$, $t = 1, \ldots, T$, where N_t is the original number of sunspots in the year t. The graph of the transformed series appears in Figure 3. Most of the published examples of fitting NN models to sunspot series deal with the original and not the square-root transformed series.

We use the observations for the period 1700–1979 to estimate the model and the remaining ones for a forecast evaluation. We begin the AR-NN modelling of the series by selecting the relevant lags using the variable selection procedure described in Section 4.2. We use a third-order polynomial approximation to the true model. Applying SBIC, lags 1,2, and 7 are selected whereas AIC yields the lags 1,2,4,5,6,7,8,9, and 10. We proceed with the lags selected by the SBIC. However, the residuals of the estimated linear AR model are strongly autocorrelated. The serial correlation is removed by also including y_{t-3} in the set of selected variables. When building the AR-NN model we select the input variables for each hidden unit separately using the specification test described in Section 4.4. Linearity is rejected at any reasonable significance level and the p-value of the linearity test minimized with lags 1 2, and 7 as input variables. The sequence of including hidden units is

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

discontinued after adding the second hidden unit, see Table 2, and the final estimated model is

$$
\begin{aligned}
y_t = &\underset{(0.83)}{-0.17} + \underset{(0.09)}{0.85}\, y_{t-1} + \underset{(0.12)}{0.14} y_{t-2} - \underset{(0.06)}{0.31}\, y_{t-3} + \underset{(0.05)}{0.08}\, y_{t-7} \\
&+ \underset{(7.18)}{12.80} \times F \left[\underset{(0.23)}{0.46} \left(\underset{(-)}{0.29} y_{t-1} - \underset{(0.83)}{0.87}\, y_{t-2} + \underset{(0.09)}{0.40}\, y_{t-7} - \underset{(0.05)}{6.68} \right) \right] \\
&+ \underset{(0.48)}{2.44} \times F \left[\underset{(8.45\times 10^3)}{1.17 \times 10^3} \left(\underset{(-)}{0.83} y_{t-1} - \underset{(0.12)}{0.53}\, y_{t-2} - \underset{(0.08)}{0.18}\, y_{t-7} + \underset{(7.18)}{0.38} \right) \right] + \hat{\varepsilon}_t.
\end{aligned} \tag{44}
$$

$$
\hat{\sigma} = 1.89 \quad \hat{\sigma}/\hat{\sigma}_L = 0.70 \quad R^2 = 0.89 \quad pLJB = 1.8 \times 10^{-7}
$$

$$
pARCH(1) = 0.94 \quad pARCH(2) = 0.75 \quad pARCH(3) = 0.90 \quad pARCH(4) = 0.44,
$$

where the figures in parentheses below the estimates are standard deviation estimates, $\hat{\sigma}$ is the residual standard deviation, $\hat{\sigma}_L$ is the residual standard deviation of the linear AR model, R^2 is the determination coefficient, $pLJB$ is the p-value of the Lomnicki-Jarque-Bera test of normality, and $pARCH(j)$, $j = 1, \ldots, 4$, is the p-value of the LM test of no ARCH against ARCH of order j. The estimated correlation matrix of the linear term and the output of the hidden units is

$$
\hat{\mathbf{\Sigma}} = \begin{pmatrix} 1 & -0.30 & 0.74 \\ -0.30 & 1 & -0.19 \\ 0.74 & -0.19 & 1 \end{pmatrix}. \tag{45}
$$

It is seen from (45) that there are no redundant hidden units in the model as none of the correlations is close to unity in absolute value. Figure 3 illuminates the contributions of the two hidden units to the explanation of y_t. The linear unit can only represent a symmetric cycle, so that the hidden units must handle the nonlinear part of the cyclical variation in the series. It is seen from Figure 3 that the first hidden unit is activated at the beginning of every upswing, and its values return to zero before the peak. The unit thus helps explain the very rapid recovery of the series following each trough. The second hidden unit is activated roughly when the series is obtaining values higher than its mean. It contributes to characterizing another asymmetry in the sunspot cycle: the peaks and the troughs have distinctly different shapes, peaks being rounder than troughs. The switches in the value of the hidden unit from zero to unity and back again are quite rapid (γ_2 large), which is the cause of the large standard deviation of the estimate of γ_2, see the discussion in Section 4.3.

