# Modelling and Forecasting Economic Time Series with Single Hidden-Layer Feedforward Autoregressive Artificial Neural Networks

# STOCKHOLM SCHOOL OF ECONOMICS
## EFI, THE ECONOMIC RESEARCH INSTITUTE

***EFI Mission***

EFI, the Economic Research Institute at the Stockholm School of Economics, is a scientific institution which works independently of economic, political and sectional interests. It conducts theoretical and empirical research in management and economic sciences, including selected related disciplines. The Institute encourages and assists in the publication and distribution of its research findings and is also involved in the doctoral education at the Stockholm School of Economics.

EFI selects its projects based on the need for theoretical or practical development of a research domain, on methodological interests, and on the generality of a problem.

***Research Organization***

The research activities are organized in twenty Research Centers within eight Research Areas. Center Directors are professors at the Stockholm School of Economics.

| | |
|---|---|
| *ORGANIZATION AND MANAGEMENT* | |
| Management and Organisation; (A) | Prof Sven-Erik Sjöstrand |
| Center for Ethics and Economics; (CEE) | Adj Prof Hans de Geer |
| Center for Entrepreneurship and Business Creation; (E) | Prof Carin Holmquist |
| Public Management; (F) | Prof Nils Brunsson |
| Information Management; (I) | Prof Mats Lundeberg |
| Center for People and Organization; (PMO) | Prof Jan Löwstedt |
| Center for Innovation and Operations Management; (T) | Prof Christer Karlsson |
| *ECONOMIC PSYCHOLOGY* | |
| Center for Risk Research; (CFR) | Prof Lennart Sjöberg |
| Economic Psychology; (P) | Prof Lennart Sjöberg |
| *MARKETING* | |
| Center for Information and Communication Research; (CIC) | Adj Prof Bertil Thorngren |
| Center for Consumer Marketing; (CCM) | Acting Prof Magnus Söderlund |
| Marketing, Distribution and Industrial Dynamics; (D) | Prof Lars-Gunnar Mattsson |
| *ACCOUNTING, CONTROL AND CORPORATE FINANCE* | |
| Accounting and Managerial Finance; (B) | Prof Lars Östman |
| Managerial Economics; (C) | Prof Peter Jennergren |
| *FINANCE* | |
| Finance; (FI) | Prof Clas Bergström |
| *ECONOMICS* | |
| Center for Health Economics; (CHE) | Prof Bengt Jönsson |
| International Economics and Geography; (IEG) | Prof Mats Lundahl |
| Economics; (S) | Prof Lars Bergman |
| *ECONOMICS STATISTICS* | |
| Economic Statistics; (ES) | Prof Anders Westlund |
| *LAW* | |
| Law; (RV) | Prof Erik Nerep |

*Chairman of the Board:* Prof Håkan Lindgren. *Director:* Associate Prof Bo Sellstedt.

***Adress***

EFI, Box 6501, S-113 83 Stockholm, Sweden • Internet: www.hhs.se/efi/
Telephone: +46(0)8-736 90 00 • Fax: +46(0)8-31 62 70 • E-mail efi@hhs.se

# Modelling and Forecasting Economic Time Series with Single Hidden-Layer Feedforward Autoregressive Artificial Neural Networks

by

Gianluigi Rech

STOCKHOLM SCHOOL OF ECONOMICS
EFI, THE ECONOMIC RESEARCH INSTITUTE

Dissertion for the Degree of Doctor of Philosophy, Ph.D in Econometrics, Stockholm School of Economics 2002


ISBN nr 91-7258-588-9

*Keywords*: neural networks
nonlinear time series
nonparametric variable selection
misspecification tests
parameter constancy
autocorrelation
Lagrange multiplier test
model specification
forecasting

Printed by
Xerox Printcenter
Handelshögskolan i Stockholm

Distributed by:
EFI, The Economic Research Institute
Stockholm School of Economics
P O Box 6501, SE 1-113 83 Stockholm, Sweden
www.hhs.se/efi

# Contents

# Preface

This thesis would not have been completed without the help of my supervisor, Prof. Timo Teräsvirta. He has proven himself full of interesting ideas, hints and endless patience. His interest in the field of time series is as strong as his endurance. His perseverance in leading me through the difficult path of scientific research has been priceless.

In addition to the support provided by my advisor, many people have contributed to the ideas developed in this dissertation. This is particularly true for my colleagues at the Stockholm School of Economics. They gave me useful comments during seminars and coffee breaks, widening my perspectives and giving me the opportunity to improve the quality of my work. I have really learnt a lot from all the members of the Department of Economic Statistics. In particular, I would like to thank Marcelo Medeiros for the great help he gave me in essay III, Sune Karlsson for always giving me a sensible comment on my work and priceless help with LATEX, and Changli He for his bits of chinese philosophy. I would also like to thank Rolf Tschernig for his help in essay I and during my visit to the Humboldt University of Berlin, and Allan Timmermann for providing me with the data set for the exogenous example in the second essay of the thesis. Monica Peijne, Carina Morton-Fincham and Pernilla Watson have all simplified my practical day-to-day problems, allowing me to concentrate on writing this dissertation. I am also indebted to Impact 92 and Gary Watson for his valuable help with checking the language.

I would also like to thank the publisher, Marcel Dekker, for giving me the permission to reprint essay I, which has been published in *Communications in Statistics, Theory and Methods*.

Financial support from the Tore Browaldh Foundation and the Stockholm School of Economics is gratefully acknowledged. I am very thankful to Sweden, which gave me the opportunity of writing my dissertation and of broadening my cultural horizons.

September 2001,

Gianluigi Rech

# Introduction

## 0.1 Background

Neural network models (hereafter ANN) are a class of nonlinear models which have been widely applied to different subjects. The idea comes from cognitive science, where they have been used to characterize processes which occur in the neural structure of the brain. In the latest years, an impressive number of publications has appeared claiming considerable successes in modelling time series, financial and high-frequency data in particular, by ANN modelling. The use of ANN models is based on a particularly interesting feature of ANN, the capability of approximating any Borel-measurable function to any degree of accuracy as Hornik, Stinchcombe, and White (1989), Cybenko (1989), Funahashi (1989) and others have shown. Statisticians have raised the question of whether ANN really perform better than other econometric models and can be considered more than "black boxes". Such skepticism is due to the lack of statistical procedures for model specification and evaluation. Their absence makes the assessment of the model's significance difficult and the interpretation of results very uncertain.

Some attempts have been made in this direction. Swanson and White (1995) estimate ANN models by a nonlinear stepwise method. After estimating the linear part of the ANN model, they fix the corresponding coefficients. Then a hidden unit is included, and regressors are selected one by one until the SBIC criterion (Rissanen (1978); Schwarz (1978)) cannot be improved. Then a second hidden unit is added and the process repeated. The procedure stops either when four hidden units are included in the model or the value of SBIC for the model with $q$ hidden units is greater than for $q-1$, $q \leq 4$. Anders and Korn (1999) stress the first problem the model builder faces, the selection of the relevant input variables to the model. For such an issue, they compare strategies based on hypothesis testing, information criteria and cross validation, but they do not suggest any way of testing the validity of the model. A more comprehensive approach can be found in Refenes and Zapranis (1999). The authors propose a three-step procedure, divided into model selection, variable significance testing and model adequacy testing. They suggest removing "irrelevant" variables from the model using confidence intervals, but they do not design a clear-cut strategy. As to model checking, they suggest analyzing residuals by any of the conventional tests like Ljung-Box and Durbin-Watson.

In this thesis, a new approach to ANN modelling is proposed, where the focus is on ANN as a statistical tool for time series analysis. Essay I considers the problem of the choice of the regressors to base the ANN model on. In essay II, the model is defined, its estimation procedure explained, and tests for its validity are developed. A simulation study and three applications on real data benchmark the model. Essay III contains a forecasting exercise on 30 time series, ranging on several fields, from economy to ecology. The approach developed in this paper is compared to linear modelling and to other three well-known neural network

modelling procedures: Information Criterion Pruning (ICP), Cross-validation pruning (CVP) and Bayesian regularization (BRP) .

### 0.1.1 Essay I: A simple variable selection technique for nonlinear models

In econometric modelling, the first issue is the choice of the regressors. Extensive literature has been devoted to the problem of variable selection in linear models. Parametric model selection criteria such as FPE (Akaike (1969)), AIC (Akaike (1974)), and SBIC are often employed. The alternative is the nonparametric approach based on kernel estimators. Cross-validation (Vieu (1995) and Yao and Tong (1994)) and nonparametric FPE criteria (Auestad and Tjøstheim (1990) and Tjøstheim and Auestad (1994)) are the most popular. As to the nonlinear case, the nonparametric approach can be directly utilized, whereas for the nonlinear case the parametric criteria require to narrow the choice of the nonlinear models to a certain class. For instance we can assume that our time series is generated by a smooth transition autoregressive (STAR) model. Whenever this is considered a restriction, either because the shape of the nonlinear function is unknown or for the large number of nonlinear models to be estimated if known, a different criterion is needed. If we decide to fit an ANN to the data, relying on its capability of approximating any Borel-measurable function as proved in Hornik, Stinchcombe, and White (1989), then reducing the dimension of the observation vector will prevent the researcher from performing an extensive comparison by estimating a large number of models for choosing the best one. Additionally, there does not exist an universally accepted criterion for comparing ANN models.

In this essay, I propose a simple variable selection technique based on the linearization of the regression function. The hypothesis of linearity is tested by a Lagrange multiplier test based on the Taylor expansion of the unknown function. If rejected, the Taylor expansion of order $k$ is utilized to estimate by ordinary least squares all the possible linearized models, and the combination of regressors leading to the lowest value of the model selection criterion is selected. In small samples, a less restrictive criterion as AIC is utilized in place of SBIC. In a simulation study, the performances of this method are compared to a refined version of the nonparametric FPE developed in Tschernig and Yang (2000). Results demonstrate that the selection technique is computationally much less demanding and performs well compared to FPE.

### 0.1.2 Essay II: Modelling and forecasting economic time series with single hidden-layer feedforward autoregressive artificial neural networks

In this essay, a unified framework for ANN modelling is proposed. It is based on the following model:

$$\begin{aligned} y_t &= \alpha_0 + \sum_{i=1}^{k} \alpha_i y_{t-i} + \sum_{j=1}^{q} \beta_j \psi(\widetilde{\boldsymbol{\gamma}}_j' \mathbf{w}_t) + u_t \qquad (1) \\ &= \boldsymbol{\alpha}' \mathbf{w}_t + \sum_{j=1}^{q} \beta_j \psi(\widetilde{\boldsymbol{\gamma}}_j' \mathbf{w}_t) + u_t, \ t = 1, ..., T. \end{aligned}$$

where $\{u_t\}$ is assumed to be a sequence of independent, normally distributed (n.i.d.) variables with mean zero, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_k, \alpha_0)'$, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_q)'$ are linear parameters, $\mathbf{w}_t = (w_{1t}, w_{2t}, ..., w_{kt}, 1)'$ is the $(k+1)\times 1$ vector of input variables and $\widetilde{\gamma}_j = (\widetilde{\gamma}_{j1}, \widetilde{\gamma}_{j2,...,} \widetilde{\gamma}_{j,k+1})'$ is

the $(k+1)\times 1$ vector of parameters or "weights" of the $j$th "hidden unit". The first difference with the ANN models usually employed in other modelling approaches is the presence of a linear part $\boldsymbol{\alpha}'\mathbf{w}_t$. The first step in my modelling prodedure is the selection of regressors $\mathbf{w}_t$. This is done by using the methodology developed in essay I, which performs at the same time linearity testing. The second step is the choice of the number of hidden units, $q$. The hypothesis of no hidden units (linear model) is tested at a given significance level $\alpha$. If rejected, a model with a linear part and one hidden unit is estimated. Then, the hypothesis of no additional hidden units is tested at the significance level $\alpha/2$. If rejected, a model with two hidden units is estimated. The procedure continues halving the significance level again to $\alpha/4$, $\alpha/8$, ..., stopping the procedure at the first acceptance of the null hypothesis of no additional hidden units. This procedure allows both to fit a linear model when nonlinearity is not present and, letting the significance level of the test for additional hidden units converge to zero as $q \to \infty$, keeps the dimension of the model under control.

Once estimated, model (1) needs to be evaluated. For such purpose, I developed specific tests for the hypothesis of no error autocorrelation and parameter constancy, while additional nonlinearity is already checked when I choose the number of hidden units. The tests are similar to the ones in Eitrheim and Teräsvirta (1996) and they are based on the Lagrange Multiplier approach.

A Monte-Carlo study gives an idea of the performances of the overall procedure. The performances of the procedure for specification, estimation and evaluation are analyzed in a simulation study carried out in samples of moderate and large sizes. The size distortion of the tests for no autocorrelation and parameter constancy is negligible, and the power good even in small samples.

Forecasting performances on actual series are also considered. In two classical benchmarks, the lynx and sunspot series, the modelling procedure is carried out in practice. Finally, following Qi (1999) and Racine (2001), the modelling procedure is also applied to a series with exogenous regressors. Qi and Racine extend the analysis of the monthly S&P 500 index return, from January 1954 to December 1992, already analyzed by Pesaran and Timmermann in (Pesaran and Timmermann (1995)). I apply my modelling procedure to this data set and compare my results to the ones obtained in the 2 aforementioned papers, obtaining similar findings to those of Racine.

### 0.1.3 Essay III: Forecasting with artificial neural network models

In the last essay, the methodology developed in essay II and the most popular methods for ANN estimation are compared by forecasting performances on 30 time series of different types, at 3 different forecasting horizons h = 1, 6 and 12 steps ahead. The idea comes from Swanson and White (1995). In Swanson and White (1997), the authors perform a comparative study between ANN and several other models (a random walk, a random walk with drift and a linear autoregressive model with exogenous regressors) by forecasting performances. In Zhang, Patuwo, and Hu (1998), a comprehensive review of the forecasting application in the ANN literature is carried out. In this work, attention is focused more on comparison among performances coming from different estimation methods than performances of diverse models. A total of 30 series is analyzed, from different subject areas. Twelve of them are economic time series: macroeconomic monthly data as US money demand and financial data as exchange rates. The remaining ones are borrowed from ecology: riverflow, temperature, ozone concentration, etc., including the two well-known case studies analyzed in Tong (1990), the lynx and sunspot series. Early stopping, pruning, information criterion pruning,

cross-validation pruning, interactive pruning, regularization, weight decay, and Bayesian regularization are the approaches I have considered here. The methodology developed in essay II is compared with them. The AR approximation to Box & Jenkins' ARMA modelling is another benchmark, because it is of interest to see whether or not neural network models actually produce forecasts superior to the ones obtained by linear models. The findings are that 1) the linear models outperform the ANN models and 2) albeit selecting and estimating much more parsimonious models, the ANN approach developed in this thesis stands up well with in comparison to other more sophisticated ANN models.

### 0.1.4 Final Remarks

In essay I, the part related to nonparametric statistics (theory and simulations) has been carried out by Rolf Tschernig. In essay III, the programs which produced forecasting results and comparisons for the linear, ICP, CVP and BRP models have been developed by Marcelo Medeiros.

# Bibliography


AKAIKE, H. (1969): "Fitting Autoregressive Models for Prediction," *Annals of the Institute of Statistical Mathematics*, 21, 243–47.

——— (1974): "A New Look at the Statistical Model Identification," *IEEE Transactions of Automatic Control*, AC-19, 716–23.

ANDERS, U., AND O. KORN (1999): "Model Selection in Neural Networks," *Neural Networks*, 12, 309–323.

AUESTAD, B., AND D. TJØSTHEIM (1990): "Identification of Nonlinear Time Series: First Order Characterization and Order Determination," *Biometrika*, 77, 669–87.

CYBENKO, G. (1989): "Approximations by Superimpositions of a Sigmoidal Function," *Mathematics of Control, Signals, and Systems*, 9, 2–54.

EITRHEIM, Ø., AND T. TERÄSVIRTA (1996): "Testing the Adequacy of Smooth Transition Autoregressive Models," *Journal of Econometrics*, 74, 59–75.

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

HORNIK, K., M. STINCHCOMBE, AND H. WHITE (1989): "Multi-Layer Feedforward Networks are Universal Approximators," *Neural Networks*, 2, 359–66.

PESARAN, M. H., AND A. TIMMERMANN (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 & Economic Statistics*, 17, 419–429.

RACINE, J. (2001): "On the Nonlinear Predictability of Stock Returns Using Financial and Economic Variables," *Journal of Business & Economic Statistics*, 19, 380–382.

