Journal of Forecasting
J. Forecast. **25**, 49–75 (2006)
Published online in Wiley InterScience
(www.interscience.wiley.com). **DOI**: 10.1002/for.974

Building Neural Network Models for Time Series: A Statistical Approach

MARCELO C. MEDEIROS,[1] TIMO TERÄSVIRTA[2]* AND GIANLUIGI RECH[2]

[1] *Department of Economics, Pontifical Catholic University of Rio de Janeiro, Brazil*

[2] *Department of Economic Statistics, Stockholm School of Economics, Sweden*

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 simple 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. All the tests are entirely based on auxiliary regressions and are easily implemented. 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.

KEY WORDS model misspecification; neural computing; nonlinear forecasting; nonlinear time series; smooth transition autoregression

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. 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.

* Correspondence to: Timo Teräsvirta, Department of Economic Statistics, Stockholm School of Economics, Box 6501, SE-113 83 Stockholm, Sweden. E-mail: timo.terasvirta@hhs.se

The artificial neural network (NN) 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 NN models in (macro)economics, and several financial applications appeared in a recent special issue of *IEEE Transactions on Neural Networks* (Abu-Mostafa *et al.*, 2001). The use of the NN model in applied work is generally motivated by a mathematical result stating that under mild regularity conditions, a relatively simple NN model is capable of approximating any Borel-measurable function to any given degree of accuracy; see, for example, Fine (1999) and the references therein. 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 NN literature (Fine, 1999). 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 single hidden-layer feedforward NN time series models. The 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 NN model as a statistical nonlinear model and apply statistical inference to the problem of specifying the model or, as the NN 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,b) 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 NN model. Refenes and Zapranis (1999) developed a computer-intensive strategy based on statistics to select the variables and the number of hidden units of the NN 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. The proposed methodology is based on auxiliary regressions and its implementation is straightforward. Moreover, selecting the right combination of variables and parameters is of great importance when forecasting from NN models is considered, as these models have a natural tendency to overfit. Usually, an over-parametrized NN model will have a very good performance in-sample but will perform poorly in an out-of-sample forecasting exercise. We shall consider the differences between our strategy and the others mentioned here in later sections of the paper.

The plan of the paper is as follows. The next section describes the model and a statistical interpretation of it. A model specification strategy, consisting of specification, estimation and evaluation of the model, is described in the third section. The results concerning a Monte-Carlo experiment are reported in the fourth section. Two applications with real data sets, as well as a review of other applications, are presented in the fifth section. A final section contains concluding remarks.

THE AUTOREGRESSIVE NEURAL NETWORK MODEL

The autoregressive neural network (AR-NN) model is defined as

DOI: 10.1002/for

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

where $G(\mathbf{x}_t; \psi)$ is a nonlinear function of the variables $\mathbf{x}_t$ with parameter vector $\psi \in \mathbb{R}^{(q+2)h+q+1}$ defined as $\psi = [\boldsymbol{\alpha}', \lambda_1, \ldots, \lambda_h, \tilde{\boldsymbol{\omega}}_1', \ldots, \tilde{\boldsymbol{\omega}}_h', \beta_1, \ldots, \beta_h]'$. 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 *et al.* (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.

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

$$y_t = \boldsymbol{\alpha}'\tilde{\mathbf{x}}_t + \sum_{i=1}^{h} \lambda_i F(\gamma_i(y_{t-d} - c_i)) + \varepsilon_t \tag{3}$$

where $\gamma_i = \tilde{\omega}_{1i}$ and $c_i = \beta_i/\tilde{\omega}_{1i}$. When $h = 1$, equation (3) defines a special case of an ordinary LSTAR model considered in Teräsvirta (1994). When $\gamma_i \to \infty$, $i = 1, \ldots, h$, model (3) becomes a self-exciting threshold autoregressive (SETAR) model with a switching intercept and $h + 1$ regimes.

An AR-NN model can thus be interpreted either as a semiparametric 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 $\beta_1 \leq \ldots \leq \beta_h$ or $\lambda_1 \geq \ldots \geq \lambda_h$. The second source of under-identification 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. The difficulty is dealt with by applying statistical inference in the model specification; see the next section. For further discussion of the identifiability of NN models see, for example, Hwang and Ding (1997) and the references therein.

DOI: 10.1002/for

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

$$F(\gamma_i(\boldsymbol{\omega}_i'\mathbf{x}_t - c_i)) = \left(1 + e^{-\gamma_i(\boldsymbol{\omega}_i'\mathbf{x}_t - c_i)}\right)^{-1} \tag{4}$$

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

$$\omega_{i1} = \sqrt{1 - \sum_{j=2}^{q} \omega_{ij}^2} > 0, \quad i = 1, \ldots, h \tag{5}$$

The parameter vector ψ of model (1) becomes $\psi = [\boldsymbol{\alpha}', \lambda_1, \ldots, \lambda_h, \gamma_1, \ldots, \gamma_h, \omega_{12}, \ldots, \omega_{1q}, \ldots, \omega_{h2}, \ldots, \omega_{hq}, c_1, \ldots, c_h]'$. In this case the first two identifying restrictions discussed above can be defined as, first, $c_1 \leq \ldots \leq c_h$ or $\lambda_1 \geq \ldots \geq \lambda_h$ and, second, $\gamma_i > 0$, $i = 1, \ldots, h$.

STRATEGY FOR BUILDING AR-NN MODELS

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 also suggest the use of in-sample misspecification tests for the purpose. Similar tests are routinely applied in evaluating STAR models (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 can be found in Trapletti *et al.* (2000).