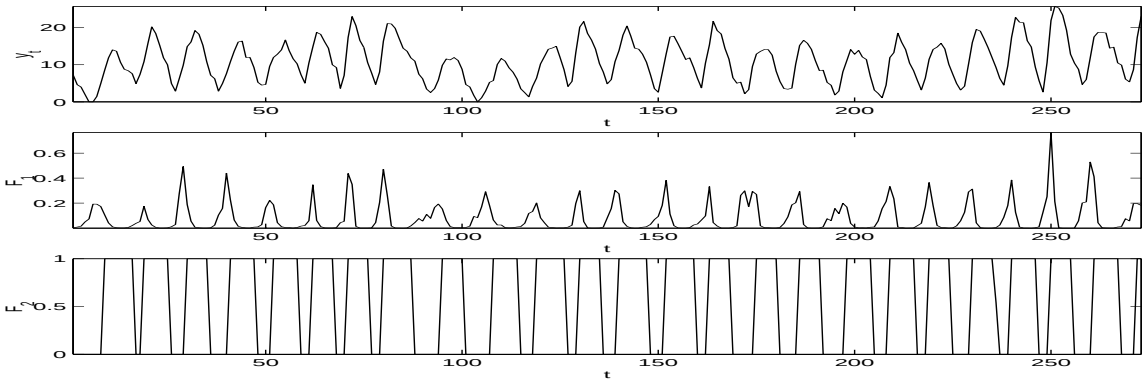


Figure 3: Panel (a): Transformed sunspot time series, 1700–1979. Panel (b): Output of the first hidden unit in (44). Panel (c): Output of the second hidden unit in (44).

Table 2: Test of no additional hidden units: minimum p-value of the set of tests against each null model.

	Number of hidden units under the null hypothesis		
	0	1	2
p-value	3×10^{-14}	2×10^{-9}	0.019

The results of the misspecification tests of model (44) in Table 3 indicate no model misspecification. In order to assess the out-of-sample performance of the estimated model we compare our forecasting results with the ones obtained from the two SETAR models, the one reported in Tong (1990, p. 420) and the other in Chen (1995), an artificial neural network (ANN) model with 10 hidden neurons and the first 9 lags as input variables, estimated with Bayesian regularization (MacKay 1992a, MacKay 1992b), and a linear autoregressive model with lags selected using SBIC. The SETAR model estimated by Chen (1995) is one in which the threshold variable is a nonlinear function of lagged values of the time series whereas it is a single lag in Tong´s model.

Table 4 shows the results of the one-step-ahead forecasting for the period 1980-2001. The

Table 3: Tests of no error autocorrelation and parameter constancy for model (44).

	LM Test for q-th order serial correlation						LM type test of parameter constancy			
	Lag						K			
	1	2	3	4	8	12	1	2	3	4
p-value	0.55	0.61	0.34	0.49	0.47	0.22	0.98	0.95	0.93	0.88

Table 4: One-step ahead forecasts, their root mean square errors, and mean absolute errors for the annual number of sunspots from a set of time series models, for the period 1980-2001.