REFENES, A.-P. N., AND A. D. ZAPRANIS (1999): "Neural Model Identification, Variable Selection and Model Adequacy," *Journal of Forecasting*, 18, 299–332.

RISSANEN, J. (1978): "Modeling by Shortest Data Description," *Automatica*, 14, 465–471.

SCHWARZ, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–64.

SWANSON, N. R., AND H. WHITE (1995): "A Model-Selection Approach to Assessing the Information in the Term Structure Using Linear Models and Artificial Neural Networks," *Journal of Business & Economic Statistics*, 13, 265–275.

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

TJØSTHEIM, D., AND B. AUESTAD (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.

TSCHERNIG, R., AND L. YANG (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.

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

ZHANG, G., B. E. PATUWO, AND M. Y. HU (1998): "Forecasting with Artificial Neural Networks: The State of the Art," *International Journal of Forecasting*, 14, 35–62.

# Essay 1

# A simple variable selection technique for nonlinear models

## 1.1 Introduction

Selecting a subset of variables is a problem that has been extensively considered in linear models. The problem often occurs in connection with autoregressive models. In that case, the variables have an ordering, and sequential tests may be applied to choosing the maximum lag if it is finite. If one is also interested in finding the relevant lags, model selection criteria such as FPE (Akaike (1969)), AIC (Akaike (1974)), SBIC (Rissanen (1978); Schwarz (1978)) and many others may be applied.

The variable selection problem also occurs in nonlinear models. In some situations, the functional form of the model may be unknown. The problem of finding the right subset of variables if it exists is then very important. This is because selecting too small a subset leads to misspecification whereas choosing too many variables aggravates the "curse of dimensionality". One way of solving the problem has been to use nonparametric methods based on local estimators. For kernel estimators, Vieu (1995) and Yao and Tong (1994) considered variable selection based on cross-validation. On the other hand, Auestad and Tjøstheim (1990) and Tjøstheim and Auestad (1994) suggested nonparametric FPE criteria. Their technique was further refined by Tschernig and Yang (2000) who, among other things, also showed consistency of nonparametric FPE.

Even when the nonlinear model is parametric, a nonparametric variable selection technique could still be useful in many situations. For instance, if the researcher intends to fit a neural network model to a set of data, then reducing the dimension of the observation vector before actually fitting any model to data is advisable, if at all possible. Nonparametric variable selection would save the researcher from the effort of estimating a possibly large number of neural network models with different combinations of variables before choosing the final one.

In this paper we propose a simple variable selection technique that instead of local estimation uses global parametric least squares estimation. This technique has several advantages. First, it is very simple and only requires OLS regression. This makes it computationally quick and easy. For the same reason the technique is applicable in large samples where the computational burden of nonparametric techniques becomes prohibitive. In simulations it is found to work very well in that situation. In small samples, its performance compares favourably with its nonparametric competitors, although there exist functional forms for

which the nonparametric technique is found to be superior. Finally, our technique also allows to check possible linearity of the function. In any case, even researchers who prefer nonparametric techniques may easily run our variable selection procedure as a check. If the results are quite different, then that may reveal something about the functional form of the true relationship. This is useful if nonparametric model selection is used as a first step towards building parametric models.

The plan of the paper is as follows. The idea and the theoretical motivation are presented in Section 2. The model selection procedure is outlined in Section 3. Results from a small-sample simulation study appear in Section 4, and our conclusions can be found in Section 5.

## 1.2 Global approximation of the nonlinear function

We assume that the data may be adequately described by the following nonlinear model

$$y_t = f(\mathbf{u}_t;\theta) + \varepsilon_t, \quad t = 1, ..., T \tag{1.1}$$

with parameter vector $\theta \in \Theta$ and the vector of regressors $\mathbf{u}_t$ such that $E[\varepsilon_t \mid \mathcal{F}_t] = 0$, where $\mathcal{F}_t = \{\mathbf{u}_t, \mathbf{u}_{t-1}, ...\}$ is the information set available at time $t$. We further allow $\mathbf{u}_t$ to include $q \geq 0$ redundant variables $\mathbf{z}_t = (z_{t1}, ..., z_{tq})'$ in addition to the correct variables $\mathbf{x}_t = (x_{t1}, ..., x_{tp})'$, $p \geq 1$. One therefore has $\mathbf{u}_t = (\mathbf{x}_t', \mathbf{z}_t')'$. Variable selection requires to identify $\mathbf{x}_t$ or to eliminate $\mathbf{z}_t$ without knowledge of the functional or parametric class of the conditional mean function $f(\cdot)$.

The basic idea of the variable selection procedures suggested in this paper is to conduct the selection on a parametric function $g(\cdot)$ which can approximate the true function $f(\cdot)$ well but is much simpler to estimate. A well-known class of simple approximating functions are series expansions

$$g(\mathbf{w};\beta) = \sum_{i=1}^{L} \beta_i g_i(\mathbf{w}_i)$$

with parameters $\beta_i$, known basis functions $g_i(\cdot)$ and $\mathbf{w}_i$ being in general subvectors of $\mathbf{w}$. Due to the linearity one can estimate the parameters $\beta_i$, $i = 1, \ldots, L$ by ordinary least squares. Of course, the quality of approximation depends on the choice of the basis functions $g_i(\cdot)$ and the length of the expansion $L$. Our choice of approximating function may be viewed as a member of the above class. In order to define $g_i(\cdot)$, assume that the sample space $\mathcal{U}$ is compact and that $f(\mathbf{u}_t;\theta)$ is continuous in $\mathcal{U}$. Then it follows from the Stone-Weierstrass theorem that $f(\mathbf{u}_t;\theta)$ can be uniformly approximated by a polynomial in the components of $\mathbf{u}_t$, see Royden (1963), pp. 150-151. Thus, using a general $k$th-order polynomial we can write:

$$
\begin{aligned}
f(\mathbf{x}_t, \mathbf{z}_t; \theta) &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{tj} + \sum_{j=1}^{q} \gamma_j z_{tj} + \sum_{j_1=1}^{p} \sum_{j_2=j_1}^{p} \beta_{j_1 j_2} x_{tj_1} x_{tj_2} \\
&\quad + \sum_{j_1=1}^{q} \sum_{j_2=j_1}^{q} \gamma_{j_1 j_2} z_{tj_1} z_{tj_2} + \sum_{j_1=1}^{p} \sum_{j_2=j_1}^{q} \psi_{j_1 j_2} x_{tj_1} z_{tj_2} \\
&\quad + \sum_{j_1=1}^{p} \sum_{j_2=j_1}^{p} \sum_{j_3=j_2}^{p} \beta_{j_1 j_2 j_3} x_{tj_1} x_{tj_2} x_{tj_3} + \ldots + \\
&\quad + \sum_{j_1=1}^{q} \sum_{j_2=j_1}^{q} \cdots \sum_{j_k=j_{k-1}}^{q} \gamma_{j_1 \ldots j_k} z_{tj_1} z_{tj_2} \cdots z_{tj_{k-1}} z_{tj_k} + R_k(\mathbf{u}_t) \\
&= g(\mathbf{x}_t, \mathbf{z}_t; k, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\psi}) + R_k(\mathbf{u}_t) \qquad (1.2)
\end{aligned}
$$

where $\max\{p, q\} \leq k$ (for notational reasons; this is not a restriction), $R_k(\mathbf{u}_t)$ is the remainder, and the $\boldsymbol{\beta}$'s, $\boldsymbol{\gamma}$'s, and $\psi$'s are parameters. Polynomial $g$ in (1.2) contains all possible combinations of $x_{ti}$ and $z_{ti}$ up to order $k$.

Since the true data-generating process is a function of $\mathbf{x}_t = (x_{t1}, \ldots, x_{tp})'$ only, it can be adequately approximated by a polynomial in the components of $\mathbf{x}_t$. This means that all terms in (1.2) containing elements of $\mathbf{z}_t$ have zero coefficients, i.e. all $\boldsymbol{\gamma}$'s and $\psi$'s are zero. Moreover, the remainder term is a function of $\mathbf{x}_t$ only: $R_k(\mathbf{u}_t) \equiv R_k(\mathbf{x}_t)$, $\forall\ t$. Thus, the relevant $k$th-order polynomial of type (1.2) is given by

$$
\begin{aligned}
f(\mathbf{x}_t; \theta) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{tj} + \sum_{j_1=1}^{p} \sum_{j_2=j_1}^{p} \beta_{j_1 j_2} x_{tj_1} x_{tj_2} + \sum_{j_1=1}^{p} \sum_{j_2=j_1}^{p} \sum_{j_3=j_2}^{p} \beta_{j_1 j_2 j_3} x_{tj_1} x_{tj_2} x_{tj_3} \\
+ \ldots + R_k(\mathbf{x}_t). \qquad (1.3)
\end{aligned}
$$

One may write equation (1.2) in matrix notation as

$$
\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{\mathbf{x}}\boldsymbol{\gamma} + \mathbf{R}_k(\mathbf{X}) + \boldsymbol{\varepsilon} \qquad (1.4)
$$

where $\mathbf{X}$ is a $T \times m(k)$ matrix whose $t-$th row contains an intercept and $m(k)$ products of elements of $\mathbf{x}_t$ and $\mathbf{Z}_{\mathbf{x}}$ a $T \times n(k)$ matrix whose $t-$th row consists of $n(k)$ products including at least one element of $\mathbf{z}_t$. Using the additional notation $\mathbf{W} = [\ \mathbf{X} \quad \mathbf{Z}_{\mathbf{x}}\ ]$, $\boldsymbol{\delta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$ and $\boldsymbol{\varepsilon}^* = \mathbf{R}_k(\mathbf{X}) + \boldsymbol{\varepsilon}$, one can write (1.4) in compact form

$$
\mathbf{y} = \mathbf{W}\boldsymbol{\delta} + \boldsymbol{\varepsilon}^*. \qquad (1.5)
$$

Note that the true parameter vector is $\boldsymbol{\delta} = (\boldsymbol{\beta}', \mathbf{0}')'$. If $\mathbf{W}$ is of full column rank, the OLS estimator of $\boldsymbol{\delta}$ in (1.5)

$$
\widehat{\boldsymbol{\delta}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y} \qquad (1.6)
$$

exists.

From the Stone-Weierstrass theorem it follows that there exists an order $k < \infty$ such that the remainder is negligible. Setting $\mathbf{R}_k(\mathbf{X}) \equiv \mathbf{0}$, one obtains

$$
\hat{\boldsymbol{\delta}} = \boldsymbol{\delta} + (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\varepsilon}. \qquad (1.7)
$$

We further assume that the following conditions (see White (1984), p. 119) hold:

(i) $\{(\mathbf{w}_t, \varepsilon_t)'\}$ is a stationary ergodic sequence ($\mathbf{w}_t$ is the $t-$th row of $\mathbf{W}$)

(ii) (a) $E\{\mathbf{w}_0\varepsilon_0 \mid \mathcal{F}_{-m}\} \to 0$ in quadratic mean as $m \to \infty$ where $\mathcal{F}_t$ is the information set containing all information about $\mathbf{w}_t$ and $\boldsymbol{\varepsilon}_t$ up until $t$,

(b) $E \mid \varepsilon_t w_{ti} \mid^2 < \infty$, $i = 1, ..., m(k) + n(k) + 1$,

(c) $\mathbf{V}_T = var\left(T^{-1/2}\mathbf{W}'\boldsymbol{\varepsilon}\right)$ is uniformly positive definite,

(d) $\sum_{j=0}^{\infty} var\,(R_{oij})^{1/2} < \infty, i = 1, ..., m(k) + n(k) + 1$, where $R_{oij} = E\,(w_{0i}\varepsilon_0 \mid \mathcal{F}_{-j}) - E\left(w_{0i}\varepsilon_0 \mid \mathcal{F}_{-(j+1)}\right), i = 1, ..., m(k) + n(k) + 1$

(iii) (a) $E \mid w_{ti} \mid^2 < \infty$, $i = 1, ..., m(k) + n(k) + 1$,

(b) $E\mathbf{w}_t\mathbf{w}_t'$ is positive definite.

It now follows from Theorem 5.16 in White (1984) that the OLS estimator (1.6) is consistent for $\boldsymbol{\delta}$ and asymptotically normally distributed. Thus, the larger $T$, the more likely is it that the estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\psi}$ are statistically insignificant. By finding the $n\,(k)$ insignificant parameter estimates of the product terms in the $k$th-order polynomial involving elements of $\mathbf{z}_t$, one is therefore able to identify all elements of $\mathbf{z}_t$ and exclude them from the variable set.

A key issue is how to find a polynomial such that the remainder under the null hypothesis is negligible. Suppose we select $k_\delta < k$ (the "adequate" order). In that case, (1.7) becomes

$$\hat{\boldsymbol{\delta}} = \boldsymbol{\delta} + \left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\mathbf{R}_{k_\delta}(\mathbf{X}) + \left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\boldsymbol{\varepsilon}. \tag{1.8}$$

Since $\left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\mathbf{R}_{k_\delta}(\mathbf{X})$ is not converging to a null vector, one obtains biased estimates independently of the sample size $T$. On the other hand, increasing $k_\delta$ leads to a larger estimation variance. As a consequence, an appropriate choice of $k_\delta$ is essential in practice to minimize and balance the opposite effects of bias and variance. This issue will be further discussed in the next section.

While the above approximation result is a general one, some nonlinear functions may not have polynomial approximations that satisfy assumption (i) required for consistency and asymptotic normality of $\hat{\boldsymbol{\delta}}$. This is because the existence of moments of order $2k$ for the variables involved in (1.3) is required for the assumption to be valid. For univariate autoregressive models, one also has to assume that the errors are serially uncorrelated to avoid violating assumption (ii) (a). Note that all moments of $y_t$ exist when $\{\varepsilon_t\}$ is a sequence of zero-mean independent identically distributed stochastic variables such that all moments of their distribution are finite, and when the autoregressive process possesses a convergent Volterra expansion, see Priestley (1981), pp. 869-871.

The assumptions stated above restrict the applicability of our theory. It should be noticed, however, that the polynomial approximation to the data generating process (1.1) obtained by setting $\mathbf{R}_k\,(\mathbf{X}) \equiv 0$ in (1.4) is never assumed to be a data-generator. It only serves as a variable-selection device as will be explained in the next section. The time series are always generated by model (1.1). On the other hand, the assumption of a compact sample space is a technical one and does not in our case seem to have practical implications.

## 1.3 The model selection procedure

Using the previous results we now propose two variable selection strategies for models with an unknown, possibly nonlinear conditional mean function $f(\cdot)$. We first describe the simple

variant. First, one chooses an appropriate order $k$ for polynomial (1.2) which is used to approximate the unknown function. As we have pointed out above, there exists a trade-off between the bias due to the remainder $\mathbf{R}_k(\mathbf{X})$ and the variance. Therefore, the choice of $k$ is important. A rule-of-thumb that yields positive results in our simulations discussed in Section 4 is to choose $k$ such that there are at least about four times as many observations as there are regressors in the polynomial. Note, however, that when $T \longrightarrow \infty$, we should allow $\{m(k) + n(k) + 1\} / T \longrightarrow 0$.

After having selected $k$, the next step is to select the appropriate variables. If the total number of possible variables $p + q$ is not too large, then this can be done by conducting a full search through all possible subsets of $\mathbf{u}_t$ as follows. First regress $y_t$ on all variables in equation (1.2) and compute the value of an appropriate model selection criterion. We use the Schwarz model selection criterion (SBIC) which was independently proposed by Rissanen (1978) and Schwarz (1978). For linear models it has been shown to be consistent under quite general conditions. For alternative model selection criteria see, for example, Judge, Griffiths, Hill, Lütkepohl, and Lee (1984), pp. 862-874, or Teräsvirta and Mellin (1986). Next, omit one regressor from the original model, regress $y_t$ on all products of variables remaining in the polynomial, and compute the value of SBIC. Repeat this by omitting each regressor in turn. Continue by simultaneously omitting two regressors from the original model. Proceed until the regression only consists of a polynomial of a function with a single regressor. Leave this out as well to check for white noise with a possibly nonzero mean. This amounts to estimating $\sum_{i=1}^{p+q} \binom{p+q}{i} + 1 = 2^{p+q}$ linear models by ordinary least squares. Select the combination of variables that yields the lowest value of SBIC. Due to the consistency of SBIC in linear models, this simple variable selection procedure is consistent if a $k_\delta$ is chosen which is equal to or larger than $k$ in the case $\mathbf{R}_k(\mathbf{x}_t) \equiv \mathbf{0}$.