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 *et al.* (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 $G(\mathbf{x}_t; \psi)$ in (1) by a general kth-order polynomial. By the Stone–Weierstrass theorem, the approxi-

DOI: 10.1002/for

mation can be made arbitrarily accurate if some mild conditions, such as the parameter space ψ being compact, are imposed on function $G(\mathbf{x}_t;\ \psi)$. Thus the AR-NN model, itself a universal approximator, is approximated by another function. This yields

$$G(\mathbf{x}_t;\psi) = \boldsymbol{\pi}'\tilde{\mathbf{x}}_t + \sum_{j_1=1}^{q}\sum_{j_2=j_1}^{q}\theta_{j_1 j_2}x_{j1,t}x_{j2,t} + \ldots + \sum_{j_1=1}^{q}\ldots\sum_{j_k=j_{k-1}}^{q}\theta_{j_1\ldots j_k}x_{j_1,t}\ldots x_{j_k,t} + R(\mathbf{x}_t;\psi) \tag{6}$$

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

In equation (6), 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 (6). In order to do that, we first regress y_t on all variables on the right-hand side of equation (6) assuming $R(\mathbf{x}_t;\ \psi)$ 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 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 $\Sigma_{i=1}^{q}\binom{q}{i} + 1$ linear models by ordinary least squares (OLS). Note that by following this procedure, the variables for the whole NN 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 applied successfully even in large samples when nonparametric model selection becomes computationally infeasible. Another alternative, using equation (6) as a starting point, is to apply the PcGets procedure; see Krolzig and Hendry (2001) for details.

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 previously on ε_t. 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 penalizing the log-likelihood, is avoided. Second, the starting values of the parameter estimates are carefully chosen, and we discuss the details of this later.

The AR-NN model is similar to many linear or nonlinear time series models in that the information matrix of the log-likelihood function is block diagonal such that we can concentrate the likelihood and first estimate the parameters of the conditional mean. Conditional maximum likelihood is thus equivalent to nonlinear least squares. The maximum likelihood estimator of the parameters of the conditional mean equals

DOI: 10.1002/for

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

where $q_t(\boldsymbol{\psi}) = (y_t - G(\mathbf{x}_t; \boldsymbol{\psi}))$.[2] Under standard regularity conditions, the maximum likelihood estimator $\tilde{\boldsymbol{\psi}}$ is almost surely consistent for $\boldsymbol{\psi}$ and

$$T^{1/2}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \rightarrow \mathrm{N}\left(\mathbf{0}, -\underset{T\rightarrow\infty}{\operatorname{plim}}\ \mathbf{A}(\boldsymbol{\psi})^{-1}\right) \tag{7}$$

where $\mathbf{A}(\boldsymbol{\psi}) = \dfrac{1}{\sigma^2 T}\dfrac{\partial^2 Q_T(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}$; see Trapletti *et al.* (2000).

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

$$\hat{\mathbf{B}}(\hat{\boldsymbol{\psi}}) = \left(\sum_{t=1}^{T}\hat{\mathbf{h}}_t\hat{\mathbf{h}}_t'\right)^{-1}\left(\sum_{t=1}^{T}\hat{\varepsilon}_t^2\hat{\mathbf{h}}_t\hat{\mathbf{h}}_t'\right)\left(\sum_{t=1}^{T}\hat{\mathbf{h}}_t\hat{\mathbf{h}}_t'\right)^{-1} \tag{8}$$

where $\hat{\mathbf{h}}_t = \dfrac{\partial_{qt}(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}$, 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.

Concentrated maximum likelihood

In order to reduce the computational burden we can apply concentrated maximum likelihood to estimate ψ as follows. Consider the ith iteration and rewrite model (1) as

$$\mathbf{y} = \mathbf{Z}(\boldsymbol{\phi})\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{9}$$

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

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

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

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{Z}(\boldsymbol{\phi})'\mathbf{Z}(\boldsymbol{\phi})\right)^{-1}\mathbf{Z}(\boldsymbol{\phi})'\mathbf{y} \tag{10}$$

The remaining parameters $\boldsymbol{\phi}$ are estimated conditionally on $\boldsymbol{\theta}$ by applying the Levenberg–Marquardt algorithm, which completes the ith iteration. This form of concentrated maximum likelihood was proposed by Leybourne *et al.* (1998) in the context of STAR models and it substantially reduces the

DOI: 10.1002/for

dimensionality of the iterative estimation problem. Numerical issues such as the choice of starting values are discussed in the Appendix.

Determining the number of hidden units

The number of hidden units included in a NN 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.

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 an 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) and Hansen (1996).

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. In carrying out the sequence of tests we simply assume that the smaller (null) model and thus the corresponding log-likelihood function are always correctly specified. This is a standard assumption when the null model is a linear model, but we retain it even when our null model is already an AR-NN model. Admittedly, from the point of view of application, neural network models are generally regarded as approximations to a true relationship between the variables involved; for discussion see, for example, Anders *et al.* (1998). Since the only aim of testing is to find a parsimonious but adequate AR-NN model to serve as an approximation to the unknown true relationship for forecasting, we do not consider the assumption of correct specification for this purpose to be harmful or misleading. Following Teräasvirta and Lin (1993) we derive a test that is repeated until the first nonrejection of the null hypothesis. Assume now that our AR-NN model (1) contains $h + 1$ hidden units and write it as follows:

$$y_t = \boldsymbol{\alpha}'\tilde{\mathbf{x}}_t + \sum_{i=1}^{h} \lambda_i F(\gamma_i(\boldsymbol{\omega}_i'\mathbf{x}_t - c_i)) + \lambda_{h+1} F(\gamma_{h+1}(\boldsymbol{\omega}_{h+1}'\mathbf{x}_t - c_{h+1})) + \varepsilon_t \tag{11}$$

DOI: 10.1002/for

Assume further that we have not rejected the hypothesis of model (11) containing h hidden units and want to test for the $(h+1)$th hidden unit. The appropriate null hypothesis is

$$\mathrm{H}_0: \gamma_{h+1} = 0 \tag{12}$$

whereas the alternative is $\mathrm{H}_1: \gamma_{h+1} > 0$. Under (12), the $(h+1)$th hidden unit is identically equal to a constant and merges with the intercept in the linear unit.