		AR-NN		NN model		SETAR model (Tong 1990)		SETAR model (Chen 1995)		AR model	
Year	Observation	Forecast	Error	Forecast	Error	Forecast	Error	Forecast	Error	Forecast	Error
1980	154.6	153.4	1.2	136.9	17.7	161.0	-6.4	134.3	20.3	159.8	-5.2
1981	140.4	128.4	12.0	130.5	9.9	135.7	4.7	125.4	15.0	123.3	17.1
1982	115.9	95.8	20.1	101.1	14.8	98.2	17.7	99.3	16.6	99.6	16.3
1983	66.6	76.7	-10.1	88.6	-22.0	76.1	-9.5	85.0	-18.4	78.9	-12.3
1984	45.9	29.8	16.1	45.8	0.1	35.7	10.2	41.3	4.7	33.9	12.0
1985	17.9	21.9	-4.0	29.5	-11.6	24.3	-6.4	29.8	-11.9	29.3	-11.4
1986	13.4	13.5	-0.1	9.5	3.9	10.7	2.7	9.8	3.6	10.7	2.7
1987	29.4	23.7	5.7	25.2	4.2	20.1	9.3	16.5	12.9	23.0	6.4
1988	100.2	86.7	13.5	76.8	23.4	54.5	45.7	66.4	33.8	61.2	38.9
1989	157.6	161.6	-3.9	152.9	4.6	155.8	1.8	121.8	35.8	159.2	-1.6
1990	142.6	159.7	-17.1	147.3	-4.7	156.4	-13.8	152.5	-9.9	175.5	-32.9
1991	145.7	118.2	27.5	121.2	24.5	93.3	52.4	123.7	22.0	119.1	26.6
1992	94.3	98.1	-3.8	114.3	-20.0	110.5	-16.2	115.9	-21.7	118.9	-24.6
1993	54.6	64.8	-10.2	71.0	-16.4	67.9	-13.3	69.2	-14.6	57.9	-3.3
1994	29.9	21.0	8.9	32.9	-3.0	27.0	2.9	35.7	-5.8	29.9	-0.1
1995	17.5	14.9	2.6	19.2	-1.7	18.4	-0.9	18.9	-1.4	17.6	-0.1
1996	8.6	19.2	-10.6	10.2	-1.6	18.1	-9.5	11.6	-3.0	15.7	-7.1
1997	21.5	17.6	3.9	21.3	0.2	12.3	9.2	11.8	9.7	16.0	5.5
1998	64.3	64.6	-0.3	67.6	-3.3	46.7	17.6	58.5	5.8	52.5	11.8
1999	93.3	113.0	-19.7	105.2	-11.9	105.7	-12.5	122.7	-29.4	109.2	-15.9
2000	119.6	102.4	17.2	101.8	17.8	99.5	20.1	102.7	16.8	115.1	4.4
2001	111	102.9	8.1	112.5	-1.5	110.2	0.8	112.5	-1.5	121.0	-10
RMSE			12.2		12.8		18.1		17.3		15.9
MAE			9.9		9.9		12.9		14.3		12.1

results, summarized by the root mean squared error (RMSE) and mean absolute error (MAE) measures, are quite favourable for our AR-NN model. Turning away from the neural network models, the less than impressive performance of the SETAR models may raise questions about their feasibility. However, as Tong (1990, p. 421) has pointed out, these models are at their best in forecasting several years ahead because they are able to reproduce the distinct nonlinear structure of the sunspot series clearly better than the linear autoregressive models.

In order to find out whether or not model (44) generates more accurate one-step-ahead forecasts than the other models we have applied the modified Diebold-Mariano test (Diebold and Mariano 1995) of Harvey, Leybourne and Newbold (1997) to these series of forecasts. Table 5 shows the values of the statistic and the corresponding p-values. The null hypothesis of no difference in the theoretical MAE or RMSE between the AR-NN model and a competitor can be rejected only when the competitor is any of the SETAR models. The AR-NN model thus appears somewhat better than the SETAR alternatives but not better than the linear AR model and the NN one obtained by Bayesian regularization.

We also compared multi-step forecasts made by our model and the alternative models described above. The forecasts were made according to the following procedure.

Table 5: Modified Diebold-Mariano test of the null of no difference between the forecast errors of the different models.

Comparison	MDM Statistic	p-value
Squared Errors		
AR-NN vs. NN	0.41	0.34
AR-NN vs. SETAR (Tong 1990)	1.42	0.08
AR-NN vs. SETAR (Chen 1995)	1.89	0.04
AR-NN vs. AR	1.29	0.10
Absolute Errors		
AR-NN vs. NN	0.21	0.42
AR-NN vs. SETAR (Tong 1990)	1.52	0.07
AR-NN vs. SETAR (Chen 1995)	2.10	0.02
AR-NN vs. AR	1.10	0.15

1. For $t = 1980, \ldots, 1988$, compute the out-of-sample forecasts of one to 8-step-ahead of each model, $\hat{y}_t(k)$, and the associated forecast errors denoted by $\hat{\varepsilon}_t(k)$ where k is the forecasting horizon.