Sometimes the unknown function in (1.1) may be at least approximately linear. Therefore it may be a good idea to begin the variable selection procedure by testing linearity. This can be done by testing the null hypothesis that the coefficients of all the terms in polynomial (1.2) of order higher than one equal zero. Teräsvirta, Lin, and Granger (1993) used this idea in testing linearity against a single hidden layer feedforward artificial neural network model. If the null hypothesis is not rejected, then the model selection simplifies to variable selection in linear regression using subset regressions. This means saving computer time and making the selection procedure more efficient. As before, a suitable model selection criterion such as SBIC may be applied to the problem.

In practice, $k$ is unknown and has to be chosen by the researcher. As our simulation results show, this choice may sometimes be difficult in large samples. We therefore propose another, more complex version with automatic estimation of $k$. In this variant, one only has to choose an upper bound $k_{max}$. Next, for each subset of variables one fits all polynomials with orders $1 \leq k \leq k_{max}$ and chooses $\hat{k}$ for which the estimated polynomial exhibits the lowest value of the chosen model selection criterion. This should on average minimize the joint impact of bias and variance. Note that the estimated order $\hat{k}$ may vary across the subsets of variables considered. Now $k_{max}$ may be chosen in the same way as $k$ in the simple procedure. The variable selection procedure with automatic order choice is also expected to work when a finite order representation for $f(\cdot)$ does not exist. This conjecture is confirmed by our simulation results. Note that the automatic order choice also checks for linearity since $\hat{k} = 1$ corresponds to a linear model.

In order to evaluate the performance of our two procedures we shall compare them to the nonparametric FPE procedure by Tschernig and Yang (2000). This is done by simulation. The latter procedure is selected as a benchmark because it is consistent while requiring weaker

moment assumptions than ours, that is, the function $f(\cdot)$ only needs to be differentiable up to order four. This is achieved by using local estimation techniques. Instead of increasing the order of the polynomial with increasing sample size as for the global estimator (1.2), the order of the polynomial is fixed while the latter is estimated only locally. One thus estimates the value of $f(\mathbf{u})$ at $\mathbf{u}^0$ by estimating a first-order polynomial with observations lying in a neighbourhood of $\mathbf{u}^0$. Clearly, the smaller the neighbourhood determined by a so-called bandwidth parameter, the smaller the bias but the larger the estimation variance. With increasing sample size, the approximation error is reduced by decreasing the size of the neighbourhood instead of increasing the order $k$ of the general polynomial (1.2).

The trade-off between bias and variance allows one to derive an asymptotically optimal bandwidth. Using recent results of Yang and Tschernig (1999), it can be estimated by plug-in methods. The nonparametric CAFPE proposed by Tschernig and Yang (2000) and used in the simulation study is given in equation $(A.5)$ in the Appendix.

## 1.4 A simulation study

To find out how the selection procedure functions in practice, we conduct a simulation study. We simulate both nonlinear autoregressive models and models with exogenous regressors. The autoregressive data-generating processes (DGP) are defined as follows:

(i) Artificial Neural Network model with two lags and a single hidden unit (**ANN**)

$$Y_t = -0.5 + \frac{1}{1+\exp\{-2(Y_{t-1} - 3Y_{t-2} - 0.05)\}} + 0.1\varepsilon_t, \quad \varepsilon_t \sim NID\,(0,1) \tag{1.9}$$

(ii) Nonlinear Additive AR(2) process (**NLAR1**)

$$Y_t = -0.4 \cdot \frac{3 - Y_{t-1}^2}{1 + Y_{t-1}^2} + 0.6 \cdot \frac{3 - (Y_{t-2} - 0.5)^3}{1 + (Y_{t-2} - 0.5)^4} + 0.1\varepsilon_t, \quad \varepsilon_t \sim NID(0,1) \tag{1.10}$$

(iii) Nonlinear Additive AR(4) proces (**NLAR1_14**)

$$Y_t = -0.4 \cdot \frac{3 - Y_{t-1}^2}{1 + Y_{t-1}^2} + 0.6 \cdot \frac{3 - (Y_{t-4} - 0.5)^3}{1 + (Y_{t-4} - 0.5)^4} + 0.1\varepsilon_t, \quad \varepsilon_t \sim NID(0,1) \tag{1.11}$$

(iv) Nonlinear AR(2) process (**NLAR4**)

$$Y_t = 0.9 \cdot \frac{1}{1 + Y_{t-1}^2 + Y_{t-2}^2} - 0.7 + 0.1\varepsilon_t \tag{1.12}$$

$\varepsilon_t \sim$ triangular density, $E\varepsilon_t = 0$, density positive for $|\varepsilon_t| < 0.1\sqrt{6}$

(v) Logistic Smooth Transition AR(2) process (**LSTAR**)

$$\begin{aligned} Y_t = {} & 1.8Y_{t-1} - 1.06Y_{t-2} + (0.02 - 0.90Y_{t-1} + 0.795Y_{t-2}) \\ & \times \frac{1}{1+\exp\{-100(Y_{t-1} - 0.02)\}} + 0.1\varepsilon_t, \quad \varepsilon_t \sim NID\,(0,1) \end{aligned} \tag{1.13}$$

(vi) Nonlinear AR(2) process (**SIN**)

$$Y_t = \frac{\sin(\pi Y_{t-1}) + \sin(4\pi Y_{t-2})}{2} + 0.1\varepsilon_t, \quad \varepsilon_t \sim NID(0,1)\,. \tag{1.14}$$

Simulating these processes did not reveal any instability. The first 400 observations of each series were discarded. Models (4.2), (4.3) and (4.6) are additive; models (4.1), (4.4) and (4.5) are not. Model (4.5) is the same as that in Teräsvirta (1994), except for the error variance, which is larger. The sine model (4.6) may be expected to be a problematic one, as it is not well approximated by a combination of low-order polynomials of its variables. We also simulated models with exogenous regressors with the same functional form as for the autoregressive processes, the lags being replaced by normally distributed exogenous regressors generated by a stationary first-order vector autoregression. The results from both cases are rather similar, and we therefore only report those based on univariate models. The remaining results are available from the authors upon request.

We consider three sample sizes, $T = 100, 200, 2000$, and perform 500 replications in each experiment. We compare the two polynomial-based procedures with the nonparametric CAFPE method. Since for $T = 2000$ the CAFPE procedure is computationally very demanding, the number of replications in that case has been limited to 100. In general, the set of potential lags includes lags up to lag 6 except for the simple procedure and $T = 100$ where $p + q = 5$.

Table B.1 contains the results for $T = 100$. For each of the three procedures and six models we present the numbers of correct lag selection (C), of not selecting all the relevant lags (underfitting: U) and of selecting all the relevant plus at least one redundant lag (overfitting: O). Moreover, for the simple method, $k = 3$, and the results of linearity testing are presented as well. First note that both the procedure based on estimating $k$ and the one based on fixing $k = 3$ a priori, lead to similar results. For models (1.10) and (1.11) all procedures work very well. Comparing the number of correct selections, one finds that for the neural network model (1.9) and the logistic STAR model (1.13) the proposed polynomial-based procedures work better than the CAFPE method, while for the model (1.12) and the sine model (1.14) the opposite is true. The failure of the global approximation procedure in the case of the sine model (1.14) is due to the low-order polynomial that does not allow for an adequately accurate approximation of the sine function. In this case the CAFPE procedure has a clear advantage. It may be noted that most realizations from the nonlinear model (1.12) seem linear, and selecting the correct lags is more difficult than in most other models. Note that the ranking of the models may easily be changed by changing the error variance. Our results also indicate that the nonparametric approach works better when the DGP is additive than when it is not. Finally, the CAFPE procedure tends to overfit more often than our global estimation procedure.

Doubling the number of observations to 200 leads to a general increase of correct selections with two exceptions. The first one is the sine model (1.14), when we apply the simple polynomial-based procedure with $k = 3$. The second exception is the neural network model (1.9) when the lags are selected by the CAFPE procedure. Note that underfitting decreases substantially in all cases.

Increasing the sample size to 2000 while letting $k = 3$, considerably weakens the performance of our simple procedure, as can be seen from Table B.3. The results in Table B.4 show that setting $k = 5$ is now a much better choice, except for model (1.12). These results stress the importance of choosing $k$ carefully. The proposed automatic order choice procedure performs remarkably well in all examples, even for the sine model (1.14). Moreover, our

lag selection procedure with automatic order-choice substantially outperforms the CAFPE procedure that frequently overfits. Note that results on the latter procedure are based on 100 replications only. Thus, in large samples, or as long as the functional form can be sufficiently well approximated by a low-order polynomial, our variable selection method with automatic order choice leads to satisfactory results.

## 1.5 Conclusions

Our variable selection procedure is specifically designed for nonlinear models with unknown functional form. One of the main advantages of this technique is that it is simple and computationally feasible, because it is based on ordinary least squares. It is applicable already in small samples and the computational burden remains tolerable even when the series are long. In the latter situation the currently available nonparametric techniques become computationally expensive or even prohibitive. The standard subset selection procedure for linear models constitutes a special case of our technique. When the unknown function can be sufficiently well approximated by a low-order polynomial, or when the sample size is large enough, the proposed variable selection method with automatic order choice compares quite favourably with the nonparametric CAFPE procedure. It is therefore a useful statistical tool for investigators facing variable selection problems in empirical work.

# Bibliography


AKAIKE, H. (1969): "Fitting Autoregressive Models for Prediction," *Annals of the Institute of Statistical Mathematics*, 21, 243–47.

——— (1974): "A New Look at the Statistical Model Identification," *IEEE Transactions of Automatic Control*, AC-19, 716–23.

AUESTAD, B., AND D. TJØSTHEIM (1990): "Identification of Nonlinear Time Series: First Order Characterization and Order Determination," *Biometrika*, 77, 669–87.

JUDGE, G. G., W. E. GRIFFITHS, R. C. HILL, H. LÜTKEPOHL, AND T.-C. LEE (1984): *The Theory and Practice of Econometrics*. New York: Wiley.

PRIESTLEY, M. B. (1981): *Spectral Analysis and Time Series*. New York, Academic Press.

RISSANEN, J. (1978): "Modeling by Shortest Data Description," *Automatica*, 14, 465–471.

ROYDEN, H. L. (1963): *Real Analysis*. New York, Macmillan.

SCHWARZ, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–64.

SILVERMAN, B. W. (1986): *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

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

TERÄSVIRTA, T., C.-F. LIN, AND C. W. J. GRANGER (1993): "Power of the Neural Network Linearity Test," *Journal of Time Series Analysis*, 14, 209–220.

TERÄSVIRTA, T., AND I. MELLIN (1986): "Model Selection Criteria and Model Selection Tests in Regression Models," *Scandinavian Journal of Statistics*, 13, 159–171.

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

TSCHERNIG, R., AND L. YANG (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.

WHITE, H. (1984): *Asymptotic Theory for Econometricians*. Academic Press, Orlando, FL.

YANG, L., AND R. TSCHERNIG (1999): "Multivariate Bandwidth Selection for Local Linear Regression," *Journal of the Royal Statistical Society B*, 61, 793–815.

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

# Appendix A

# The nonparametric CAFPE

Denote by $\mathbf{w}$ a $(m \times 1)$ subvector of $\mathbf{u}$, $m \leq p+q$. A local linear estimate $\widehat{f}(\mathbf{w}, h)$ of the function at $\mathbf{w}$ using the bandwidth $h$ is given by the estimated constant $\hat{c}_0$ of a linear Taylor expansion fitted locally around $\mathbf{w}$

$$\{\widehat{c}_0, \widehat{\mathbf{c}}\} = \arg\min_{\{c_0, \mathbf{c}\}} \sum_{t=1}^{T} \{y_t - c_0 - (\mathbf{w}_t - \mathbf{w})'\mathbf{c}\}^2 K_h(\mathbf{w}_t - \mathbf{w}) \tag{A.1}$$

where $K(\cdot)$ denotes a standard kernel function and $K_h(\mathbf{w}_t - \mathbf{w}) = h^{-m} \prod_{i=1}^{m} K\left((w_{t,i} - w_i)/h\right)$. The integrated mean squared error can then be estimated by

$$\widehat{A}(h) = T^{-1} \sum_{t=1}^{T} \left\{y_t - \widehat{f}(\mathbf{w}_t, h)\right\}^2 w(\mathbf{u}_t) \tag{A.2}$$

where the integration is restricted to the domain of the weight function $w(\cdot)$ which is defined for the full vector $\mathbf{u}$. Furthermore, define the term

$$\widehat{B}(\widehat{h}_B) = T^{-1} \sum_{t=1}^{T} \left\{y_t - \widehat{f}(\mathbf{w}_t, \widehat{h}_B)\right\}^2 w(\mathbf{u}_t)/\widehat{\mu}(\mathbf{w}_t, \widehat{h}_B) \tag{A.3}$$

where $\widehat{\mu}(\cdot)$ is a Gaussian kernel estimator of the density $\mu$ using Silverman's (Silverman (1986)) rule-of-thumb bandwidth $\widehat{h}_B = h(m+2, \hat{\sigma}, T)$ and

$$h(k, \sigma, n) = \sigma\{4/k\}^{1/(k+2)} n^{-1/(k+2)}.$$

Moreover, $\widehat{\sigma} = \left(\prod_{j=1}^{m} \sqrt{Var(\mathbf{w}_j)}\right)^{1/m}$ denotes the geometric mean of the standard deviation of the regressors.

The local linear estimate of the FPE is then given by

$$AFPE = \widehat{A}(\widehat{h}_{opt}) + 2K(0)^m T^{-1} \widehat{h}_{opt}^{-m} \widehat{B}(\widehat{h}_B) \tag{A.4}$$

where the plug-in bandwidth is computed from

$$\widehat{h}_{opt} = \left\{m||K||_2^{2m} \widehat{B}(\widehat{h}_B) T^{-1} \widehat{C}(\widehat{h}_C)^{-1} \sigma_K^{-4}\right\}^{1/(m+4)}$$

with $||K||_2^2 = \int K^2(u)du$, $\sigma_K^2 = \int K(u)u^2 du$. Note that the second term in $(A.4)$ serves as a penalty term to punish overfitting.

The estimation of $C$ involves second derivatives which are estimated with a local quadratic estimator that excludes all cross derivatives. It is a simplification of the partial local cubic estimator of Yang and Tschernig (1999). The bandwidth estimate $\widehat{h}_C$ is given by $h(m+4, 3\widehat{\sigma}, T)$.

Based on theoretical reasons and Monte Carlo evidence provided in Tschernig and Yang (2000), the authors suggest to use the corrected FPE

$$CAFPE = AFPE\left\{1 + mT^{-4/(m+4)}\right\} \tag{A.5}$$

where the correction increases the probability of correct fitting. One then chooses that variable vector $\mathbf{w}^*$ for which $AFPE$ or $CAFPE$ are minimized. Tschernig and Yang (2000) show that both criteria are consistent if the nonlinear process is stationary and $\beta$-mixing and if some regularity conditions hold.

# Appendix B

# Tables and Figures

| | Taylor expansion-based methods | | | | | | | | | | CAFPE | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | k=3 | | | | | | | Estimated $\widehat{k}$ | | | | | |
| | Linear | | | Nonlinear | | | L+N | | | | | | |
| Model | C | U | O | C | U | O | C | C | U | O | C | U | O |
| ANN | 161 | 56 | 7 | 224 | 52 | 0 | 385 | 371 | 101 | 28 | 102 | 92 | 306 |
| NLAR1 | 0 | 0 | 0 | 499 | 0 | 1 | 499 | 498 | 0 | 2 | 495 | 0 | 5 |
| NLAR1_14 | 0 | 0 | 0 | 500 | 0 | 0 | 500 | 499 | 0 | 1 | 499 | 0 | 1 |
| NLAR4 | 35 | 439 | 0 | 0 | 26 | 0 | 35 | 51 | 444 | 5 | 119 | 361 | 20 |
| LSTAR | 215 | 11 | 3 | 260 | 10 | 1 | 475 | 471 | 13 | 16 | 247 | 13 | 240 |
| SIN | 11 | 28 | 0 | 0 | 271 | 0 | 11 | 4 | 496 | 0 | 294 | 189 | 17 |

**Notes:** The results for all three procedures are based on 500 replications; C: the number of correct lag selections; U: number of underfitting models (not all correct lags in the model); O: number of overfitting models (all correct lags plus some additional ones). L+N stands for the sum of linear and nonlinear models. The maximum number of lags to be selected equals five (lags 1 to 5) for $k = 3$ and six (lags 1 to 6) for the case where $k$ is estimated.