We assume that under (12) the parameters of (11) can be estimated consistently. Model (11) is only identified under the alternative, which means that the standard asymptotic inference is not available. This problem is circumvented as in Luukkonen *et al.* (1988) by expanding the $(h+1)$th hidden unit into a Taylor series around the null hypothesis (12). The order of expansion is a compromise between a small approximation error (high order) and availability of data (short time series necessarily imply a relatively low order). Using a third-order Taylor expansion, rearranging and merging terms results in the following model:

$$y_t = \boldsymbol{\pi}'\tilde{\mathbf{x}}_t + \sum_{i=1}^{h} \lambda_i F(\gamma_i(\boldsymbol{\omega}_i'\mathbf{x}_t - c_i)) + \sum_{i=1}^{q}\sum_{j=i}^{q} \theta_{ij} x_{i,t} x_{j,t} + \sum_{i=1}^{q}\sum_{j=i}^{q}\sum_{k=j}^{q} \theta_{ijk} x_{i,t} x_{j,t} x_{k,t} + \varepsilon_t^* \tag{13}$$

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 is $\mathrm{H}_0': \theta_{ij} = 0$, $i = 1, \ldots, q$; $j = i, \ldots, q$, $\theta_{ijk} = 0$, $i = 1, \ldots, q$; $j = i, \ldots, q$; $k = j, \ldots, q$. Note that under $\mathrm{H}_0: \varepsilon_t^* = \varepsilon_t$, so that the properties of the error process remain unchanged under the null hypothesis. It is important to stress that the use of a first-order Taylor expansion is not possible because it is linear in $\mathbf{x}_t$. The second-order approximation is not useful either because the logistic function is odd and thus its second derivative evaluated under the null is zero. Higher-order approximations will introduce many regressors, reducing the degrees of freedom, and will not improve the statistical properties of the test. Under $\mathrm{H}_0: \gamma_{h+1} = 0$ and standard regularity conditions, completed with the assumption $E|x_{t,i}|^{\delta} < \infty$, $i = 1, \ldots, q$, for some $\delta > 6$, the LM type statistic

$$LM = \frac{1}{\hat{\sigma}^2} \sum_{t=1}^{T} \hat{\varepsilon}_t \mathbf{v}_t' \left\{ \sum_{t=1}^{T} \hat{\mathbf{v}}_t \hat{\mathbf{v}}_t' - \sum_{t=1}^{T} \hat{\mathbf{v}}_t \hat{\mathbf{h}}_t' \left(\sum_{t=1}^{T} \hat{\mathbf{h}}_t \hat{\mathbf{h}}_t' \right)^{-1} \sum_{t=1}^{T} \hat{\mathbf{h}}_t \hat{\mathbf{v}}_t' \right\}^{-1} \sum_{t=1}^{T} \hat{\mathbf{v}}_t \hat{\varepsilon}_t \tag{14}$$

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

$$\hat{\mathbf{h}}_t = \left. \frac{\partial G(\mathrm{x}_t; \psi)}{\partial \psi'} \right|_{\psi = \hat{\psi}} = \left[\tilde{\mathbf{x}}_t', \hat{F}_1, \ldots, \hat{F}_h, \hat{\lambda}_1 \frac{\partial \hat{F}_1}{\partial \gamma_1}, \ldots, \hat{\lambda}_h \frac{\partial \hat{F}_h}{\partial \gamma_h}, \hat{\lambda}_1 \frac{\partial \hat{F}_1}{\partial \tilde{w}_{12}'}, \ldots, \hat{\lambda}_1 \frac{\partial \hat{F}_1}{\partial \tilde{w}_{1q}'}, \ldots, \right.$$

$$\left. \hat{\lambda}_h \frac{\partial \hat{F}_h}{\partial \tilde{w}_{h2}'}, \ldots, \hat{\lambda}_h \frac{\partial \hat{F}_h}{\partial \tilde{w}_{hq}'}, \hat{\lambda}_1 \frac{\partial \hat{F}_1}{\partial c_1}, \ldots, \hat{\lambda}_h \frac{\partial \hat{F}_h}{\partial c_h} \right]'$$

with $\hat{F}_i \equiv F(\hat{\gamma}_i(\hat{\boldsymbol{\omega}}_i'\mathbf{x}_t - \hat{c}_i))$ and

$$\frac{\partial \hat{F}_i}{\partial \gamma_i} = (\hat{\boldsymbol{\omega}}_i'\mathbf{x}_t - \hat{c}_i)[2\cosh(\hat{\gamma}_i(\hat{\boldsymbol{\omega}}_i'\mathbf{x}_t - \hat{c}_i))]^{-2}, \quad i = 1, \ldots, h$$

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

$$\frac{\partial \hat{F}_i}{\partial c_i} = -\gamma_i[2\cosh(\hat{\gamma}_i(\hat{\omega}_i'\mathbf{x}_t - \hat{c}_i))]^{-2}, \quad i = 1, \ldots, h$$

DOI: 10.1002/for

and $\boldsymbol{v}_t = [x_{1,t}^2, x_{1,t}x_{2,t}, \ldots, x_{i,t}x_{j,t}, \ldots, x_{1,t}^3, \ldots, x_{i,t}x_{j,t}x_{k,t}, \ldots, x_{h,t}^3]$, 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 is 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 $SSR_0 = \Sigma_{t=1}^{T} \tilde{\varepsilon}_t^2$. The new residuals $\tilde{\varepsilon}_t$ are orthogonal to $\hat{\mathbf{h}}_t$.
2. Regress $\tilde{\varepsilon}_t$ on $\hat{\mathbf{h}}_t$ and $\hat{\boldsymbol{v}}_t$. Compute the sum of squared residuals $SSR_1 = \Sigma_{t=1}^{T} \hat{v}_t^2$.
3. Compute the χ^2 statistic.

$$LM_{\chi^2}^{hn} = T\frac{SSR_0 - SSR_1}{SSR_0} \tag{15}$$

or the F version of the test

$$LM_F^{hn} = \frac{(SSR_0 - SSR_1)/m}{SSR_1/(T-n-m)} \tag{16}$$

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.

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 *et al*. (1993). A natural alternative to our procedure is the one first suggested in White (1989) and investigated later in Lee *et al.* (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}, $j = 1, \ldots, q$. This solves the identification problem as the extra neurons are observable, and the null hypothesis $\lambda_{h+1} = \ldots = \lambda_{h+q} = 0$ 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.