2. For each forecasting horizon, compute the RMSE and the MAE statistics.

Table 6 shows the root mean squared error and the mean absolute errors for the annual number of sunspots from a total of forecasts each made by each model for forecast horizons from 2 to 8 years. Several interesting facts emerge from the results. The forecastability of sunspots using the linear AR model deteriorates very slowly with the forecast horizon. This is clearly due to the extraordinarily persistent cycle in the series. As to the AR-NN model that also contains a linear unit, the advantage in forecast accuracy compared to the AR model is clear at short horizons but vanishes at the seven-year horizon. The large NN model obtained by Bayesian regularization does not contain a linear unit and fares less well in this comparison. For the two SETAR models, forecastability deteriorates quite slowly with the forecast horizon after a quick initial decay. The accuracy of forecasts, however, measured by the root mean squared error or the mean absolute error, is somewhat inferior to that of the linear AR model.

Table 6: Multi-step ahead forecasts, their root mean square errors, and mean absolute errors for the annual number of sunspots from a set of time series models, for the period 1981-2000.

Horizon	AR-NN		NN model		SETAR model (Tong 1990)		SETAR model (Chen 1995)		AR	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
2	18.4	14.6	20.7	16.7	31.6	21.0	27.2	21.6	26.5	18.8
3	21.6	14.6	24.3	19.3	38.4	25.2	33.6	24.8	28.2	19.9
4	22.2	15.6	27.3	21.6	42.2	26.4	31.8	23.6	27.8	20.2
5	22.4	14.0	32.4	23.2	42.2	27.0	30.6	21.6	26.9	19.1
6	20.6	14.0	36.5	25.3	41.6	26.4	31.9	23.0	26.8	19.7
7	27.5	18.4	42.2	30.2	43.3	30.3	34.0	25.0	27.5	19.8
8	25.1	20.0	39.6	30.1	45.2	35.0	33.8	26.0	26.7	19.6

6.2 Example 2: Financial Prediction

Our second example has to do with forecasting stock returns. We have chosen it because our results can be compared with ones from previous studies and because this is a multivariate example. Pesaran and Timmermann (1995) provided evidence in favour of monthly US stock returns being predictable. They constructed a linear model containing nine economic variables and showed that using the model for managing a portfolio consisting of either the S&P 500 index or bonds gave results superior to the ones obtained from a simple random walk model. The choice between stocks and bonds was reconsidered every month, and profits were reinvested. The time period extended from January 1954 to December 1992. Later, Qi (1999) applied a neural network model based on Bayesian regularization to the same data set and obtained results vastly superior to the ones Pesaran and Timmermann (1995) had reported. Recently, however, Maasoumi and Racine (2001) found that with no model could one come close to the level of accumulated wealth Qi's model generated even though some models had a similar in-sample performance. When Racine (2001) reproduced her experiment, he was unable to demonstrate similar results for the neural network model.

Following the others, we respecify our model for each observation period. Thus, our modelling strategy is applied as follows:

1. For $t = 1, \ldots, T$; with $T = 1960.1$ to 1992.12

   (a) Select the variables with the procedure described in Section 4.2 using a third-order Taylor expansion.

   (b) Test linearity with all the selected variables in the transition function using the heteroskedasticity-

robust version of the linearity test.

(c) If linearity is rejected, estimate an AR-NN model. Otherwise, estimate a linear regression including all the covariates. The number of hidden units is determined using the heteroskedasticity-robust version of the LM test. The initial significance level of the tests equals 0.05.