Table B.1: MONTE CARLO RESULTS OF LAG SELECTION PROCEDURES GIVEN 100 OBSERVATIONS

| | Taylor expansion-based methods | | | | | | | | | | CAFPE | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | k=3 | | | | | | | Estimated $\hat{k}$ | | | | | |
| | Linear | | | Nonlinear | | | L+N | | | | | | |
| Model | C | U | O | C | U | O | C | C | U | O | C | U | O |
| ANN | 22 | 1 | 6 | 454 | 17 | 0 | 476 | 471 | 23 | 6 | 48 | 34 | 418 |
| NLAR1 | 0 | 0 | 0 | 500 | 0 | 0 | 500 | 500 | 0 | 0 | 499 | 0 | 1 |
| NLAR1_14 | 0 | 0 | 0 | 500 | 0 | 0 | 500 | 500 | 0 | 0 | 500 | 0 | 0 |
| NLAR4 | 96 | 376 | 3 | 0 | 25 | 0 | 96 | 200 | 283 | 17 | 275 | 167 | 58 |
| LSTAR | 30 | 0 | 1 | 469 | 0 | 0 | 499 | 499 | 0 | 1 | 211 | 1 | 288 |
| SIN | 3 | 29 | 1 | 0 | 467 | 0 | 3 | 0 | 499 | 1 | 487 | 1 | 12 |

**Notes:** The maximum number of lags to be selected equals six (lags 1 to 6).

Table B.2: MONTE CARLO RESULTS OF LAG SELECTION PROCEDURES GIVEN 200 OBSERVATIONS

| | Taylor expansion-based method | | | | | | |
|---|---|---|---|---|---|---|---|
| | $k = 3$ | | | | | | |
| | Linear | | | Nonlinear | | | L + N |
| Model | C | U | O | C | U | O | C |
| ANN1 | 0 | 0 | 0 | 500 | 0 | 0 | 500 |
| NLAR1 | 0 | 0 | 0 | 384 | 0 | 116 | 384 |
| NLAR1_14 | 0 | 0 | 0 | 111 | 0 | 389 | 111 |
| NLAR4 | 170 | 0 | 10 | 298 | 22 | 0 | 468 |
| LSTAR1 | 0 | 0 | 0 | 500 | 0 | 0 | 500 |
| SIN1 | 0 | 0 | 0 | 31 | 469 | 0 | 31 |

**Notes:** The maximum number of lags to be selected equals six (lags 1 to 6).

Table B.3: MONTE CARLO RESULTS OF THE SIMPLE TAYLOR LAG SELECTION PROCEDURE GIVEN 2000 OBSERVATIONS

| | Taylor expansion-based methods | | | | | | | | | | CAFPE | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | k=5 | | | | | | | Estimated $\hat{k}$ | | | | | |
| | Linear | | | Nonlinear | | | L+N | | | | | | |
| Model | C | U | O | C | U | O | C | C | U | O | C | U | O |
| ANN | 0 | 0 | 0 | 500 | 0 | 0 | 500 | 500 | 0 | 0 | 1 | 0 | 99 |
| NLAR1 | 0 | 0 | 0 | 500 | 0 | 0 | 500 | 500 | 0 | 0 | 67 | 0 | 33 |
| NLAR1_14 | 0 | 0 | 0 | 500 | 0 | 0 | 500 | 500 | 0 | 0 | 100 | 0 | 0 |
| NLAR4 | 150 | 0 | 5 | 0 | 345 | 0 | 150 | 499 | 0 | 1 | 73 | 0 | 27 |
| LSTAR | 0 | 0 | 0 | 500 | 0 | 0 | 500 | 500 | 0 | 0 | 59 | 0 | 41 |
| SIN | 0 | 0 | 0 | 472 | 28 | 0 | 472 | 452 | 48 | 0 | 100 | 0 | 0 |

**Notes:** The maximum number of lags to be selected equals six (lags 1 to 6) The number of replications when simulating the Taylor expansion-based method is 500, whereas the corresponding number for the CAFPE procedure is 100.

Table B.4: MONTE CARLO RESULTS OF LAG SELECTION PROCEDURE GIVEN 2000 OBSERVATIONS

# Essay 2

# Modelling and forecasting economic time series with single hidden-layer feedforward autoregressive artificial neural networks

## 2.1 Introduction

Neural networks (herafter NN) are a class of nonlinear models which have been applied to a wide spectrum of disciplines like psycology, computer science, engineering, linguistic and economics. They originate from the work of cognitive scientists who have been inspired by the neural structure of the brain. They seem particularly suited to study topics like pattern recognition, classification and nonlinear features in the data. Whereas the connection between them and statistical inference is generally well accepted, they are often built without applying statistical tools to the data of interest. A particularly interesting feature is their capability of approximating any Borel-measurable function to any degree of accuracy as pointed out in Hornik, Stinchcombe, and White (1989). The quality of the approximation depends on the specification of the model, the quality of its estimation procedure and the set of tests to evaluate it.

In this paper, the purpose is to develop a statistically consistent procedure to model time series data. The time series is first tested for linearity; if rejected, a NN model with a linear part and one hidden unit is estimated. The presence of additional hidden units is tested; if detected, a model with two hidden units is estimated. The process continues similarly and when the number of units suffices, tests for the adequacy of the estimated model are carried out and the model evaluated. The novelty of the whole methodology is that it utilizes statistical tools from the beginning to the end. It can be applied to any series of data. In this thesis I have been mainly interested in time series; therefore, this essay will be mainly based on this topic. Finally, an application with exogenous regressors is discussed in the last paragraph.

In time series applications, the first issue is the choice of the lags to be included in the model. I do this using a semiparametric method as in Rech, Teräsvirta, and Tschernig (2001). The authors attained satisfactory results from simulations with different types of nonlinear functions even in small samples.

The choice of the number of hidden units determines the accuracy of the approximation

to the true data generating process (hereafter DGP). It is, therefore, important to utilize an appropriate method for this choice. White (1989b) suggests to choose the number of hidden units using a model selection criterion like AIC (Akaike (1974)) or SBIC (Rissanen (1978); Schwarz (1978)). These criteria can always be utilized but lack theoretical underpinnings in nonlinear modelling. Here a NN model with $q$ hidden units is tested against a model with $q+Q$ ones. A sequence of Lagrange Multiplier tests is carried out from specific to general until the first acceptance of the null hypothesis of $q$ hidden units.

The estimation procedure is carried out following White (1989c). In order to improve the accuracy of the procedure, a set of starting values is chosen linearizing the model and using ordinary least squares (herafter OLS). The nonlinear least squares procedure is subsequently carried out by the Broyden-Fletcher-Goldfarb-Shanno algorithm (hereafter BFGS) in two different ways: full estimation, and estimation conditional on previous hidden units estimates, which I call frozen estimation. A subset of parameters can be restricted to certain values in both cases.

Evaluating a model requires, as in Eitrheim and Teräsvirta (1996), developing specific LM tests for the hypothesis of no error autocorrelation and parameter constancy, while additional nonlinearity is already checked when the number of hidden units is chosen. A Monte-Carlo study gives an idea of the performances of the overall procedure.

The whole device is applied to the Canadian Lynx and Sunspot series (endogenous cases) and to a series of stock returns modelled on exogenous financial and economic variables as in Qi (1999). I compare the performances in forecasting a linear model and the ANN models by four numerical benchmarks.

## 2.2 Neural network models

A single hidden-layer feedforward artificial neural network model (hereafter ANN) may be seen as a "net of knots" which connects several inputs and one or more outputs. In case of $q$ links, the model can be written:

$$y_t = \beta_0 + \sum_{j=1}^{q} \beta_j \psi\left(\widetilde{\gamma}_j' \mathbf{w}_t\right) \tag{2.1}$$

where $\mathbf{w}_t = (w_{1t}, w_{2t}, ..., w_{kt}, 1)'$ is the $(k+1) \times 1$ vector of input variables and $\widetilde{\gamma}_j = \left(\widetilde{\gamma}_{j1}, \widetilde{\gamma}_{j2}, ..., \widetilde{\gamma}_{j,k+1}\right)'$ is the $(k+1) \times 1$ vector of parameters or "weights" of the $j$th "hidden unit". The choice of notation $\widetilde{\gamma}_j$ will be clarified in the estimation section. Functions $\psi(\widetilde{\gamma}_j' \mathbf{w}_t)$ are the hidden units or so-called "activation functions". In this paper, I choose the logistic, $\psi(z) = (1 + \exp(-z))^{-1}$, $z \in R$, which takes values between zero and one.

The use of ANN models is motivated by the mathematical result, see Hornik, Stinchcombe, and White (1989), Funahashi (1989), or Cybenko (1989), that a single hidden layer feedforward network with an arbitrary sigmoid activation function is capable of approximating any Borel measurable function to any desirable degree of accuracy, if sufficiently many hidden units are included in function (2.1). Adding noise, I can view the resulting ANN model as a nonlinear regression model

$$y_t = \beta_0 + \sum_{j=1}^{q} \beta_j \psi(\widetilde{\gamma}_j' \mathbf{w}_t) + u_t \tag{2.2}$$

where $\{u_t\}$ is assumed to be a sequence of independent, normally distributed (n.i.d.) variables with mean zero. It should be noted, however, that the approximation result requires the

assumption that $\mathbf{w}_t$ contains exactly the relevant variables. In practice, they are unknown and have to be selected in some suitable fashion.

Considering model (2.2) as a starting point for the analysis, some additional assumptions have to be made. In this paper, I consider the univariate case and want to model a single time series $\mathbf{y} = \{y_t\}$, $t = 1, ..., T$. For notational simplicity, I assume for the time being that $\mathbf{w}_t = (y_{t-1}, y_{t-2}, ..., y_{t-k}, 1)'$. In practice, the true set of lags included in $\mathbf{w}_t$ is unknown and has to be determined from the data. Whenever $\mathbf{w}_t$ would include some exogenous variables, or be completely exogenous to the dependent variable $y_t$, the overall device for modelling the series works unchanged, as demonstrated in the application to stock returns with exogenous regressors.

In economic applications, it is usually important to include a linear part in the model. Our model becomes

$$\begin{aligned} y_t &= \alpha_0 + \sum_{i=1}^{k} \alpha_i y_{t-i} + \sum_{j=1}^{q} \beta_j \psi(\tilde{\boldsymbol{\gamma}}_j' \mathbf{w}_t) + u_t \\ &= \boldsymbol{\alpha}' \mathbf{w}_t + \sum_{j=1}^{q} \beta_j \psi(\tilde{\boldsymbol{\gamma}}_j' \mathbf{w}_t) + u_t, \ t = 1, ..., T. \end{aligned} \tag{2.3}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_k, \alpha_0)'$. I call model (3.5) an autoregressive ANN model of order $k$, AR-ANN($k$) for short. Leisch, Trapletti, and Hornik (1999) showed that (3.5) is weakly stationary and ergodic if the roots of the characteristic polynomial $1 - \sum_{i=1}^{k} \alpha_i z^i$ lie outside the unit circle. The stationarity assumption is required in the following because the modelling techniques I am going to propose in this paper depend on it.

## 2.3 Modelling cycle

In this paper I consider the modelling of economic time series with artificial neural networks. In order to proceed in a systematic fashion, I have to construct a modelling cycle for building ANN models. This consists of stages of specification, estimation and evaluation. The last-mentioned stage consists of in-sample evaluation. If an evaluated model is deemed successful it also has to be checked out-of-sample. Thus a forecasting stage is required to complete the cycle. For similar modelling cycles see, for example: Box and Jenkins (1970) for ARMA models; Tsay (1989) for univariate and Tsay (1998) for vector SETAR models; and Teräsvirta (1994), Teräsvirta (1998) for STAR and STR models. In the following I shall consider these stages in the ANN framework.

### 2.3.1 Specification and estimation of the ANN model

As is obvious from equation (3.5), two specification problems demand particular attention in ANN models. One is the selection of variables, that is, the elements of $\mathbf{w}_t$. In the univariate case this amounts to choosing the right lags into the model. The other problem is selecting the number of hidden units or activation functions. Since the ANN model is just a richly parametrized approximation to an unknown nonlinear model, the number of hidden units cannot be assumed known in advance. These problems have been discussed in the ANN modelling literature that Anders and Korn (1999) have recently surveyed. They mention three approaches to the problem: *regularization, pruning*, and *network training*. The first

one may be characterized as a penalized nonlinear least squares approach in which adding another parameter and thus lowering the minimum of the objective function carries a positive penalty. This approach implies estimating a large number of ANN models in search of the minimum of the penalized objective function. Besides, how the size of the penalty should be determined is not completely clear.

*Pruning* may be characterized as a general-to-specific or top-down method. Reed (1993) has surveyed this technique. The idea is to begin with a large model, and gradually diminish its size by reducing the number of hidden units and also by setting parameters in hidden units equal to zero. The latter operation may be done by likelihood ratio tests if the likelihood principle is applied in the estimation of the parameters; usually it is not. Because the starting-point is general, this method requires a large number of observations and a large number of ANN estimations. Besides, if the original number of hidden units exceeds the true one, the method may suffer from the fact that the ANN model to be estimated is not identified. I shall return to this point later.

*Network training* is an estimation method. The time series is split into two subseries: the training series and the validation series following it. The estimation is carried out with the training series and the estimated model is used for minimizing the sum of squared prediction errors or some similar criterion in the validation set. The estimation is repeated as long as the criterion values decrease. When they begin to increase the estimation is stopped. This procedure does not correspond to any conventional optimization method. It is also unclear how the decisions to set coefficients to zero and add or remove hidden units are taken. No universally accepted rules or statistical criteria seem to exist for this purpose.

In this paper, I shall consider a statistical approach to ANN model specification. This is what Anders and Korn (1999) also discussed. In their paper, the authors performed a simulation study comparing five different selection strategies for selecting the complexity of a neural network model. Such strategies were based on cross-validation, information criteria (AIC and NIC (Stone (1977))), the neural network test proposed in White (1989a) and the neural network based linearity test of Teräsvirta, Lin, and Granger (1993). Comparing all strategies, they found that the last-mentioned one led to the most reliable results. My approach may be seen as an extension to the work of Anders and Korn (1999) in that I also discuss evaluation issues that they did not highlight in their article. An advantage of statistical methods is that the modelling can proceed according to statistical principles. Although modelling is rarely free from subjective elements, they may play a less dominant role here than they do in the techniques mentioned above. A discussion of specification has to be preceded by considering estimation, to which I now turn.