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

DOI: 10.1002/for

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

$$LM^r_{\chi^2} = T - SSR_1 \tag{17}$$

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 previously are just quasi-maximum likelihood estimates. Under mild regularity conditions they are still consistent and asymptotically normal.

Evaluation of the estimated model

After a model has been estimated it has to be evaluated. The specification test for determining the number of hidden units is also a misspecification test of the estimated model and thus an in-sample evaluation tool. We shall discuss another such test but before doing that we would like to emphasize the need for checking stationarity. The modelling strategy we propose applies to stationary variables. The estimated AR-NN model is not stationary when the lag polynomial of the linear component contains nonstationary roots. (Note that neural network models without a linear unit trivially satisfy the stationary condition.) This has an adverse effect on forecasting several periods ahead. Thus, the roots of the lag polynomial have to be computed, and nonstationary roots must lead to respecification of the model.

Test of no serial correlation

As already mentioned, the adequacy of the final model is tested by testing the hypothesis of no additional hidden units. In this subsection we discuss another misspecification test, one for testing the null hypothesis of no serial correlation in the errors. Rejecting this hypothesis suggests that the errors contain forecastable structure, which should not be the case in the AR-NN model intended for forecasting. The test is an application of the results in Eitrheim and Teräsvirta (1996) and Godfrey (1988, pp. 112–121) and may be performed after finding the number of hidden units. We assume that the errors in equation (1) follow an rth-order autoregressive process defined as

$$\varepsilon_t = \boldsymbol{\pi}'\boldsymbol{\nu}_t + u_t \tag{18}$$

where $\boldsymbol{\pi}' = [\pi_1, \ldots, \pi_r]$ is a parameter vector, $\boldsymbol{\nu}_t' = [\varepsilon_{t-1}, \ldots, \varepsilon_{t-r}]$, and $u_t \sim \text{NID}(0, \sigma^2)$. 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 (18) for observation t, given the fixed starting values, has the form

$$l_t = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left\{y_t - \sum_{j=1}^{r}\pi_j y_{t-j} - G(\mathbf{x}_t;\boldsymbol{\psi}) + \sum_{j=1}^{r}\pi_j G(\mathbf{x}_{t-j};\boldsymbol{\psi})\right\}^2 \tag{19}$$

DOI: 10.1002/for

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

$$\begin{aligned}\frac{\partial l_t}{\partial \pi_j} &= \left(\frac{u_t}{\sigma^2}\right)\{y_{t-j} - G(\mathbf{x}_{t-j};\boldsymbol{\psi})\},\ j = 1,\ldots,r \\ \frac{\partial l_t}{\partial \boldsymbol{\psi}} &= -\left(\frac{u_t}{\sigma^2}\right)\left\{\frac{\partial G(\mathbf{x}_t,\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} - \sum_{j=1}^{r}\pi_j \frac{\partial G(\mathbf{x}_{t-j};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right\}\end{aligned} \tag{20}$$

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

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

where $\hat{\boldsymbol{\nu}}_t' = [\hat{\varepsilon}_{t-1}, \ldots, \hat{\varepsilon}_{t-r}]$, $\hat{\varepsilon}_{t-j} = y_{t-j} - G(\mathbf{x}_{t-j};\ \hat{\boldsymbol{\psi}})$, $j = 1, \ldots, r$, $\hat{\mathbf{h}}_t = \dfrac{\partial G(\mathbf{x}_t;\hat{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}}$, and $\hat{\sigma}^2 = (1/T)\ \Sigma_{t=1}^{T}\hat{\varepsilon}_t^2$. The *LM* statistic is (14) 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.

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 earlier. After selecting the variables we select the number of hidden units sequentially. We begin testing linearity against a single hidden unit as described above at significance level α. The model under the null hypothesis is simply a linear AR 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 $\alpha\rho$, $0 < \rho < 1$. 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 $\alpha\rho^2$. The sequence is terminated at the first nonrejection of the null hypothesis. The significance level is reduced at each step of the sequence and converges to zero. In our applications later, we use $\rho = 1/2$. 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 $\rho = 1/2$ then $\alpha^* \leq 0.187$. Note that if, instead of our LM type test, we 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 vari-

DOI: 10.1002/for

ables. 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 extremely high. Because of this we suggest another technique that combines sequential testing for hidden units and variable selection. Consider equation (11). Instead of just testing a single null hypothesis as is done within (11), 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 $(h + 1)$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 $(h + 1)$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 earlier.

Compared to our first strategy, this one adds to the flexibility and on 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 and controlling stationarity. If the model does not pass the checks, the model builder has to reconsider the specification. 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. The results are of course 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 situation, see Clements and Hendry (1999, chapter 2). In our view, out-of-sample and in-sample evaluations of the estimated model are complementary rather than competing model evaluation techniques.

Discussion and comparisons

It is useful to compare our modelling strategy with other bottom-up approaches available in the literature. Swanson and White (1995, 1997a,b) apply a model selection criterion as follows. They start with a linear model, adding potential variables to it until SBIC indicates that the model cannot be further improved. Then they estimate models with a single hidden 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 above 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 (4) and (5).

DOI: 10.1002/for

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 *et al.* (1979). The model for which the value of this model selection criterion is minimized is selected.

Refenes and Zapranis (1999) propose adding hidden units into the model sequentially. The number of units, however, is selected only after adding all units up to a predetermined maximum number, so the procedure is not genuinely sequential. The choice is made by applying the network information criterion (Murata *et al*., 1994). 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, 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 above. 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.

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 earlier 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 our misspecification test is considered. In both experiments we make use of the following model:

$$\begin{aligned} y_t &= 0.10 + 0.75y_{t-1} - 0.05y_{t-4} \\ &\quad + 0.80F(2.24(0.45y_{t-1} - 0.89y_{t-4} + 0.09)) \\ &\quad - 0.70F(1.12(0.44y_{t-1} + 0.89y_{t-4} + 0.35)) + \varepsilon_t \\ \varepsilon_t &= \kappa\varepsilon_{t-1} + u_t, \quad u_t \sim NID(0, \sigma^2) \end{aligned} \tag{21}$$

DOI: 10.1002/for

In the first experiment, $\kappa = 0$. 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.