The first model is estimated with the data extending to the end of 1959, and the whole modelling procedure is repeated after adding another month to the sample. It is seen from Figure 4 that the composition of variables varies quite considerably, although there are periods of stability, such as the years 1978-1984, for example. (For a detailed description of the variables, see Pesaran and Timmermann (1995).) Not a single one of the nine variables appears in every model, however. Perhaps quite predictably when the sample is small, linearity is not rejected at the 5% level, see Figure 5. There is a single period between 1984 and 1988 when the model selection strategy yields a neural network (NN) model and another one at the end of the period. This already leads one to expect only minor differences in wealth at the end of the period between the strategies based on the linear and the NN model. In fact, the linear model containing all variables and the NN strategy (either a linear model with a subset of variables or an NN model) lead to a different investment decision in only 10 cases out of 396. Out of these 10, our technique yielded a correct direction forecast in four cases and the linear model in remaining six.

The accumulated wealth is shown in Table 7. The linear model gives the best results. Our NN model (Panel E) is slightly better than Bayesian regularization NN model of Racine (2001) (Panel D) for no or low transaction costs. For high transaction costs, the relationship is the opposite. Thus, our NN modelling strategy compares well with the Qi-Racine approach but is not any better than a linear model with a constant composition of variables. The main reason for the linear model doing well is that there is not much structure to be modelled in the relationship between the returns and the explanatory variables. A nonlinear model cannot therefore be expected to do better than a linear one. Furthermore, NN models most often require a large sample to perform well, and in this example a clear majority of samples must be considered small.

It has been pointed out, see for instance Fama (1998), that the accumulated wealth compar-

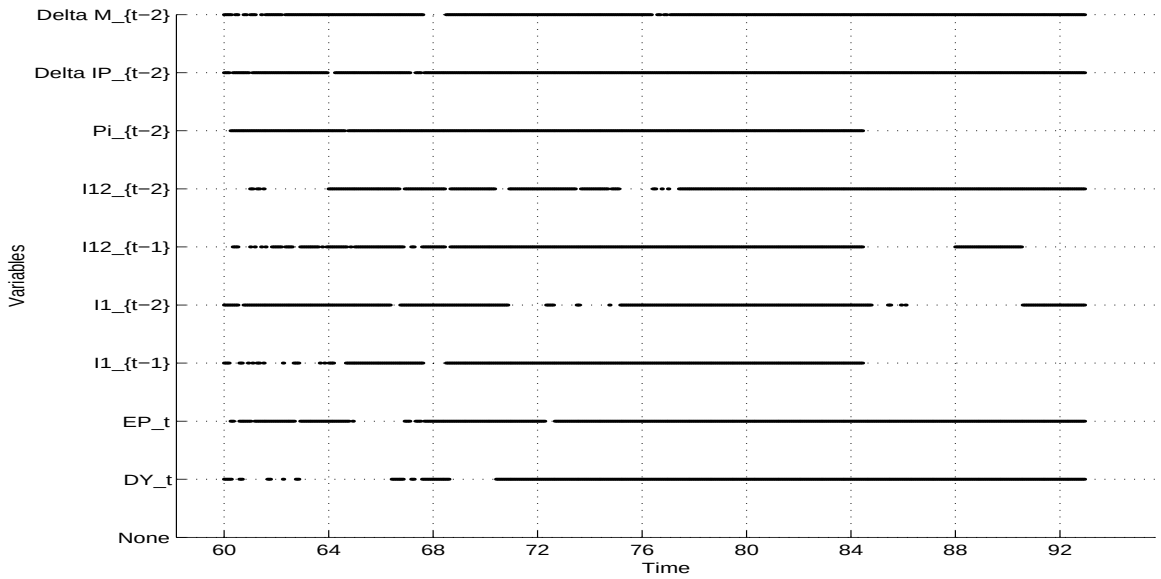


Figure 4: Variables selected using the AIC.