### 2.3.2 ANN estimation procedure

Assume for the time being that the correct set of variables $\mathbf{w}_t$ has been selected. Variable selection will be discussed later. Consider the simplest case of an ANN model with a linear part and a single nonlinear hidden unit:

$$y_t = \boldsymbol{\alpha}'\mathbf{w}_t + \beta_1\psi(\tilde{\boldsymbol{\gamma}}_1'\mathbf{w}_t) + u_t, t = 1, ..., T, u_t \sim n.i.d.(0, \sigma_u^2) \tag{2.4}$$

where $\mathbf{w}_t = (w_{t1}, ..., w_{tk}, 1)'$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_k, \alpha_0)'$, and $\tilde{\boldsymbol{\gamma}}_1 = (\tilde{\gamma}_{11}, \tilde{\gamma}_{12}, ..., \tilde{\gamma}_{1,k+1})'$. In the time series case, $\mathbf{w}_t = (y_{t-1}, y_{t-2}, ..., y_{t-k}, 1)'$. Since this paper is focused on autoregressive models, I shall consider $\mathbf{w}_t$ a set of lags, demonstrating exogenous variables modelling in an application in the last part of the paper. I assume $\psi(\tilde{\boldsymbol{\gamma}}_1'\mathbf{w}_t)$ to be a logistic function:

$\psi(\widetilde{\gamma}_1'\mathbf{w}_t) = \left(1+\exp(-\widetilde{\gamma}_1'\mathbf{w}_t)\right)^{-1}$. I reparametrize the sigmoid as follows:

$$\begin{aligned}\psi\left(\widetilde{\gamma}_1'\mathbf{w}_t\right) &= \psi\left(\gamma_1^c, \mathbf{w}_t\right) \\ &= \left(1+\exp\left\{-\left(\gamma_{11}w_{t1}+\gamma_{12}w_{t2}+...+\gamma_{1,k}w_{tk}-c_1\right)\right\}\right)^{-1} \end{aligned} \tag{2.5}$$

where $\boldsymbol{\gamma}_1^c = \left(\gamma_{11}, \gamma_{12}, ..., \gamma_{1,k}, -c_1\right) = \left(\widetilde{\gamma}_{11}, \widetilde{\gamma}_{12}, ..., \widetilde{\gamma}_{1,k}, \widetilde{\gamma}_{1,k+1}\right)$, $\gamma_{ji} > 0$ for at least one $i$, $i = 1, ..., k$. The restriction on $\gamma_{ji}$ is an identification condition, removing one of the sources of lack of identifiability considered in Hwang and Ding (1997). Now, rewriting (2.5)

$$\begin{aligned}\psi\left(\boldsymbol{\gamma}_1^c, \mathbf{w}_t\right) &= \left(1+c_1^*\exp\left\{-\left(\gamma_{11}w_{t1}+\gamma_{12}w_{t2}+...+\gamma_{1,k}w_{tk}\right)\right\}\right)^{-1} \\ &= \psi\left(\boldsymbol{\gamma}_1, \mathbf{w}_t\right) \end{aligned} \tag{2.6}$$

where $c_1^* = \exp\{c_1\}$ and $\boldsymbol{\gamma}_1 = \left(\gamma_{11}, \gamma_{12}, ..., \gamma_{1,k}, c_1^*\right)'$, I obtain a new parametrization that makes the estimation numerically less volatile. In the following, the model with a linear part and a single nonlinear hidden unit is expressed as follows:

$$y_t = \boldsymbol{\alpha}'\mathbf{w}_t + \beta_1\psi(\boldsymbol{\gamma}_1, \mathbf{w}_t) + u_t, t = 1, ..., T, u_t \sim n.i.d.(0, \sigma_u^2) \tag{2.7}$$

White (1989c) proposes a two-step estimation procedure for ANN models that has asymptotic efficiency equivalent to that of the NLLS. I also rely on a two-step procedure, but its first step is quite different from that of White. It is actually based on a linearization of the nonlinear model; this entails drawing randomly the parameters of the nonlinear part of each hidden unit, taking care of the parameter bounds. In the following, the $\gamma_{ji}$s are randomly drawn by an $U[-1, 1]$ distribution and the location parameters $c_j$s from a $U[-0.05, 0.05]$. Since I do not know the magnitude of the true parameters in advance, each draw is multiplied by 1.5 to the power of $0, 1, ..., d_{MAX}$ to cover the parameter space fully. For example, if I think that $|\gamma_{ji}| < 100$ for each $i, j$, I choose $d_{MAX}$ such that $1.5^{d_{MAX}} \approx 100$.

For simplicity, I first consider an ANN model with a linear part and a single nonlinear hidden unit. The estimation procedure can then be described as follows:

- Choose $N > 1000$; draw the $N/10$ sets of random values as described above. This will generate a grid of $N$ different sets of starting-values for $\gamma_{1i}, i = 1, 2, .., k$, and $c_1$, hence $c_1^* = exp(c_1)$. Estimate $\boldsymbol{\alpha}$ and $\beta_1$ by OLS conditionally on the remaining parameters and compute the sum of squared residuals for each such regression.

- Keep $2 \leq n \leq n_{\max}$ sets of initial estimates corresponding to the $n$ lowest sums of squared residuals. The exact values are chosen by the model builder (I use $n_{\max} = 20$). For each of the $n$ sets of starting-values, re-estimate the parameters as follows:

- Rescale the parameters by dividing them by their initial estimates (this means multiplying the corresponding variables with these estimates; all starting-values are thereby set to unity). This guarantees that the parameters to be estimated are of the same magnitude, which speeds up convergence. Different magnitudes in the gradients of the linear and nonlinear parameters may lead to numerical problems too. If rescaling is not sufficient, divide the exponents of the logistic functions by a certain number. Usually the standard deviation, as suggested in Granger and Teräsvirta (1993), p.123, is appropriate, but it is often necessary to try with several values before obtaining convergence of the nonlinear optimization procedure.

- Estimate the ANN model using the BFGS algorithm. BFGS is a numerically more stable algorithm than many of those inverting the Hessian at every step. To improve accuracy, provide BFGS with the analytical first derivatives. After rescaling, numerical problems usually do not arise. If they do, or if convergence of the algorithm cannot be reached after a certain number of iterations, the nonlinear optimization procedure returns an error code, and the set of parameter estimates is discarded.

- Select the set of parameter estimates which yields the smallest value of the objective function and compute their approximate t-values using the approximate covariance matrix of the parameters as in Davidson and MacKinnon (1993), p.154:

$$\mathbf{t} = \left\{ diag \left[ \widehat{cov} \left( \left( \hat{\boldsymbol{\alpha}}, \hat{\beta}_1, \hat{\boldsymbol{\gamma}}_1 \right)' \right) \right] \right\}^{-\frac{1}{2}} \left( \hat{\boldsymbol{\alpha}}, \hat{\beta}_1, \hat{\boldsymbol{\gamma}}_1 \right)'$$

$$\text{where } \widehat{cov} \left( \left( \hat{\boldsymbol{\alpha}}, \hat{\beta}_1, \hat{\boldsymbol{\gamma}}_1 \right)' \right) \approx \sigma_u^2 GG'$$

  Approximate the covariance matrix of the estimates by the outer product of the matrix of first derivatives, where $G = (g_1, g_2, ..., g_T)$ is the score. Numerical problems often occur computing the Hessian numerically, therefore choose an approximation based on the outer product of the analytical score.

Consider now the general case of an ANN model with $q$ hidden units:

$$y_t = \boldsymbol{\alpha}' \mathbf{w}_t + \sum_{j=1}^{q} \beta_j \psi(\boldsymbol{\gamma}_j, \mathbf{w}_t) + u_t, t = 1, ..., T, u_t \sim n.i.d.(0, \sigma_u^2) \tag{2.8}$$

where $\boldsymbol{\gamma}_j = \left( \gamma_{j1}, \gamma_{j2}, ..., \gamma_{jk}, c_j^* \right)'$, $j = 1, ..., q$.

When $q > 1$, the ANN model differs from standard nonlinear regression models in that it is not globally identified. This occurs both because the hidden units are exchangeable, and that there exist $q!$ observationally equivalent models. I avoid this problem by re-ordering the hidden units after estimation according to the criterion $\beta_q > \beta_{q-1} > ... > \beta_2 > \beta_1$. See Hwang and Ding (1997) for a general discussion on identification of ANN models.

The likelihood function often tends to be rather flat; see, for example, Stinchcombe (1995), where an ANN model is represented as a linear combination of elements of approximately flat sets. The problem of interdependence among some parameters of the model will render the estimation much more difficult than in the ANN model with a single hidden unit. It is not uncommon to end up with a local maximum of the log-likelihood function instead of the global one. This is why obtaining good starting-values is important and why several sets of starting-values should be applied at the estimation stage. In the case of $q > 1$ hidden units I have to draw a larger number of starting-values than in the estimation of a model with a single hidden unit, and this has to be done in a sensible way to avoid the estimation problems discussed above. This is done in two ways:

First, define a grid of $qN/3$ different starting values for all $\gamma_{qi}, i = 1, 2, .., k$, and $c_q^*$. This entails keeping the estimates of the nonlinear parameters of the first $q-1$ hidden units fixed to their estimates. Estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by OLS conditionally on the remaining parameters and compute the sum of squared residuals of that regression.

Second, define a grid of $2qN/3$ different starting values for all $\gamma_{ji}$, and $c_j^*$, $i = 1, 2, .., k$, $j = 1, 2, ..., q$. Draw random values for the nonlinear parameters of *all* the hidden units.

Estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by OLS conditionally on the remaining parameters drawn randomly, and compute the sum of squared residuals of that regression as above. Store the $2 < n < n_{\max}$ sets of starting-values corresponding to the lowest sum of squared residuals among the $qN$ ones. Estimate all parameters by BFGS, and re-order the nonlinear hidden units in the model such that $\beta_q > \beta_{q-1} > ... > \beta_2 > \beta_1$. Check the t-values of the parameter estimates, and for each hidden unit, remove the lags corresponding to low t-values and re-estimate the model. Repeat this if necessary.

Despite several re-estimations, parameter estimates with low t-values may still be present in the model. In such case, it is useful to compute the approximated covariance matrix of the estimated parameters by the outer product of the gradient matrices. If some of them are heavily correlated, their corresponding estimated variances are very high and the t-values low. Removing any regressor can be useless; therefore, one can keep them all or remove the one with the lowest t-value. For a practical application, see the Sunspot series case.

I call this procedure **full estimation**. When $q > 1$, I may want to alleviate the computational burden by simplifying the estimation procedure. I do this by applying **frozen or conditional estimation.** Such a method entails estimating an ANN model with $q$ hidden units keeping all the parameter estimates of the $q-1$ ones from the previous rounds fixed but also re-estimating the linear part. The procedure is as follows. First, estimate a model with a linear part and a single hidden unit, and compute $\hat{v}_t^1 = y_t - \hat{\beta}_1 \psi(\hat{\boldsymbol{\gamma}}_1, \mathbf{w}_t)$. Next, estimate a model for $\hat{v}_t^1$ with a linear part and a single hidden unit, and compute $\hat{v}_t^2 = \hat{v}_t^1 - \hat{\beta}_2 \psi(\hat{\boldsymbol{\gamma}}_2, \mathbf{w}_t)$. A general step consists of estimating $\boldsymbol{\alpha}_q$ and $\beta_q$ in

$$\hat{v}_t^{q-1} = \boldsymbol{\alpha}_q' \mathbf{w}_t + \beta_q \psi(\boldsymbol{\gamma}_q, \mathbf{w}_t) + u_t^{q-1}, t = 1, ..., T.$$

The final estimates of parameters of the linear part are those obtained at the last stage. This technique is simple but suboptimal. Whether the loss of performance in forecasting is small enough to justify the use of this technique will be investigated.

### 2.3.3 Specification

#### Approximating hidden units by Taylor series expansions

Much of our ANN modelling is based on the notion of using Taylor approximations to the logistic functions that constitute the hidden units of our single hidden-layer feedforward ANN model. Consider the logistic function

$$\psi(\tilde{\boldsymbol{\gamma}}_j' \mathbf{w}_t) = \left(1 + \exp\left\{-\tilde{\boldsymbol{\gamma}}_j' \mathbf{w}_t\right\}\right)^{-1} \tag{2.9}$$

that has a convergent Taylor expansion for any parameter vector $\tilde{\boldsymbol{\gamma}}_j$ everywhere in the sample space. Assuming that $\mathbf{w}_t$ is a $(p \times 1)$ vector of regressors and expanding it into a Taylor series at the $h$th order around an arbitrary fixed point $\mathbf{w}_t^0$ in the sample space yields

$$\begin{aligned} \psi(\tilde{\boldsymbol{\gamma}}_j' \mathbf{w}_t) &= \psi(\tilde{\boldsymbol{\gamma}}_j' \mathbf{w}_t^0) + \frac{\partial \psi(\tilde{\boldsymbol{\gamma}}_j' \mathbf{w}_t)}{\partial \mathbf{w}_t'} |_{\mathbf{w}_t = \mathbf{w}_t^0} \left(\mathbf{w}_t - \mathbf{w}_t^0\right) \\ &\quad + \left(\mathbf{w}_t - \mathbf{w}_t^0\right)' \frac{\partial^2 \psi(\tilde{\boldsymbol{\gamma}}_j' \mathbf{w}_t)}{\partial \mathbf{w}_t \partial \mathbf{w}_t'} |_{\mathbf{w}_t = \mathbf{w}_t^0} \left(\mathbf{w}_t - \mathbf{w}_t^0\right) + ... + R_h\left(\tilde{\boldsymbol{\gamma}}_j, \mathbf{w}_t; \mathbf{w}_t^0\right) \end{aligned}$$

where $R_h\left(\tilde{\boldsymbol{\gamma}}_j, \mathbf{w}_t; \mathbf{w}_t^0\right)$ is the remainder of the $h$th order expansion such that $R_h\left(\tilde{\boldsymbol{\gamma}}_j, \mathbf{w}_t; \mathbf{w}_t^0\right) \to 0$ as $h \to \infty$. Merging terms of the same order and reparametrizing yields

$$\psi(\tilde{\boldsymbol{\gamma}}_j' \mathbf{w}_t) = \delta_0 + \sum_{i=1}^{p} \delta_i w_{ti} + \sum_{i=0}^{p} \sum_{l=i}^{p} \delta_{il} w_{ti} w_{tl} + ... + R_h\left(\tilde{\boldsymbol{\gamma}}_j, \mathbf{w}_t; \mathbf{w}_t^0\right) \tag{2.10}$$

It will also be useful to expand (2.6) into a Taylor series around the point $\boldsymbol{\gamma}_{j0}$=0 in the parameter space. This gives, after merging terms and reparametrizing,

$$\psi(\boldsymbol{\gamma}_j, \mathbf{w}_t) = \kappa_0 + \sum_{i=1}^{p} \kappa_i w_{ti} + \sum_{i=0}^{p} \sum_{l=i}^{p} \kappa_{il} w_{ti} w_{tl} + ... + \overline{R}_h \left(\boldsymbol{\gamma}_j, \mathbf{w}_t; 0\right) \tag{2.11}$$

where $\overline{R}_h \left(\boldsymbol{\gamma}_j, \mathbf{w}_t; 0\right)$ is another remainder of the $h$th order expansion. I shall make use of these approximations at different stages of our modelling cycle.

### Selecting the variables to the model

As discussed above, the builder of an ANN model faces two specification problems. First, the set of variables to be included in the model has to be selected. Second, the number of hidden units has to be determined. I begin with the variable selection problem. Rech, Teräsvirta, and Tschernig (2001), hereafter RTT, consider this problem in a rather general case where the functional form of the model is unknown. As a matter of fact, I have a similar situation in the sense that the ANN models may be viewed as approximations to unknown functional forms. I am also able to apply the technique of RTT to this situation. If I am interested in the first $k = p + q$ lags of the series, the problem is to find the correct lag vector $\mathbf{w}_t^{(IN)} = \left(y_{t-i_1}, ..., y_{t-i_p}\right)'$ from the set $\mathcal{Y} = \{y_{t-j} : j = 1, ..., p+q\}$. Let $\mathbf{w}_t^{(OUT)} = \left(y_{t-i_{p+1}}, ..., y_{t-i_{p+q}}\right)'$ be the vector of the $q$ remaining lags. Disregarding the ordering in the vector, I can then rewrite $\mathbf{w}_t = \left(\mathbf{w}_t^{(IN)\prime}, \mathbf{w}_t^{(OUT)\prime}, 1\right)'$. Then, applying (2.11) and merging terms, (2.8) can be written as

$$\begin{aligned} f(\mathbf{w}_t; \theta) \;=\; & \beta_0 + \sum_{j=1}^{p} \beta_j w_{tj}^{(IN)} + \sum_{j=1}^{q} \gamma_j w_{tj}^{(OUT)} + \sum_{j_1=1}^{p} \sum_{j_2=j_1}^{p} \beta_{j_1,j_2} w_{t,j_1}^{(IN)} w_{t,j_2}^{(IN)} \\ & + \sum_{j_1=1}^{q} \sum_{j_2=j_1}^{q} \gamma_{j_1,j_2} w_{t,j_1}^{(OUT)} w_{t,j_2}^{(OUT)} + \sum_{j_1=1}^{p} \sum_{j_2=1}^{q} \delta_{j_1,j_2} w_{t,j_1}^{(IN)} w_{t,j_2}^{(OUT)} \\ & + \sum_{j_1=1}^{p} \sum_{j_2=j_1}^{p} \sum_{j_3=j_2}^{p} \beta_{j_1,j_2,j_3} w_{t,j_1}^{(IN)} w_{t,j_2}^{(IN)} w_{t,j_3}^{(IN)} + ... \\ & + \sum_{j_1=1}^{q} \sum_{j_2=j_1}^{q} \cdots \sum_{j_k=j_{k-1}}^{q} \gamma_{j_1,...,j_k} w_{t,j_1}^{(OUT)} w_{t,j_2}^{(OUT)} \cdots w_{t,j_{k-1}}^{(OUT)} w_{t,j_k}^{(OUT)} \\ & + R_h \left(\mathbf{w}_t^{(IN)}, \mathbf{w}_t^{(OUT)}\right) + u_t \end{aligned} \tag{2.12}$$

where $q - p \leq h$, the order of the expansion (for notational reasons, not being a restriction), $R_h \left(\mathbf{w}_t^{(IN)}, \mathbf{w}_t^{(OUT)}\right)$ is the remainder, and the $\beta$'s, $\gamma$'s, and $\delta$'s are parameters. Expansion (2.12) contains all possible combinations of lags from 1 to $p+q$ up to order $h$. The idea of RTT is that the terms involving elements of $\mathbf{w}_t^{(OUT)}$ in (2.12) are redundant and thus have zero coefficients. This fact is used to omit the redundant variables from the model. As expansion (2.12) is not a function of the number of hidden units in (2.8), the variables may be selected before tackling the problem of choosing the number of hidden units. The variable selection procedure works as follows. First regress $y_t$ on all variables (products of lags from 1 to $p+q$) in the Taylor expansion and compute the value of an appropriate model selection criterion. I may use SBIC (Rissanen (1978), Schwarz (1978)) or AIC (Akaike (1974)). Next leave one

lag out from the original model, regress $y_t$ on all products of remaining lags in the Taylor expansion and compute the value of the criterion. Repeat this by omitting each lag in turn. Continue by simultaneously omitting two lags from the original model. Proceed until the regression only consists of a Taylor expansion of a function with a single lag. Leave this out as well to check if $y_t$ is just white noise. This amounts to estimating $\sum_{i=1}^{p+q}\binom{p+q}{i}+1 = 2^{p+q}$ linear models by ordinary least squares. The combination of lags that yields the lowest value of the model selection criterion is selected. Of the two criteria SBIC is consistent in the sense that if there is a finite set of correct lags and if the true model is in fact of type (2.12) with a zero remainder, the correct lags will asymptotically be selected with probability one. On the other hand, by using AIC I may avoid underfitting if the sample size is small and/or the signal-to-noise ratio very low. RTT show that the procedure works well already in small samples and can be applied even in large samples where the nonparametric model selection techniques become computationally infeasible.