Architecture selection

Results from simulating the modelling strategy can be found in Table I. 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 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 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.

Serial correlation test

The test of no error autocorrelation is simulated using model (21) 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 1. Again, the test is somewhat conservative for $\sigma = 0.25$ and less so for $\sigma = 0.125$. The results of power simulations do not offer any surprise: the power increases with parameter κ. They are thus omitted to save space but are available upon request.

CASE STUDIES

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-NN 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 *et al.* (1992). In this work we adopt the square-root transformation of Ghaddar and Tong (1981) and Tong (1990, p. 420). The

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

DOI: 10.1002/for

Table I. Outcomes of the experiments of selecting the number of hidden units and the set of input variables

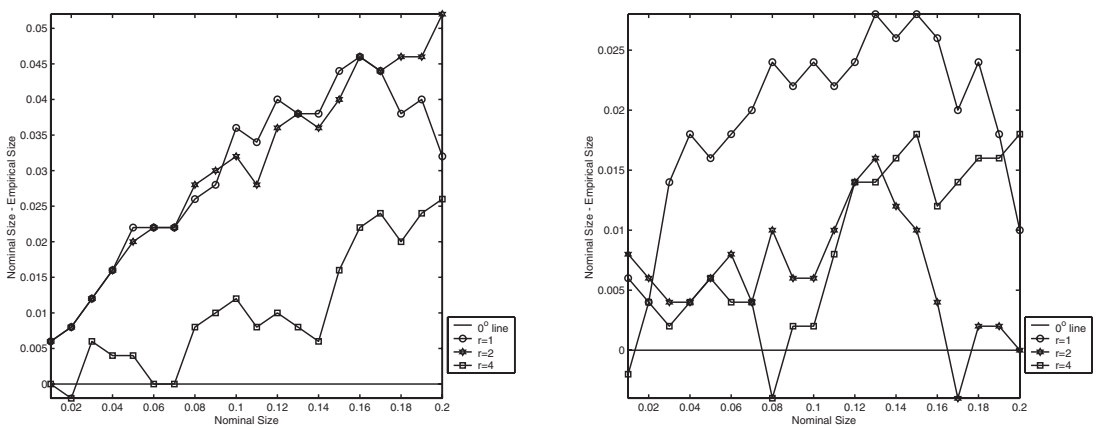
	$\sigma = 1$							
	$\alpha = 0.05$, $\rho = 1/2$							
	200 observations				1000 observations			
	Correct variables	Too many variables	Too few variables	SBIC	Correct variables	Too many variables	Too few variables	SBIC
$\hat{h} = 0$	405	6	10	483	114	0	0	394
$\hat{h} = 1$	73	2	1	17	363	0	0	106
$\hat{h} = 2$	3	0	0	0	3	0	0	0
$\hat{h} > 2$	0	0	0	0	0	0	0	0
	$\alpha = 0.10$, $\rho = 1/2$							
	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
	$\sigma = 0.5$							
	$\alpha = 0.05$, $\rho = 1/2$							
	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
	$\alpha = 0.10$, $\rho = 1/2$							
	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

DOI: 10.1002/for

Table I. *Continued*

	$\sigma = 0.125$							
	$\alpha = 0.05$, $\rho = 1/2$							
	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
	$\alpha = 0.10$, $\rho = 1/2$							
	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 test sequence starts at significance levels $\alpha = 0.05$ and 0.10 and sample sizes 200 and 1000 based on 500 replications of model (21) for $\kappa = 0$ and three different values for σ and the same using SBIC. (b) Table entries represent the number of times a given model is selected as the final specification. (c) $\hat{h}$ is the estimated number of hidden units. (d) The cases where the number of variables is correct but the combination is not the correct one appear under the heading 'Too few variables'. (e) The results concerning model selection using SBIC do not depend on the value of α.



(a) (b)

Figure 1. Size discrepancy curves of the no error autocorrelation test. Panel (a): $\kappa = 0$ and $\sigma = 0.25$. Panel (b): $\kappa = 0$ and $\sigma = 0.125$

DOI: 10.1002/for

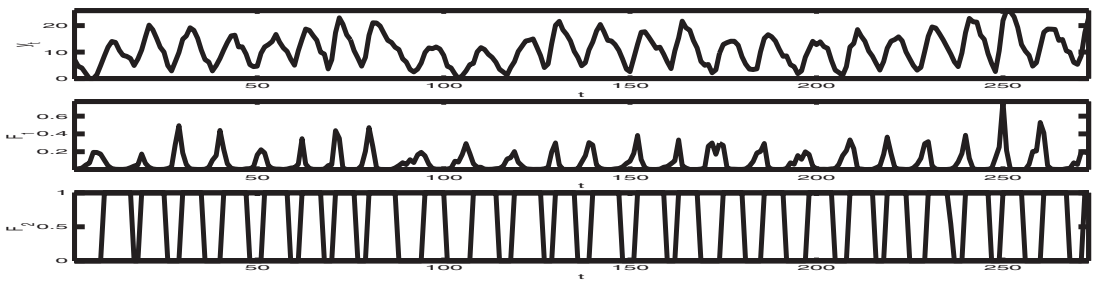


Figure 2. Panel (a): transformed sunspot time series, 1700–1979. Panel (b): output of the first hidden unit in (22). Panel (c): output of the second hidden unit in (22)

Table II. 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

transformed observations have the form $y_t = 2[\sqrt{(1+N_t)} - 1]$, $t = 1, \ldots, T$, where N_t is the original number of sunspots in the year t. The graph of the transformed series appears in Figure 2.