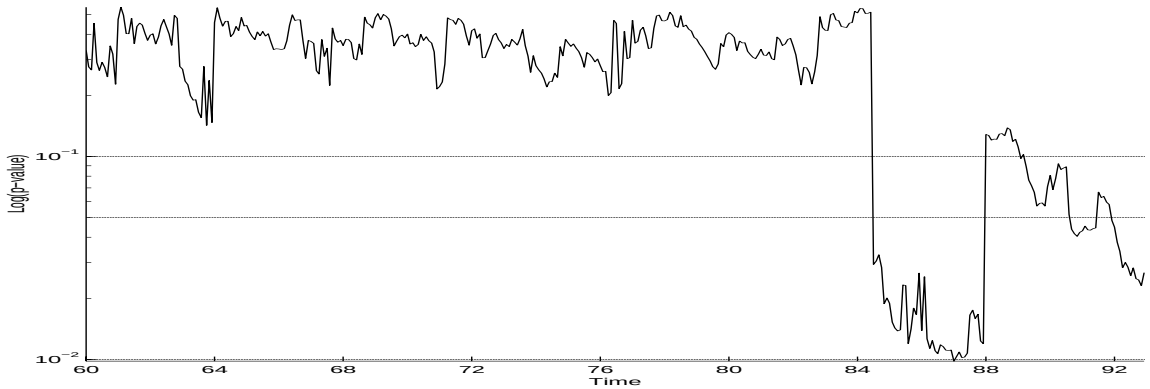


Figure 5: p-value of the linearity test (heteroskedasticity-robust version). The dashed lines are the 0.1, 0.05, and 0.01 bounds.

Table 7: Risks and profits of market, bond, and switching portfolios based on the out-of-sample forecasts of alternative models, 1960.1 to 1992.12.

Transaction costs	Mean return (%)	Std. of return	Sharpe ratio	Final wealth ($)
		Panel A: Market portfolio		
Zero	11.15	14.90	0.35	2,503
Low	11.13	14.90	0.43	2,463
High	11.11	14.89	0.43	2,424
		Panel B: Bond portfolio		
Zero	5.93	2.74	–	700
Low	4.72	2.74	–	471
High	4.72	2.74	–	471
		Panel C: Switching portfolio based on linear forecasts		
Zero	13.66	10.08	0.77	7,458
Low	12.21	10.18	0.74	4,631
High	11.23	10.34	0.63	3,346
		Panel D: Switching portfolio based NN forecasts of Racine (2001)		
Zero	13.23	10.89	0.67	6,624
Low	11.98	10.88	0.67	4,204
High	11.23	10.93	0.60	3,292
		Panel E: Switching portfolio based AR-NN forecasts		
Zero	13.50	10.12	0.75	7,054
Low	12.00	10.20	0.71	4,319
High	10.99	10.34	0.61	3,089

isons may be misleading in assessing the forecasting performance of different models because the cumulative effect of a single pair of different direction forecasts and thus investment decisions early on may grow quite large. In our case, the different decisions are few and appear relatively late in the sample. As a result, repeating the same exercise without reinvesting the profits leads to the conclusion that there is no difference in performance between the linear AR model and the models, either linear or NN, obtained by our technique.

7 Conclusion

In this paper we have demonstrated how statistical methods can be applied in building neural network models. The idea is to specify parsimonious models and keep the computational cost small. An advantage of our modelling strategy is that the modelling procedure is not a black box. Every step in model building is clearly documented and motivated. On the other hand, using this strategy requires active participation of the model builder and willingness to make decisions. Choosing the model selection criterion for variable selection and determining significance levels for the test sequence for selecting the number of hidden units are not automated, and different choices may often produce different models. Combining them in forecasting could be an interesting topic that, however, lies beyond the scope of this paper. Nevertheless, the method shows promise, and research is being carried out in order to learn more about its properties in modelling and forecasting stationary time series. The Matlab code for carrying out the modelling cycle exists and is downloadable at www.econ.puc-rio/br/mcm/nonlinear.html or www.hhs.se/stat/research/nonlinear.htm.

References

Abu-Mostafa, Y. S., Atiya, A. F., Magdon-Ismail, M. and White, H.: 2001, Introduction to the special issue on neural networks in financial engineering, *IEEE Transactions on Neural Networks* **12**, 653–655.