To select variables with certain accuracy, the order $h$ has to be at least 3 unless I assume that the unknown nonlinear function is even. If one assumes nonlinear additivity, the crossed terms in the Taylor expansion have no power and the number of regressors can be lowered to $3k$, but in general this is too strong a restriction. Keeping $h = 3$, the number of terms of the Taylor expansion becomes very large when $p + q$ increases. Such problem arises whenever $T$ is not large enough compared to $k$, hence $\{m(k) + n(k)\}/T$ is too high. For instance, when $k = 12$ the number of regressors is $12 + 12 \times 13/2 + 12 \times 13 \times 14/3! = 454$, but when dealing with monthly data, one has to encompass the first 12 lags at least. I therefore suggest the following approximated procedure. Supposing the maximum number of lags which can be encompassed is $k_{MAX_1} \leq k_{MAX}$ , where $k_{MAX}$ is the maximum lag I am interested in investigating. Run the procedure among the first $k_{MAX_1}$ lags and select a set of $p_1$ lags $y_{MAX_1,1}, ..., y_{MAX_1,p_1}$. Therefore, $q_1 = k_{MAX_1} - p_1$ lags were discarded. I fix such $q_1$ lags and set $k_{MAX_2} = k_{MAX_2} + q_1$. Since $q_1$ lags are discarded in advance, the number of lags to choose from will still be $k_{MAX_1}$. The procedure is repeated and I call the new set of selected lags $p_2$, while $q_2 = k_{MAX_2} - p_2$ lags will be discarded in turn. Again, I set $k_{MAX_3} = k_{MAX_2} + q_2$ and continue. The procedure is stopped either when $k_{MAX_{j+1}} = k_{MAX_j}$ (no additional lags can be encompassed) or $k_{MAX_{j-1}} = k_{MAX}$.

### Choosing the number of hidden units

After choosing the regressors for the model, I have to choose the number of hidden units. I assume for notational simplicity that the first $k$ lags have been selected. The ANN model with a linear part and $q + Q$ hidden units thus has the form

$$\begin{aligned} y_t &= \boldsymbol{\alpha}'\mathbf{w}_t + \sum_{j=1}^{q}\beta_j\psi(\boldsymbol{\gamma}_j, \mathbf{w}_t) + \sum_{j=q+1}^{Q}\beta_j\psi(\boldsymbol{\gamma}_j, \mathbf{w}_t) + u_t, \qquad (2.13) \\ t &= 1, ..., T, u_t \sim n.i.d.(0, \sigma_u^2) \end{aligned}$$

where $\mathbf{w}_t = (y_{t-1}, y_{t-2}, ..., y_{t-k}, 1)'$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_k, \alpha_0)'$, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_{q+Q})'$, $\boldsymbol{\gamma}_j = \left(\gamma_{j1}, \gamma_{j2}, ..., \gamma_{j,k}, c_j^*\right)'$, $j = 1, ..., q + Q$. The null hypothesis of $q$ hidden units is that any logistic function which I possibly add to the model is never activated: $H_0 : \gamma_j = 0$, $j = q + 1, ..., Q$. The alternative hypothesis is $H_1 : \gamma_j > 0$ for at least one $j$, $j = q + 1, ..., Q$. I have an identification problem because the ANN model (2.13) is only identified under $H_1$. Under $H_0$, $\beta_j$, $\gamma_{ji}$, and $c_j^*$, $j = q + 1, ..., Q$, $i = 1, ..., k$ can take any value without the value of the log-likelihood being affected. Following Teräsvirta and Lin (1993), I solve this

problem by expanding $\psi(\boldsymbol{\gamma}_j, \mathbf{w}_t)$, $j = q+1, ..., Q$, into a Taylor series around the point $\boldsymbol{\gamma}_j$=0, $j = q+1, ..., Q$ at the third order as in (2.11). Using $(Q-q)$ third-order Taylor expansions, and multiplying them by the $(Q-q)$ coefficients of the logistic functions $\beta_j$, $j = q+1, ..., Q$, I get an overall third-order expansion, and I call the coefficients of the linear, quadratic and cubic part respectively $\lambda_i, i = 1, ..., k$; $\lambda_{ij}, i = 1, ..., k; j = i, ..., k$; $\lambda_{ijl}, i = 1, ..., k; j = i, ..., k; l = j, ..., k$. This gives:

$$\begin{aligned} y_t &= \boldsymbol{\alpha}^{*\prime}\mathbf{w}_t + \sum_{j=1}^{q}\beta_j\psi(\boldsymbol{\gamma}_j, \mathbf{w}_t) + \sum_{i=1}^{k}\sum_{j=i}^{k}\lambda_{ij}w_{ti}w_{tj} \\ &\quad + \sum_{i=1}^{k}\sum_{j=i}^{k}\sum_{l=j}^{k}\lambda_{ijl}w_{ti}w_{tj}w_{tl} + u_t^* \end{aligned} \tag{2.14}$$

where $\alpha_i^* = \alpha_i + \lambda_i$, $u_t^* = R_3(\mathbf{w}_t; 0)$, where $R_3(\mathbf{w}_t; 0)$ is the rest of the overall third-order Taylor expansion of the $(Q-q)$ hidden units. The null hypothesis becomes

$$H_0^{'}: \begin{cases} \lambda_{ij} = 0, i = 1, ..., k; j = i, ..., k \\ \lambda_{ijl} = 0, i = 1, ..., k; j = i, ..., k; l = j, ..., k \end{cases} \tag{2.15}$$

I use the Lagrange Multiplier principle to derive a test for (2.15). Under the null hypothesis $\{u_t^*\}$ is a sequence of normal independent random variables because in that case $u_t^* = u_t$. The approximated conditional log-likelihood for observation $t$ is:

$$l_t(\cdot) = c - \frac{1}{2}\ln\sigma^2$$

$$-\frac{1}{2\sigma^2}\left(y_t - \boldsymbol{\alpha}^{*\prime}\mathbf{w}_t - \sum_{j=1}^{q}\beta_j\psi(\boldsymbol{\gamma}_j, \mathbf{w}_t) - \sum_{i=1}^{k}\sum_{j=i}^{k}\lambda_{ij}w_{ti}w_{tj} - \sum_{i=1}^{k}\sum_{j=i}^{k}\sum_{l=j}^{k}\lambda_{ijl}w_{ti}w_{tj}w_{tl}\right)^2$$

The residuals estimated under $H_0^{'}$ are $\widehat{v}_t = y_t - \hat{\boldsymbol{\alpha}}^{*\prime}\mathbf{w}_t - \sum_{j=1}^{q}\hat{\beta}_j\psi(\hat{\boldsymbol{\gamma}}_j, \mathbf{w}_t)$, where $\hat{\boldsymbol{\alpha}}^*$, $\hat{\beta}_j$, $j = 1, .., q$, and $\hat{\boldsymbol{\gamma}}_j$, $j = 1, .., q$ are the maximum likelihood estimates of the corresponding parameters under $H_0$. Since the information matrix is block diagonal with respect to the second derivative of $\sigma^2$, I can treat $\sigma^2$ as fixed. The resulting $TR^2$ form of the LM test based on the auxiliary regression:

$$\begin{aligned} \widehat{v}_t &= \boldsymbol{\alpha}^{*\prime}\mathbf{w}_t + \sum_{j=1}^{q}\beta_j\psi(\boldsymbol{\gamma}_j, \mathbf{w}_t) + \sum_{j=1}^{q}\sum_{i=1}^{k}\delta_{ji}\partial\psi(\boldsymbol{\gamma}_{ji}, \mathbf{w}_t)/\partial\boldsymbol{\gamma}_{ji} + \sum_{i=1}^{k}\sum_{j=i}^{k}\lambda_{ij}w_{ti}w_{tj} \\ &\quad + \sum_{i=1}^{k}\sum_{j=i}^{k}\sum_{l=j}^{k}\lambda_{ijl}w_{ti}w_{tj}w_{tl} + v_t^1 \end{aligned} \tag{2.16}$$

The test can be performed in three stages:

(i) estimate the parameters of model (2.14) by conditional maximum likelihood, and compute the estimated residuals $\widehat{v}_t$ and the corresponding sum of the squared residuals $SSR_0 = \sum\widehat{v}_t^2$;

(ii) regress $\widehat{v}_t$ on $\mathbf{w}_t$, $\psi(\widehat{\boldsymbol{\gamma}}_j, \mathbf{w}_t)$, $\left(\partial\psi(\boldsymbol{\gamma}_j, \mathbf{w}_t)/\partial\boldsymbol{\gamma}_j^{\prime}\right)_{\boldsymbol{\gamma}_j=\widehat{\boldsymbol{\gamma}}_j}$, $j = 1, ..., q$, $w_{ti}w_{tj}$, $i = 1, ..., k; j = i, ..., k$; $w_{ti}w_{tj}w_{tl}$, $i = 1, ..., k; j = i, ..., k; l = j, ..., k$; and compute the residuals $\widehat{v}_t^{\prime}$ and the corresponding sum of squared residuals $SSR = \sum\widehat{v}_t^{\prime 2}$;

(iii) compute the test statistic

$$F = \frac{(SSR_0 - SSR)/m}{SSR/(T-p-1-m)}$$

where $m$ is the number of second and third-order components in the auxiliary regression. Under the null, $F$ has approximately $F_{m,T-p-1-m}$ distribution, see e.g. Harvey (1990), p. 174-175. I use the $F$ distribution since in small samples the actual size of the $F$-test is closer to the nominal one than that of the corresponding asymptotic $\chi^2$ test. As discussed previously, the estimation of the ANN models may sometimes be difficult due to the large number of parameters, and the estimation algorithm may not converge properly. The estimated residuals $\widehat{v}_t$ may not, therefore, be fully orthogonal to the gradient matrix, which causes size distortion in the test. To remedy the problem, I regress the residuals on the gradient before carrying out the test. For the exact expression of the gradient, see Appendix A. This entails the following additional step:

(i') regress $\widehat{v}_t$ on $\mathbf{w}_t$, $\psi(\widehat{\boldsymbol{\gamma}}_j, \mathbf{w}_t)$, $\left(\partial\psi(\boldsymbol{\gamma}_j, \mathbf{w}_t)/\partial\boldsymbol{\gamma}_j\right)'_{\boldsymbol{\gamma}_j=\widehat{\boldsymbol{\gamma}}_j}$, $j = 1, ..., q$, compute the residuals $\widetilde{v}_t$ and the residual sum of squares, $SSR_0' = \sum_{t=1}^{T} \widetilde{v}_t^2$.

The OLS residuals $\widetilde{v}_t$ are orthogonal to the regressors. I perform (ii) and (iii) using $\widetilde{v}_t$ in place of $\widehat{v}_t$.

In some special cases, numerical problems may arise from high correlation among the columns of the regression matrix involving $w_{ti}w_{tj}$, $w_{ti}w_{tj}w_{tl}$, and the partial derivatives $\left(\partial\psi(\boldsymbol{\gamma}_j, \mathbf{w}_t)/\partial\boldsymbol{\gamma}_j\right)$. In addition, if in some estimated hidden unit $\psi(\widehat{\boldsymbol{\gamma}}_j, \mathbf{w}_t)$ any estimate of $\gamma_{ji}, i = 1, ..., k$ is large, the estimated logistic function is close to being a step function; its corresponding first derivatives will then be mostly constant for $t = 1, ..., T$. The solution advocated by Eitrheim and Teräsvirta (1996) and adopted here is to omit those regressors, which can be done practically without affecting the size of the test.

### Specification strategy

Summing up, I describe how to build an ANN model by using the **full** estimation method:

- The set of variables to be included in the model is selected as discussed in section 3.3.2. Given these variables, the hypothesis of no nonlinear hidden units (linear model) is tested at a given significance level $\alpha$. If rejected, a model with the linear part and one hidden unit is estimated and the approximate t-values $\mathbf{t}$ of the parameter estimates computed. The lags that have coefficient estimates with low t-values are removed and the model re-estimated. This can be done several times. If eventually, after removing all the lags but lag $i$, either $\hat{\beta}_1$ or $\hat{\gamma}_{1i}$ or both are not significant, I retain lag $i$ and the hidden unit anyway, because linearity was originally rejected.

- The hypothesis of no additional hidden units is tested at the significance level $\alpha/2$, by using the $TR^2$ form of the F-test derived by (2.16). If it is rejected, an ANN model with two hidden units is estimated, and the size of the model is reduced by checking the t-values of its estimates as above. However, at least one lag is retained in each hidden unit.

- The previous step is repeated halving the significance level again to $\alpha/4$, $\alpha/8$, ...; the procedure is discontinued at the first acceptance of the null hypothesis of no additional hidden units.

Letting the significance level of the hidden unit test converge to zero as $q \to \infty$ keeps the dimension of the model under control. It is not possible to compute analytically the overall size $\alpha^*$ of the testing procedure. The Bonferroni upper bound for $q$ steps is

$$\alpha^* \leq 1 - \prod_{i=1}^{q} (1 - \alpha/2^{i-1}) \tag{2.17}$$

For instance, for $\alpha = 0.10$, a numerical evaluation of (2.17) when $q \to \infty$ gives $\alpha^* \leq 0.187$.

If I use **frozen** estimation, the parameters of each hidden unit are estimated separately, conditionally on the previous results as described at the end of Section 3.2. As above, the significance levels are halved at each step, and the procedure ends at the first acceptance of the null hypothesis. There are indeed two important differences:

- in testing for an additional hidden unit, I approximate it by the third-order Taylor expansion terms of the lags included in the previous hidden unit *only*;
- the misspecification tests for no serial correlation, no ARCH and parameter constancy will be conditional to previous hidden units estimates.

Choosing the significance level for discarding variables is an important issue. In econometrics, we talk of parameter estimates which are significant at the 1%, 5% or 10% level. Here, the issue is very much influenced by the sample size, which is often small, and by estimation difficulties. It is unlikely that all the t-values of its components computed in one point of the parameter space $\Theta$ are greater than for another. In such situations I can only talk of t-values conditioned to a certain estimate $\hat{\theta} = \hat{\theta}_{GM}$, because the estimation procedure is based on maximazing the log-likelihood. Once $\hat{\theta}_{GM}$ is fixed, I remove the parameters whose estimates are lower compared to the whole vector $\hat{\theta}_{GM}$. Removing all the parameters, and therefore the whole hidden unit, is nonsense because rejecting the null hypoyhesis of no additional hidden units indicates that there is nonlinearity left to be modeled.