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 earlier. 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 previously. 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 discontinued after adding the second hidden unit, see Table II, and the final estimated model is

$$
\begin{aligned}
y_t = {} & \underset{(0.83)}{-0.17} + \underset{(0.09)}{0.85}\, y_{t-1} + \underset{(0.12)}{0.14}\, y_{t-2} - \underset{(0.06)}{0.31}\, y_{t-3} + \underset{(0.05)}{0.08}\, y_{t-7} \\
& + \underset{(7.18)}{12.80} \times F\left[\underset{(0.23)}{0.46}\left(\underset{(-)}{0.29}\, y_{t-1} - \underset{(0.83)}{0.87}\, y_{t-2} + \underset{(0.09)}{0.40}\, y_{t-7} - \underset{(0.05)}{6.68}\right)\right] \\
& + \underset{(0.48)}{2.44} \times F\left[\underset{(8.45 \times 10^3)}{1.17 \times 10^3}\left(\underset{(-)}{0.83}\, y_{t-1} - \underset{(0.12)}{0.53}\, y_{t-2} - \underset{(0.08)}{0.18}\, y_{t-7} + \underset{(7.18)}{0.38}\right)\right] + \hat{\varepsilon}_t \qquad (22) \\
& \hat{\sigma} = 1.89, \quad \hat{\sigma}/\hat{\sigma}_L = 0.70, \quad R^2 = 0.89, \quad pLJB = 1.8 \times 10^{-7} \\
& pARCH(1) = 0.94, \quad pARCH(2) = 0.75, \quad pARCH(3) = 0.90, \quad pARCH(4) = 0.44
\end{aligned}
$$

DOI: 10.1002/for

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

$$\hat{\mathbf{\Sigma}} = \begin{pmatrix} 1 & -0.30 & 0.74 \\ -0.30 & 1 & -0.19 \\ 0.74 & -0.19 & 1 \end{pmatrix} \tag{23}$$

It is seen from (23) that there are no redundant hidden units in the model as none of the correlations is close to unity in absolute value. Figure 2 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 2 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 above.

Table III shows the results of the test of no serial correlation described earlier. The results of the misspecification tests of model (22) indicate no model misspecification. Furthermore, checking the roots of the AR polynomial we can see that the estimated AR-NN model is stationary.

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), a NN model with 10 hidden neurons and the first 9 lags as input variables, estimated with Bayesian regularization (MacKay, 1992a,b), and a linear autoegressive 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 IV shows the results of the one-step-ahead forecasting for the period 1980–2001. The 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

Table III. Tests of no error autocorrelation in (22)

	LM test for qth-order serial correlation					
	Lag					
	1	2	3	4	8	12
p-value	0.55	0.61	0.34	0.49	0.47	0.22

DOI: 10.1002/for

Table IV. 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

Year	Observation	AR-NN		NN model		SETAR model (Tong, 1990)		SETAR model (Chen, 1995)		AR model	
		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	111.0	102.9	8.1	112.5	–1.5	110.2	0.8	112.5	–1.5	121.0	–10.0
RMSE			12.2		12.8		18.1		17.3		15.9
MAE			9.9		9.9		12.9		14.3		12.1

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 (22) 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 *et al*. (1997) to these series of forecasts. Table V 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:

(1) For $t = 1980, \ldots, 1988$, compute the out-of-sample forecasts of one to eight-step-ahead of each model, $\hat{y}_t(k)$, and the 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.

DOI: 10.1002/for

Table V. 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

Table VI. 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 1981–2001

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

Table VI 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 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 7-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.

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

DOI: 10.1002/for

for managing a portfolio consisting of either the S&P500 index or bonds gave results superior to the ones obtained from a simple random walk model. The variables are: excess returns on the S&P500 index, ρ_t; dividend yield, DY_t; earnings–price ratio; first lag of the one-month Treasury bill rate, $I1_{t-1}$; second lag of the one-month Treasury bill rate, $I1_{t-2}$; first lag of the 12-month Treasury bond rate, $I12_{t-1}$; second lag of the 12-month Treasury bond rate, $I12_{t-2}$; second lag of the year-on-year rate of inflation, π_{t-2}; second lag of the year-on-year rate of change in industrial production, ΔIP_{t-2}; and second lag of the year-on-year growth rate of narrow money stock, ΔM_{t-2}. 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 NN 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 the experiment, he was indeed unable to demonstrate similar results for the NN 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 earlier 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 3 that the composition of variables varies quite considerably, although there are periods of stability, such as the years 1978–1984, for example. 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 4. There is a single period between 1984 and 1988 when the model selection strategy yields a 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 a 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 the remaining six.

The accumulated wealth is shown in Table VII. The linear model gives the best results. Our NN model (Panel E) is slightly better than the 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

DOI: 10.1002/for

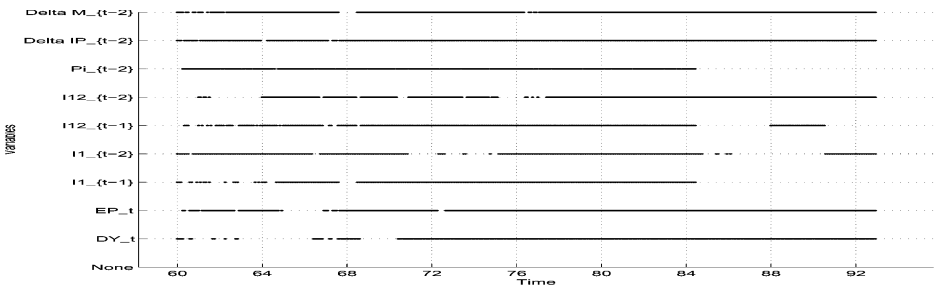


Figure 3. Variables selected using the AIC

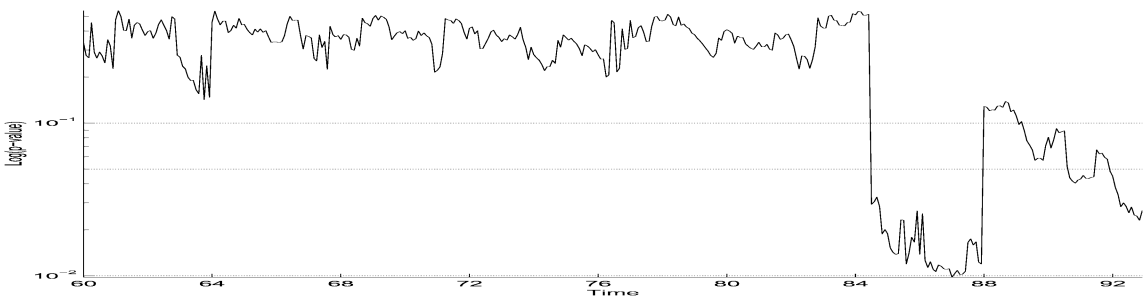


Figure 4. *p*-value of the linearity test (heteroskedasticity-robust version). The dashed lines are the 0.1, 0.05 and 0.01 bounds

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 Fama (1998), that the accumulated wealth comparisons 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 AR model and the models, either linear or NN, obtained by our technique.