Anders, U. and Korn, O.: 1999, Model selection in neural networks, *Neural Networks* **12**, 309–323.

Auestad, B. and Tjøstheim, D.: 1990, Identification of nonlinear time series: First order characterization and order determination, *Biometrika* **77**, 669–687.

Balkin, S. D. and Ord, J. K.: 2000, Automatic neural network modeling for univariate time series, *International Journal of Forecasting* **16**, 509–515.

Bates, D. M. and Watts, D. G.: 1988, *Nonlinear Regression Analysis and its Applications*, Wiley, New York.

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

Bishop, C. M.: 1995, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford.

Chen, R.: 1995, Threshold variable selection in open-loop threshold autoregressive models, *Journal of Time Series Analysis* **16**, 461–481.

Chen, X., Racine, J. and Swanson, N. R.: 2001, Semiparametric ARX neural-network models with an application to forecasting inflation, *IEEE Transactions on Neural Networks* **12**, 674–683.

Clements, M. P. and Hendry, D. F.: 1999, *Forecasting Non-stationary Economic Time Series*, MIT Press, Cambridge, MA.

Cybenko, G.: 1989, Approximation by superposition of sigmoidal functions, *Mathematics of Control, Signals, and Systems* **2**, 303–314.

Davidson, R. and MacKinnon, J. G.: 1998, Graphical methods for investigating the size and power of hypothesis tests, *The Manchester School* **66**, 1–26.

Diebold, F. X. and Mariano, R. S.: 1995, Comparing predictive accuracy, *Journal of Business and Economic Statistics* **13**, 253–263.

Eitrheim, . and Teräsvirta, T.: 1996, Testing the adequacy of smooth transition autoregressive models, *Journal of Econometrics* **74**, 59–75.

Fama, E. F.: 1998, Market efficiency, long-term returns, and behavioral finance, *Journal of Financial Economics* **49**, 283–306.

Fine, T. L.: 1999, *Feedforward Neural Network Methodology*, Springer, New York.

Funahashi, K.: 1989, On the approximate realization of continuous mappings by neural networks, *Neural Networks* **2**, 183–192.

Gallant, A. R. and White, H.: 1992, On learning the derivatives of an unknown mapping with multilayer feedforward networks, *Neural Networks* **5**, 129–138.

Ghaddar, D. K. and Tong, H.: 1981, Data transformations and self-exciting threshold autoregression, *Journal of the Royal Statistical Society* **C30**, 238–248.

Godfrey, L. G.: 1988, *Misspecification Tests in Econometrics*, Vol. 16 of *Econometric Society Monographs*, second edn, Cambridge University Press, New York, NY.

Golub, G., Heath, M. and Wahba, G.: 1979, Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics* **21**, 215–223.

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

Harvey, D., Leybourne, S. and Newbold, P.: 1997, Testing the equality of prediction mean squared errors, *International Journal of Forecasting* **13**, 281–291.

Haykin, H.: 1999, *Neural Networks: A Comprehensive Foundation*, second edn, Prentice-Hall, Oxford.

Hwang, J. T. G. and Ding, A. A.: 1997, Prediction intervals for artificial neural networks, *Journal of the American Statistical Association* **92**, 109–125.

Kuan, C. M. and White, H.: 1994, Artificial neural networks: An econometric perspective, *Econometric Reviews* **13**, 1–91.

Kurková, V. and Kainen, P. C.: 1994, Functionally equivalent feedforward neural networks, *Neural Computation* **6**, 543–558.

Lee, T.-H., White, H. and Granger, C. W. J.: 1993, Testing for negleted nonlinearity in time series models. a comparison of neural network methods and alternative tests, *Journal of Econometrics* **56**, 269–290.

Leybourne, S., Newbold, P. and Vougas, D.: 1998, Unit roots and smooth transitions, *Journal of Time Series Analysis* **19**, 83–97.

Lin, C. F. J. and Teräsvirta, T.: 1994, Testing the constancy of regression parameters against continuous structural change, *Journal of Econometrics* **62**, 211–228.