### 2.3.4 Evaluation of the model

After the ANN model has been estimated it has to be evaluated. This means that the assumptions under which the model has been estimated have to be checked. They include the assumption of no serial correlation in the error process and that of constant parameters. The assumption of constant conditional variance has to be tested as well. In Refenes and Zapranis (1999), the authors recently developed a three-step methodology for ANN modelling. It consists of model selection, variable significance testing and model adequacy testing. Being the last step based on residual testing, in such paper (p.304), the authors argue that "any of the conventional tests for randomness can be applied here, e.g. Box-Pierce, Ljung-Box, Durbin-Watson, F-test". As discussed in Eitrheim and Teräsvirta (1996), Ljung-Box is inapplicable to STAR models because its asymptotic properties are unknown if the test is based on residuals of a nonlinear model. The same problem arises with the residuals of ANN models, but statistical tests of no autocorrelation and parameter constancy in the ANN framework may be derived following the ideas in Eitrheim and Teräsvirta (1996). The alternative to parameter constancy then, is that parameters change smoothly over time. An important special case, a sudden change or structural break, is nested in this alternative hypothesis. The test concerns the intercept, the coefficients in the linear unit and the coefficients of the hidden units. The parameters of the logistic functions could also be

subjected to test. The possible variation in a logistic function is, however, "second-order variation" in the sense that would be more difficult to detect and, if it is not dramatic, would generally affect the outcome less than changes in the "connection strengths" coefficients in front of the logistic functions. Such tests will be considered in the following subsections.

### Test of serial independence

The evaluation stage consists of checking the residuals of the model and considering its long-term properties. The hypothesis of no error autocorrelation can be tested as discussed in Godfrey (1988), p.116-117, and Eitrheim and Teräsvirta (1996). It should be noted that the standard portmanteau test of Ljung and Box (1978) is not applicable because its asymptotic null distribution is unknown if the true model is a nonlinear ANN one.

I proceed as in Eitrheim and Teräsvirta (1996) and consider the following nonlinear autoregressive model with autocorrelated errors:

$$y_t = G(\mathbf{w}_t; \mathbf{\Phi}) + u_t \tag{2.18}$$

where $G(\mathbf{w}_t; \mathbf{\Phi})$ is at least twice continuously differentiable (this always holds for ANN models) and

$$u_t = \mathbf{a}'\mathbf{v}_t + \varepsilon_t = \left(\sum_{j=1}^{r} a_j L^j\right) u_t + \varepsilon_t, \ \varepsilon_t \sim n.i.d.(0, \sigma^2)$$

where $L$ is the lag operator, $\mathbf{w}_t = (y_{t-1}, ..., y_{t-k}, 1)'$ is defined as above, $\mathbf{\Phi}$ is the set of parameters of the nonlinear function, $\mathbf{v}_t = (u_{t-1}, ..., u_{t-r})'$ and $\mathbf{a} = (a_1, ..., a_r)'$, $a_r \neq 0$. Assume that the roots of $1 - \sum_{j=1}^{r} a_j z^j = 0$ lie outside the unit circle, and, under the assumption $u_t \sim n.i.d.(0, \sigma^2)$, $\{y_t\}$ is stationary and ergodic such that the parameters of (2.18) can be consistently estimated by NLLS. Thus the hypotesis of serial independence of errors $\{u_t\}$ in (2.18) is equivalent to $\mathbf{a} = 0$. Left-multiplying (2.18) by $1 - \sum_{j=1}^{r} a_j L^j$ yields:

$$y_t = \sum_{j=1}^{r} a_j L^j y_t + G(\mathbf{w}_t; \mathbf{\Phi}) - \sum_{j=1}^{r} a_j G\left(L^j \mathbf{w}_t; \mathbf{\Phi}\right) + \varepsilon_t$$

Assuming the necessary starting values $(y_0, y_{-1}, ..., y_{-(k-1)})$ fixed and reparametrizing with $\mathbf{w}_t$, the conditional log-likelihood for observation $t$, where $t = 1, ..., T$ is

$$\begin{aligned} l_t &= c - \frac{1}{2}\ln\sigma^2 - \frac{\varepsilon_t^2}{2\sigma^2} \\ &= c - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left\{y_t - \mathbf{a}'\mathbf{w}_t - G(\mathbf{w}_t; \mathbf{\Phi}) + \sum_{j=1}^{r} a_j G(\mathbf{w}_{t-j}; \mathbf{\Phi})\right\}^2 \end{aligned}$$

Since the information matrix is block diagonal, I can treat $\sigma^2$ as fixed when deriving $l_t$:

$$\frac{\partial l_t}{\partial a_j} = \frac{\varepsilon_t}{\sigma^2}\left\{w_{t-j} - G(\mathbf{w}_{t-j}; \mathbf{\Phi})\right\}, j = 1, ..., r \tag{2.19}$$

$$\frac{\partial l_t}{\partial \mathbf{\Phi}} = -\frac{\varepsilon_t}{\sigma^2}\left\{\frac{\partial G(\mathbf{w}_t; \mathbf{\Phi})}{\partial \mathbf{\Phi}} - \sum_{j=1}^{r} a_j \frac{\partial G(\mathbf{w}_{t-j}; \mathbf{\Phi})}{\partial \mathbf{\Phi}}\right\} \tag{2.20}$$

Setting $\widehat{\mathbf{v}}_t = (\widehat{u}_{t-1}, ..., \widehat{u}_{t-r})'$, $\widehat{u}_{t-j} = w_{t-j} - G\left(\mathbf{w}_{t-j}; \widehat{\mathbf{\Phi}}\right), j = 0, 1, ..., r$, where $\widehat{\mathbf{\Phi}}$ is the maximum likelihood estimator of $\mathbf{\Phi}$, and $\widehat{\mathbf{z}}_t = \left(\partial G\left(\mathbf{w}_t; \mathbf{\Phi}\right)/\partial \mathbf{\Phi}\right)_{\mathbf{\Phi}=\widehat{\mathbf{\Phi}}}$, I can write (2.19) and (2.20) under the null hypothesis:

$$\left(\frac{\partial \widehat{l}_t}{\partial \mathbf{a}}\right)_{H_0} = \frac{1}{\sigma^2}\widehat{u}_t\widehat{\mathbf{v}}_t; \quad \left(\frac{\partial \widehat{l}_t}{\partial \mathbf{\Phi}}\right)_{H_0} = -\frac{1}{\sigma^2}\widehat{u}_t\widehat{\mathbf{z}}_t$$

In the ANN model case as defined in (3.5) with logistic activation functions,

$$G\left(\mathbf{w}_t; \mathbf{\Phi}\right) = G\left(\mathbf{w}_t; \beta, \boldsymbol{\alpha}; \boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_q\right) = \boldsymbol{\alpha}'\mathbf{w}_t + \sum_{j=1}^{q} \beta_j \psi(\boldsymbol{\gamma}_j, \mathbf{w}_t)$$

and I need to compute $\widehat{\mathbf{z}}_t = \left(\partial G\left(\mathbf{w}_t; \mathbf{\Phi}\right)/\partial \mathbf{\Phi}\right)_{H_0} = \left(\partial G\left(\mathbf{w}_t; \mathbf{\Phi}\right)/\partial \mathbf{\Phi}\right)_{\mathbf{\Phi}=\widehat{\mathbf{\Phi}}}$:

$$\left(\frac{\partial G\left(\mathbf{w}_t; \mathbf{\Phi}\right)}{\partial \boldsymbol{\alpha}'}\right)_{H_0} = \mathbf{w}_t; \quad \left(\frac{\partial G\left(\mathbf{w}_t; \mathbf{\Phi}\right)}{\partial \beta_j}\right)_{H_0} = \psi(\widehat{\boldsymbol{\gamma}}_j, \mathbf{w}_t);$$

$$\left(\frac{\partial G\left(\mathbf{w}_t; \mathbf{\Phi}\right)}{\partial \boldsymbol{\gamma}_j}\right)_{H_0} = -\widehat{\beta}_j \left(\frac{\partial \psi(\boldsymbol{\gamma}_j, \mathbf{w}_t)}{\partial \boldsymbol{\gamma}_j}\right)_{H_0}; j = 1, ..., q$$

The exact expression of the derivative $\left(\partial \psi(\boldsymbol{\gamma}_j, \mathbf{w}_t)/\partial \boldsymbol{\gamma}_j\right)$ can be found in Appendix A. Numerical problems for $\left(\partial \psi(\boldsymbol{\gamma}_j, \mathbf{w}_t)/\partial \boldsymbol{\gamma}_j\right)_{\boldsymbol{\gamma}_j=\widehat{\boldsymbol{\gamma}}_j}$ can arise as in the test for an additional hidden unit. Dropping the corresponding regressors which are highly correlated avoids singularity in the matrix $(1/T)\sum_{t=1}^{T} \widehat{\mathbf{z}}_t \widehat{\mathbf{z}}_t'$ while having a negligible effect on the values of the test statistic.

The corresponding $LM$ statistic for the hypotesis $H_0 : \mathbf{a} = 0$ is a test for serial independence both against $AR(r)$ and $MA(r)$ error processes as explained in Godfrey (1988), p.115. The test can be carried out in three stages as follows:

(i) Estimate the ANN model under the assumption of uncorrelated errors and compute the residual sum of squares $SSR_0 = \sum_{t=1}^{T} \widehat{u}_t^2$.

(ii) Regress $\widehat{u}_t$ on $\widehat{\mathbf{v}}_t$, and the gradient vector $\widehat{\mathbf{z}}_t = \left(\mathbf{w}_t, \psi(\widehat{\boldsymbol{\gamma}}_j, \mathbf{w}_t), \left(\partial \psi(\boldsymbol{\gamma}_j, \mathbf{w}_t)/\partial \boldsymbol{\gamma}_j\right)_{\boldsymbol{\gamma}_j=\widehat{\boldsymbol{\gamma}}_j}\right)'$, $j = 1, ..., q$ and compute the residual sum of squares, $SSR$.

(iii) Compute the test statistic $F_{LM} = \left\{\left(SSR_0 - SSR\right)/r\right\} / \left\{SSR/\left(T - n - r\right)\right\}$, where $n$ is the dimension of the gradient vector $\widehat{\mathbf{z}}_t$.

I use an $F$ distribution for the reasons outlined in section 3.3.3. As I did in such section, size distortion can be avoided by orthogonalizing the residuals before carring out the test. Thus I regress $\widehat{u}_t$ on the gradient $\mathbf{w}_t$, $\psi(\widehat{\boldsymbol{\gamma}}_j, \mathbf{w}_t)$, $\left(\partial \psi(\boldsymbol{\gamma}_j, \mathbf{w}_t)/\partial \boldsymbol{\gamma}_j\right)_{\boldsymbol{\gamma}_j=\widehat{\boldsymbol{\gamma}}_j}$, $j = 1, ..., q$, compute the residuals $\widetilde{u}_t$ and the residual sum of squares, $SSR_0' = \sum_{t=1}^{T} \widetilde{u}_t^2$.

The new OLS residuals $\widetilde{u}_t$ are orthogonal to the regressors. I perform (ii) and (iii) using $\widetilde{u}_t$ in place of $\widehat{u}_t$.

**Test of parameter constancy**

In this section, I also follow Eitrheim and Teräsvirta (1996). In model (2.8), I assume that the hidden units have constant parameters whereas both $\beta$s and $\alpha$s may change smoothly over time. Consider the ANN model

$$y_t = \boldsymbol{\alpha}(t)' \mathbf{w}_t + \sum_{j=1}^{q} \beta_j(t)\, \psi(\boldsymbol{\gamma}_j, \mathbf{w}_t) + u_t, t = 1, ..., T, u_t \sim n.i.d.(0, \sigma_u^2) \tag{2.21}$$

where $\boldsymbol{\beta}(t) = \beta_0 + \boldsymbol{\lambda}_1 F_j(t, \gamma_1, c_1)$, where $\boldsymbol{\lambda}_1$ is a $((q+1) \times 1)$ vector and $\boldsymbol{\alpha} = \boldsymbol{\alpha}(t) = \alpha_0 + \boldsymbol{\lambda}_2 F_j(t, \gamma_1, c_1)$, where $\boldsymbol{\lambda}_2$ is a $((\dim(\mathbf{w}_t) - 1) \times 1)$ vector. The null hypothesis of parameter constancy in (2.8) arises when the function $F_j$ does not change over time, $H_0 : F_j(t, \gamma_1, c_1) \equiv$ constant. I consider the following three functional forms for $F_j$:

$$F_1(t, \gamma_1, c_1) = (1 + \exp\{-\gamma_1 (t - c_1)\})^{-1} - 1/2 \tag{2.22}$$

$$F_2(t, \gamma_1, c_1) = 1 - \exp\left\{-\gamma_1 (t - c_1)^2\right\} \tag{2.23}$$

$$F_3(t, \gamma_1, \underline{c}_1) = (1 + \exp\{-\gamma_1 (t - c_1)(t - c_2)(t - c_3)\})^{-1} - 1/2 \tag{2.24}$$

Subtracting $1/2$ makes the functions (2.22) and (2.24) equal to zero under the null hypothesis $\gamma_1 = 0$ (the alternative is $\gamma_1 > 0$). $F_1$, $F_2$, $F_3$ correspond respectively to a monotonic change, a symmetric nonmonotonic one and a possibly nonmonotonic, nonsymmetric change. I derive a test statistic for the most general case, hence I'll test the hypothesis of constant parameters against the functional form $F_3$. I consider model (2.24). To circumvent the identification problem, I take the first order Taylor expansion of (2.24) about $\gamma_1 = 0$:

$$T_{13}(t, \gamma_1, \underline{c}_1) = \frac{1}{4}\gamma_1 \left(t^3 + c_{12}t^2 + c_{11}t + c_{10}\right) + R(t, \gamma_1, \underline{c}_1) \tag{2.25}$$

and approximate $\boldsymbol{\beta}(t)$ and $\boldsymbol{\alpha}(t)$ using (2.25). This yields, after a reparametrization:

$$\begin{aligned} y_t &= \boldsymbol{\pi}_0' \mathbf{w}_t + \boldsymbol{\pi}_1' t \mathbf{w}_t + \boldsymbol{\pi}_2' t^2 \mathbf{w}_t + \boldsymbol{\pi}_3' t^3 \mathbf{w}_t \\ &\quad + \sum_{j=1}^{q} \left(\vartheta_{0j} + \vartheta_{1j} t + \vartheta_{2j} t^2 + \vartheta_{3j} t^3\right) \psi(\boldsymbol{\gamma}_j, \mathbf{w}_t) + u_t^* \\ &= \sum_{h=0}^{3} \boldsymbol{\pi}_h' \left(t^h \mathbf{w}_t\right) + \sum_{h=0}^{3} \sum_{j=1}^{q} \left(\vartheta_{hj} t^h \psi(\boldsymbol{\gamma}_j, \mathbf{w}_t)\right) + u_t^* \end{aligned}$$

where $\boldsymbol{\pi}_h = (\pi_{h1}, \pi_{h2}, ..., \pi_{hk})'$, $u_t^* = u_t + R_1(t, \gamma_1, \underline{c}_1)$. Under $H_0$, $u_t^* = u_t$ and $\boldsymbol{\pi}_h = 0$; $\vartheta_{hj} = 0, h = 1, .., 3; j = 1, ..., q$. Neglecting $R_1(\cdot)$, the local approximation to the pseudo log-likelihood function in a neighbourhood of $H_0$ for each observation $t$ is:

$$l_t = const - \frac{1}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \left\{ y_t - \sum_{h=0}^{3} \boldsymbol{\pi}_h' \left(t^h \mathbf{w}_t\right) - \sum_{j=1}^{q} \sum_{h=0}^{3} \left(\vartheta_{jh} t^h \psi(\boldsymbol{\gamma}_j, \mathbf{w}_t)\right) \right\}^2$$

I can treat $\sigma^2$ as fixed and compute the partial derivatives:

$$\frac{\partial l_t}{\partial \boldsymbol{\pi}_j} = \frac{1}{\sigma^2} u_t t^h \mathbf{w}_t; \; \frac{\partial l_t}{\partial \vartheta_{jh}} = \frac{1}{\sigma^2} u_t t^h \psi(\boldsymbol{\gamma}_j, \mathbf{w}_t); \; h = 0, 1, 2, 3; \; j = 1, ..., q$$

$$\begin{aligned}\frac{\partial l_t}{\partial \gamma_j} &= \frac{1}{\sigma^2}u_t\sum_{h=0}^{3}\vartheta_{jh}t^h\frac{\psi(\gamma_j,\mathbf{w}_t)}{\partial\gamma_j}\\ &= \frac{1}{\sigma^2}u_t\sum_{h=0}^{3}\vartheta_{jh}t^h\frac{\psi(\gamma_j,\mathbf{w}_t)}{\partial\gamma_j},\ j=1,...,q\end{aligned}$$

Under the null hypothesis all the terms involving powers of $t$ have zero coefficients. Using an $LM$ test (hereafter $LM_j$ depending on the functional form $H_j$) in a similar way than in testing for serial independence:

$$\widehat{\mathbf{z}}_t = \left(\mathbf{w}_t', \left(\psi(\widehat{\gamma}_1,\mathbf{w}_t),...,\psi(\widehat{\gamma}_q,\mathbf{w}_t)\right)', \left(\frac{\partial\psi}{\partial\gamma_1},...,\frac{\partial\psi}{\partial\gamma_q}\right)'_{\gamma=\widehat{\gamma}}\right)'$$

$$\widehat{\mathbf{v}}_t = \left(t\widehat{\mathbf{z}}_t', t^2\widehat{\mathbf{z}}_t', t^3\widehat{\mathbf{z}}_t'\right)'$$

$\widehat{\mathbf{z}}_t$ and $\widehat{\mathbf{v}}_t$ have respectively $(k+1)+q+q(k+1) = 1+k+q(k+2)$ and $3+3k+3q(k+2)$ columns. The components are trending, but Lin and Teräsvirta (1994) proved that the asymptotic distribution of the $LM$ test under the null hypothesis may be derived under the condition that $Eu_t^6 < \infty$. The same result holds here as the hidden units $\psi(\gamma_1,\mathbf{w}_t)$ and their derivatives are bounded. Under $H_0$, the $LM_3$ statistic is asymptotically distributed as a $\chi^2$ with $3k(q+1)$ degrees of freedom. The corresponding $F$ statistic has $3k(q+1)$ and $T-4(1+k+q(k+2))$ degrees of freedom. Whether I consider functions $F_1$ and $F_2$, hence the $LM_1$ and $LM_2$ tests, I do not include them in the alternative terms involving $\left(t^3,t^2\right)'$ and $t^3$ respectively. Likewise, the constancy of a subset of linear parameters may be tested using different specifications of the alternative hypothesis. As previously discussed, for all the tests, I omit any regressor involving any derivative with respect to $\gamma_{ji}$ whenever it does not carry useful information.