Other examples

To complete the presentation, we briefly review other studies that have relevance in the present context. Anders and Korn (1999) contains a simulation study in which neural network models constructed using various strategies are evaluated out-of-sample. One of the strategies is based on Teräsvirta *et al*. (1993) and is therefore relevant for this paper. The results indicate that it is the one that on average generates the most accurate forecasts. In a related paper, Anders *et al*. (1998) apply neural network models to pricing call options. Their conclusion is that models specified using statistical inference have the best out-of-sample performance.

DOI: 10.1002/for

Table VII. 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	2503
Low	11.13	14.90	0.43	2463
High	11.11	14.89	0.43	2424
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	7458
Low	12.21	10.18	0.74	4631
High	11.23	10.34	0.63	3346
Panel D: switching portfolio based on NN forecasts of Racine (2001)				
Zero	13.23	10.89	0.67	6624
Low	11.98	10.88	0.67	4204
High	11.23	10.93	0.60	3292
Panel E: switching portfolio based on AR-NN forecasts				
Zero	13.50	10.12	0.75	7054
Low	12.00	10.20	0.71	4319
High	10.99	10.34	0.61	3089

Rech (2002) presents a forecasting exercise in which 30 time series from different fields such as macroeconomics, climatology and biology are being forecast using linear autoregressive models as well as various neural network models. The linear autoregressive model turns out to be the best performer overall. The modelling strategy described in the present paper performs equally well as the other, computationally more involved, strategies that include early stopping and various forms of pruning.

In a recent study, Teräsvirta *et al*. (2005) apply the modelling strategy described in this paper to 47 monthly macroeconomic time series from G7 countries. Another neural network modelling strategy they use is Bayesian regularization. It turns out that on average the latter strategy generates more accurate forecasts than the AR-NN approach, without dominating it. The authors report, however, that a number of the estimated AR-NN models are nonstationary, which has a negative effect on forecast accuracy at long forecast horizons. The models specified and estimated using Bayesian regularization do not have this drawback because either they do not have a linear unit or the estimates of parameters of the linear unit are shrunk towards zero. This outcome rather stresses the importance of model evaluation, which the authors, due to the vastness of the study (in total they estimated more than 900 models and generated forecasts for four forecast horizons from each of them), had to ignore. As a whole, the forecasting performance of the neural network models was somewhat inferior to that of linear autoregressive and logistic smooth transition autoregressive models also considered in that study. Combining all the information it appears that the AR-NN modelling strategy can be successfully used to produce effective forecasting models, but the role of model evaluation should not be forgotten.

DOI: 10.1002/for

CONCLUSIONS

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 down. An advantage of our modelling strategy is that the modelling procedure is transparent rather than 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. However, after these decisions have been made, the modelling strategy can easily be automated. As a whole, 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.hhs.se/stat/research/nonlinear.htm.

APPENDIX: STARTING VALUES AND COMPUTATIONAL ISSUES

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, γ_i; $i = 1, \ldots, h$; 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. This, together with the fact that $\|\boldsymbol{\omega}_h\| = 1$, gives us a basis for discussing the choice of starting values of γ_i, $i = 1, \ldots, h$. A natural choice of initial values for the estimation of parameters in the model with h neurons is to use the previous estimates for the parameters in the first $h - 1$ hidden units and the linear unit. The starting values for the parameters γ_h, $\boldsymbol{\theta}$, $\boldsymbol{\omega}_h$ and c_h 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 $c_h^{(k)} = \text{median}(\boldsymbol{\omega}_h^{(k)'} \mathbf{x})$, 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.
(3) For $k = 1, \ldots, K$ and $n = 1, \ldots, N$, estimate $\boldsymbol{\theta}$ using (10) and compute the value of $Q_T(\psi)$ for each combination of starting values. Choose those values of the parameters that minimize $Q_T(\psi)$.

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

Even from appropriate starting values it may sometimes 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 such that $\boldsymbol{\omega}_i'\mathbf{x}_t$ lies in a small neighbourhood of c_i. When the sample size is

DOI: 10.1002/for

not very large, there are generally few observations of this kind 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; see, for example, Bates and Watts (1988, p. 87) or Teräsvirta (1994).

ACKNOWLEDGEMENTS

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 5th Brazilian Conference on Neural Networks, Rio de Janeiro, April 2001, the ESF-EMM Network First Annual Meeting, Arona, September 2001, the 20th 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 University of Rio de Janeiro, and PUC-Rio. We wish to thank the participants of these occasions, Hal White in particular, for helpful comments. Our thanks also go to Chris Chatfield, Dick van Dijk and Marcelo Fernandes for useful remarks, and Allan Timmermann for the data used in the second empirical example of the paper and for fruitful suggestions. Comments from two anonymous referees have been helpful. The responsibility for any errors or shortcomings in the paper remains ours.

REFERENCES

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

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

Anders U, Korn O, Schmitt C. 1998. Improving the pricing of options: a neural network approach. *Journal of Forecasting* **17**: 369–388.

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

Balkin SD, Ord JK. 2000. Automatic neural network modeling for univariate time series. *International Journal of Forecasting* **16**: 509–515.

Bates DM, Watts DG. 1988. *Nonlinear Regression Analysis and its Applications.* John Wiley & Sons: New York.