Luukkonen, R., Saikkonen, P. and Teräsvirta, T.: 1988, Testing linearity in univariate time series models, *Scandinavian Journal of Statistics* **15**, 161–175.

Maasoumi, E. and Racine, J.: 2001, Entropy and predictability of stock market returns, *Journal of Econometrics* **107**, 291–312.

MacKay, D. J. C.: 1992a, Bayesian interpolation, *Neural Computation* **4**, 415–447.

MacKay, D. J. C.: 1992b, A practical Bayesian framework for backpropagation networks, *Neural Computation* **4**, 448–472.

Murata, N., Yoshizawa, S. and Amari, S.-I.: 1994, Network information criterion - determining the number of hidden units for an artificial neural network model, *IEEE Transactions on Neural Networks* **5**, 865–872.

Pesaran, M. H. and Timmermann, A.: 1995, Predictability of stock returns: robustness and econoic significance, *Journal of Finance* **50**, 1201–1228.

Qi, M.: 1999, Nonlinear predictability of stock returns using financial and economic variables, *Journal of Business and Economic Statistics* **17**, 419–429.

Racine, J.: 2001, On the nonlinear predictability of stock returns using financial and economic variables, *Journal of Business and Economic Statistics* **19**, 380–382.

Rech, G., Teräsvirta, T. and Tschernig, R.: 2001, A simple variable selection technique for nonlinear models, *Communications in Statistics, Theory and Methods* **30**, 1227–1241.

Reed, R. D. and Marks II, R. J.: 1999, *Neural Smithing*, MIT Press, Cambridge, MA.

Refenes, A. P. N. and Zapranis, A. D.: 1999, Neural model identification, variable selection, and model adequacy, *Journal of Forecasting* **18**, 299–332.

Ripley, B. D.: 1996, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge.

Schwarz, G.: 1978, Estimating the dimension of a model, *Annals of Statistics* **6**, 461–464.

Stone, M.: 1977, An asymptotic equivalence of choice of model by cross-validatiob and Akaike's criterion, *Journal of the Royal Statistical Society B* **39**, 44–47.

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 and Economic Statistics* **13**, 265–275.

Swanson, N. R. and White, H.: 1997a, Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models, *International Journal of Forecasting* **13**, 439–461.

Swanson, N. R. and White, H.: 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.

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

Teräsvirta, T. and Lin, C.-F. J.: 1993, Determining the number of hidden units in a single hidden-layer neural network model, *Research Report 1993/7*, Bank of Norway.

Teräsvirta, T. and Mellin, I.: 1986, Model selection criteria and model selection tests in regression models, *Scandinavian Journal of Statistics* **13**, 159–171.

Teräsvirta, T., Lin, C. F. and Granger, C. W. J.: 1993, Power of the neural network linearity test, *Journal of Time Series Analysis* **14**, 309–323.

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 University Press, Oxford.

Trapletti, A., Leisch, F. and Hornik, K.: 2000, Stationary and integrated autoregressive neural network processes, *Neural Computation* **12**, 2427–2450.

Tschernig, R. and Yang, L.: 2000, Nonparametric lag selection for time series, *Journal of Time Series Analysis* **21**, 457–487.

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* M. Casdagli and S. Eubank (eds), *Nonlinear Modeling and Forecasting*, Addison-Wesley.

White, H.: 1980, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* **48**, 817–838.

White, H.: 1989, An additional hidden unit test for negleted nonlinearity in multilayer feedforward networks, *Proceedings of the International Joint Conference on Neural Networks*, IEEE Press, New York, NY, pp. 451–455.

White, H.: 1990, Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings, *Neural Networks* **3**, 535–550.

Wooldridge, J. M.: 1990, A unified approach to robust, regression-based specification tests, *Econometric Theory* **6**, 17–43.

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

Zapranis, A. and Refenes, A.-P.: 1999, *Principles of Neural Model Identification, Selection and Adequacy: With Applications to Financial Econometrics*, Springer-Verlag, Berlin.

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