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

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

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

Clements MP, Hendry DF. 1999. *Forecasting Non-stationary Economic Time Series.* MIT Press: Cambridge, MA.

Davies RB. 1977. Hypothesis testing when the nuisance parameter is present only under the alternative. *Biometrika* **64**: 247–254.

DOI: 10.1002/for

Daviese RB. 1987. Hypothesis testing when the nuisance parameter is present only under the alternative. *Biometrika* **73**: 33–44.
Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**: 253–263.
Eitrheim Ø, Teräsvirta T. 1996. Testing the adequacy of smooth transition autoregressive models. *Journal of Econometrics* **74**: 59–75.
Fama EF. 1998. Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics* **49**: 283–306.
Fine TL. 1999. *Feedforward Neural Network Methodology.* Springer: New York.
Ghaddar DK, Tong H. 1981. Data transformations and self-exciting threshold autoregression. *Journal of the Royal Statistical Society* **C30**: 238–248.
Godfrey LG. 1988. *Misspecification Tests in Econometrics.* Econometric Society Monographs, Vol. 16, 2nd edn. Cambridge University Press: New York.
Golub G, Heath M, Wahba G. 1979. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**: 215–223.
Hansen BE. 1996. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* **64**: 413–430.
Harvey D, Leybourne S, Newbold P. 1997. Testing the equality of prediction mean squared errors. *International Journal of Forecasting* **13**: 281–291.
Hwang JTG, Ding AA. 1997. Prediction intervals for artificial neural networks. *Journal of the American Statistical Association* **92**: 109–125.
Krolzig H-M, Hendry DF. 2001. Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics Control* **25**: 831–866.
Kuan CM, White H. 1994. Artificial neural networks: an econometric perspective. *Econometric Reviews* **13**: 1–91.
Lee T-H, White H, Granger CWJ. 1993. Testing for negleted nonlinearity in time series models. A comparison of neural network methods alternative tests. *Journal of Econometrics* **56**: 269–290.
Leybourne S, Newbold P, Vougas D. 1998. Unit roots and smooth transitions. *Journal of Time Series Analysis* **19**: 83–97.
Luukkonen R, Saikkonen P, Teräsvirta T. 1988. Testing linearity in univariate time series models, *Scandinavian Journal of Statistics* **15**: 161–175.
Maasoumi E, Racine J. 2001. Entropy and predictability of stock market returns. *Journal of Econometrics* **107**: 291–312.
MacKay DJC. 1992a. Bayesian interpolation. *Neural Computation* **4**: 415–447.
MacKay DJC. 1992b. A practical Bayesian framework for backpropagation networks. *Neural Computation* **4**: 448–472.
Murata N, Yoshizawa S, 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 MH, Timmermann A. 1995. Predictability of stock returns: robustness and economic 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 and 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. 2002. Forecasting with artificial neural network models. Working Paper Series in Economics and Finance 491, Stockholm School of Economics.
Rech G, Teräsvirta T, Tschernig R. 2001. A simple variable selection technique for nonlinear models. *Communications in Statistics, Theory and Methods* **30**: 1227–1241.
Refenes APN, Zapranis AD. 1999. Neural model identification, variable selection, and model adequacy. *Journal of Forecasting* **18**: 299–332.
Swanson NR, White H. 1995. A model selection approach to assessing the information in the term structure using linear models and artificial neural networks. *Journal of Business and Economic Statistics* **13**: 265–275.
Swanson NR, 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 NR, 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.

DOI: 10.1002/for

Teräsvirta T. 1994. Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association* **89**: 208–218.
Teräsvirta T, Lin, C-FJ. 1993. Determining the number of hidden units in a single hidden-layer neural network model. Research Report 1993/7, Bank of Norway.
Teräsvirta T, Mellin I. 1986. Model selection criteria and model selection tests in regression models. *Scandinavian Journal of Statistics* **13**: 159–171.
Teräsvirta T, Lin CF, Granger CWJ. 1993. Power of the neural network linearity test. *Journal of Time Series Analysis* **14**: 309–323.
Teräsvirta T, van Dijk D, Medeiros M. 2005. Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: a re-examination. *International Journal of Forecasting*, forthcoming.
Tjøstheim D, 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, Hornik K. 2000. Stationary and integrated autoregressive neural network processes. *Neural Computation* **12**: 2427–2450.
Tschernig R, 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, Rumelhart D. 1992. Predicting sunspots and exchange rates with connectionist networks. In *Nonlinear Modeling and Forecasting*, Casdagli M, Eubank S (eds). Addison-Wesley: Reading, MA.
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 neglected nonlinearity in multilayer feedforward networks. In *Proceedings of the International Joint Conference on Neural Networks*. IEEE Press: New York; 451–455.
Wooldridge JM. 1990. A unified approach to robust, regression-based specification tests. *Econometric Theory* **6**: 17–43.
Yao Q, Tong H. 1994. On subset selection in non-parametric stochastic regression. *Statistica Sinica* **4**: 51–70.
Zapranis A, Refenes A-P. 1999. *Principles of Neural Model Identification, Selection and Adequacy: With Applications to Financial Econometrics.* Springer-Verlag: Berlin.

Authors' biographies:

Marcelo C. Medeiros is Assistant Professor at the Department of Economics of the Pontifical Catholic University of Rio de Janeiro (PUC-Rio). His main research interest is nonlinear time series econometrics and the link between machine learning and econometric theory.

Timo Teräsvirta is Professor of Econometrics at the Stockholm School of Economics, Sweden. His main research interest is time series econometrics, nonlinear models and modelling in particular. He is a co-author of a book on nonlinear econometrics and has published a number of articles in international journals.

Gianluigi Rech obtained his PhD from the Stockholm School of Economics, Sweden in 2002. His main research interest is nonlinear time series econometrics.

Authors' addresses:

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

Timo Teräsvirta and **Gianluigi Rech**, Department of Economic Statistics, Stockholm School of Economics, Box 6501, SE-113 83, Stockholm, Sweden.

DOI: 10.1002/for