## 2.4 Monte-Carlo study

In essay II, I have discussed asymptotic theory for inference in ANN models. For applied work, it is necessary to find out how the theory works in small samples and how performances improve as the sample size grows. To this end, I have carried out a simulation study with $T = 200$ and $T = 1000$. The random generator of the standard normal distribution in GAUSS 3.2.13 was used. The first 400 observations are discarded to eliminate the initialization effects. The DGP is an ANN with two hidden units and a linear part, time-varying parameters and autocorrelated errors:

$$\begin{aligned}y_t &= H_1(t)(0.07+0.2y_{t-1}-0.1y_{t-2})\\ &\quad +0.53H_2(t)\left(1+\exp\{-(17y_{t-1}-30.8y_{t-2}-1.19)\}\right)^{-1}\\ &\quad -0.37H_3(t)\left(1+\exp\{-(20y_{t-1}+48y_{t-4}+1.40)\}\right)^{-1}+u_t \qquad (2.26)\\ u_t &= \rho u_{t-1}+\varepsilon_t, \varepsilon_t \sim N(0,10^{-2}); \rho = 0, 0.1, 0.2 \qquad (2.27)\\ H_1(t) &= (1+\exp(-2(t-100.5)))-0.5\\ H_j(t) &= \left(1-\exp\left(-2(j-1)(t-100.5(j-1))^2\right)\right), j=2,3\end{aligned}$$

Setting $\rho \equiv 0$ generates a series with no autocorrelation, setting $H_j(t) \equiv 1$, $j = 1,2,3$ gives a process with constant parameters. The generated series are stationary since the roots of the characteristic equation of the linear part, $1-0.2z+0.1z^2 = 0$, lie outside the unit circle. Following Leisch, Trapletti, and Hornik (1999), Theorem 1, stationarity is guaranteed.

### 2.4.1 Specification and estimation

The aim is to simulate the model selection procedure, including both the variable selection in RTT and the subsequent choice of the number of hidden units. To this end, I set $\rho \equiv 0$ and $H_j(t) \equiv 1, j = 1, 2, 3$ in (2.27) and (2.26). The DGP is an ANN process with two hidden units, constant parameters, and normal and independent errors. The initial significance level of the hidden unit tests are 0.05, 0.10 and 0.20. This entails that if the initial significance level is 0.10, the hypothesis of 0 hidden units (linear model) against 1 or more is tested at the 0.10 level, 1 against 2 or more at 0.05, 2 against 3 or more at 0.025. The parameter estimation is carried out by the two-step procedure described in the paper. I also show the performances of the modelling procedure when the correct set of lags is not detected. I define underfitting when at least one correct lag is not detected, and overfitting when selecting all the relevant plus at least one redundant lag. I then rerun the whole experiment estimating the parameters with the frozen estimation method. This is done to find out whether or not such a simplified procedure can be applied in modelling real time series in cases where the full estimation method may be numerically difficult due to a large number of parameters.

Results are reported in table B.1. In part A of the table, I replicate 200 times a process with $T = 200$ observations. Let's consider the case where the initial significance level $\alpha_{IN}$ is equal to 0.05 first. Once linearity is rejected at the $\alpha_{IN}$ level, the variables are selected as in RTT, and an ANN model with one hidden unit is estimated by the full NLLS procedure. The hypothesis of one hidden unit is tested at the $\alpha_{IN}/2 = 0.025$ level, and a model with two hidden units is estimated if the null is rejected. The correct number of hidden units is detected in 182 cases over 197 of correct fitting. Overfitting never occurs, and in the 3 cases of underfitting the correct number of hidden units is found in only 50% of the cases. Increasing $\alpha_{IN}$ to 0.10 leads to a higher percentage of rejections when the null is true, $178 < 182$ correct detection; likewise, I get 173 correct cases when $\alpha_{IN} = 0.10$. When $T = 1000$ (part C and D), the correct set of variables is detected in the 100% of cases. For $\alpha_{IN} = 0.05, 0.10,$ 0.20, two hidden units are detected in 188, 193 and 188 cases respectively. A comparison between the two sample sizes tells us that when the ANN model is built on the correct set of variables, the LM testing procedure for the number of hidden units is reliable even in small samples. If the frozen estimation procedure is employed, for $T = 200$ (part B) I get 118, 142, and 128 cases and 160, 162, and 165 for $T = 1000$ (part D). This is due to the approximated estimation procedure. The residuals get more power when $T$ increases leading to overestimation. This feature suggests to set a limit to the maximum number of hidden units when dealing with actual data.

It is interesting to compare the estimates for $T = 200$, $T = 1000$, full and frozen estimation. I do that by saving the parameter estimates $\hat{\boldsymbol{\theta}}_i$ for each replication, $i = 1, 2, ..., 200$ and computing their median and median absolute deviation ($MAD$) as suggested in van Dijk (1999):

$$MAD\left(\hat{\boldsymbol{\theta}}\right) = \text{median}\left(\left|\hat{\boldsymbol{\theta}} - \text{median}\left(\hat{\boldsymbol{\theta}}\right)\right|\right)$$

High values for some of the estimates of gammas will often occur, leading to outliers. Utilizing the median can be seen as a robustification. The ratio between an estimate and its $MAD$ are taken as a measure of accuracy in table B.2. For $T = 200$, full estimation, the parameter estimates are close to their true values. Increasing the sample size to $T = 1000$ make the estimates much more accurate. For both sample sizes, the well known problem of hidden unit exchangeability occurs, which in our specific case means that $\hat{\beta}_2\psi(\hat{\gamma}_2, \mathbf{w}_t) = \hat{\beta}_2\left(1 - \psi(-\hat{\gamma}_2, \mathbf{w}_t)\right)$. The signs of the parameter estimates of the second hidden unit will be reverted, $\hat{\beta}_2 \approx -\beta_2$, $\hat{\gamma}_2 \approx -\gamma_2$ and the estimate of the intercept will be

approximately $\hat{\alpha}_0 \approx \alpha_0 - \hat{\beta}_2$. Indeed, the value of the $MAD$ is not always greater than that of the parameter estimates which have been correctly estimated, although it is computed on a distribution of the estimates which is far from being the correct one. I can therefore rely on the full NLLS estimation, being aware that it may need a final reparametrization. When frozen estimation is employed, the trade-off between such simplified estimation method and lower accuracy becomes apparent. In table B.1, part B ($T = 200$) and D ($T = 1000$) the number of detected hidden units is often overestimated; it is greater than 2 in 118, 142, 128 and 160, 162, 165 cases over 197 and 200 respectively. Furthermore, some of the estimates take different values for $T = 200$ and $T = 1000$, giving the idea that the estimation does not converge. Because the overall performances clearly worsen, I can draw the conclusion that frozen estimation should be utilized in actual data only when the full one becomes computationally unfeasible.

### 2.4.2 Evaluation

In order to assess the performances of the evaluation tests, I carry out diagnostic tests on the simulated series generated from the ANN model with 2 hidden units. When the correct lags and number of hidden units are detected, I apply an $LM$ test of serial independence and three $LM$ tests of parameter constancy. In power simulations for the serial correlation test, I assume that the errors follow an $AR(1)$ process $u_t = \rho u_{t-1} + \varepsilon_t$, $\varepsilon_t \sim N(0, 10^{-2})$, $\rho = 0.1$ and $\rho = 0.2$. In power simulations for the parameter constancy tests, I let the linear parameters change over time. I investigate power and size of the evaluation tools in the cases of time-varying parameters and constant ones ($H_j(t) \equiv 1$ for any $j$ and any $t$ in (2.26), autocorrelated and *n.i.d.* errors ($\rho = 0$ in (2.27)). Results for $T = 200$ observations and $N = 1000$ replications are reported in figures B.4-B.11 using size-power and size discrepacy graphs. The latter graphical representation has been devised in Davidson and MacKinnon (1998), and I briefly outline it in the following.

Size-power and size-discrepancy graphs are based on the empirical distribution function $\widehat{F}$ of the cumulative density function of the p-values $p$ of the test statistic $\tau$, $p_j = p(\tau_j)$. At any point $x_i$ of the $(0, 1)$ interval, it is defined by

$$\widehat{F}(x_i) = \frac{1}{N} \sum_{j=1}^{N} I(p_j \leq x_i)$$

Here, $I(p_j \leq x_i)$ is the indicator function, $N = 1000$ and $x_i = 0.002, 0.004, ..., 0.01, 0.02, ..., 0.99, 0.992, ..., 0.998$, therefore $i = 1, ...107$. The finer scaling near 0 and 1 ensures that I do not miss the behaviour of the test statistic $\tau$ in the tails. Plotting $\widehat{F}(x_i)$ against $x_i$ will give us the p-value plots, which should approach 1 under the alternative hypothesis when $x_i$ increases and be very close to a a 45° line under the null hypothesis, for any value of $x_i \in (0, 1)$. In the latter case, because the tests statistics behave approximately the way they should, it becomes very difficult to distinguish between $\widehat{F}(x_i)$ and the 45° line, therefore such graphical representation is not very useful. It is more convenient to plot the difference $\widehat{F}(x_i) - x_i$ against $x_i$, the size-discrepancy plots, which convey a lot more information when, as it is the case in this simulation study, the test statistics are well behaved.

When the DGP has constant parameters (figures B.4-B.6), the size distortion of the tests for no autocorrelation and parameter constancy is negligible, and its power good even in small samples. When autocorrelation is present, a certain amount of size distortion in the parameter constancy tests is expected. This is due to the general feature that parameter

constancy tests have some power towards the hypothesis of no autocorrelation. Introducing nonconstant parameters (figures B.9-B.11) biases the size of the tests for no autocorrelation when the null is true, while their power is still a positive function of $\rho$. As to parameter constancy tests, the power is good both for $\rho = 0$ and $\rho = 0.1, 0.2$.

## 2.5 Applications

### 2.5.1 Endogenous regressors case 1: the Canadian Lynx yearly series

#### Estimation and evaluation

The series I model is the 10-based logarithm of the yearly series of the number of lynx trapped in the region of the Mackenzie River, Canada, 1821-1934. The number of observations is $T = 114$. The series is nonlinear; for considerations and modelling efforts see, for example, Tong (1990) and Teräsvirta (1994). In an earlier survey, Lim (1987) performs an extensive comparison of previous studies.

I estimate the linear and neural network models on the period 1821-1890 (70 observations), leaving out the last $t_f = 44$ observations for evaluating forecasting performances. The sample is therefore split into two subsamples: the estimation set $y_1, ..., y_{T-t_f}$ and the set used for out-of-sample forecasting $y_{T-t_f+1}, .., y_T$. Because of the small number of observations, I utilize AIC as a selection criterion for the reasons explained in 3.3.2. Assuming a linear AR model, the model with the lowest AIC among all the possible linear ones generated by lags $1, 2, ..., 12$, is

$$y_t = \underset{(2.28)}{1.07} + \underset{(12.47)}{0.81}\, y_{t-1} - \underset{(-7.23)}{0.44}\, y_{t-4} + \underset{(3.55)}{0.25} y_{t-7} + \hat{\varepsilon}_t \tag{2.28}$$

| | | | |
|---|---|---|---|
| $R^2 = 0.87$ | $sk = 0.16$ | $AIC = -3.07$ | $LB(10) = 11.33(0.13)$ |
| $s = 0.21$ | $ek = 0.96$ | $SBIC = -2.98$ | $ARCH - LM(2) = 0.50(0.61)$ |
| | $LJB = 2.70(0.26)$ | | |

where the t-values of parameter estimates are in parentheses, $s$ is the standard deviation of the residuals, sk is skewness, ek is excess kurtosis, LJB is the Lomnicki-Jarque-Bera test for normality, ARCH-LM($i$) is the Engle's test for ARCH up to the lag $i$, and LB($j$) is the Ljung-Box test statistic up to lag $j$. For each test statistic, the corresponding p-value is written in brackets. The estimated linear model is weakly stationary since the roots of its lag polinomial all lie outside the unit circle. At the 5% significance level, the Lomnicki-Jarque-Bera test does not reject normality, but the Ljung-Box test statistic up to lag 7 does detect autocorrelation in the residuals.

In order to consider an ANN model, linearity is tested first. I denote the Lagrange multiplier test for zero hidden units against at least one, and one hidden unit against at least two by $\text{LM}_{lin}$ and $\text{LM}_{hu1}$ respectively. The starting significance level of the sequence of LM tests is 0.10. With a p-value of $\text{LM}_{lin}$ equal to $3.52 \times 10^{-3}$, the null hypothesis of zero hidden units is strongly rejected at the 10% level. Using the RTT procedure and the third-order Taylor expansion with the AIC criterion, lags 1 and 3 are selected. Because lags up to 12 are encompassed, the number of regressors by the 3rd order Taylor expansion is too large compared to $T$, and the RTT procedure needs to be approximated as explained in 3.3.2. The sequence of tests for the number of hidden units leads to a model with a single

hidden unit. As the test statistic of one unit against a greater number has p-value = 0.80, it is clearly rejected and the estimated model is:

$$\begin{aligned} y_t &= \underset{(-0.36)}{-0.89} + \underset{(0.46)}{0.27} y_{t-1} + \underset{(0.65)}{0.64} y_{t-3} \\ &\quad + \underset{(0.88)}{1.98} \cdot \left(1 + \exp\left(-\left(\underset{(0.74)}{2.00} y_{t-1} \underset{(-0.73)}{-3.61}\, y_{t-3} + \underset{(-0.67)}{5.03}\right)\right)\right)^{-1} + \hat{\varepsilon}_t \end{aligned} \tag{2.29}$$

$$\begin{array}{llll} R^2 = 0.89 & sk = -0.02 & AIC = -3.25 & LM_{lin} = 3.49(3.52 \times 10^{-3}) \\ s = 0.19 & ek = 0.41 & SBIC = -3.15 & LM_{hu1} = 0.54\,(0.80) \\ & LJB = 0.46(0.79) & \rho_{lin,hu_1} = -0.57 & ARCH - LM(2) = 0.20\,(0.66) \end{array}$$

where $\rho_{lin,hu_1}$ denotes the correlation between the linear unit and the hidden one. The t-values are in parenthesis below their corresponding parameter estimates. All of them are quite low; this is not surprising since the number of observations is only 70. On the other hand, the t-values of the parameter estimates of the hidden unit are of the same magnitude, therefore I do not drop any parameter. $R^2$ increases from 0.87 to 0.89. AIC and SBIC both decrease. The neural network model shows an improvement in fitting the data, but now I need to check its adequacy by the customized evaluation tests. The correlation coefficient $\rho_{lin,hu_1}$ takes value $-0.57$, significantly different from one in absolute value: this means that the hidden unit is not redundant. A graph of the hidden unit over time (figure B.13) shows that it is active over the whole estimation period. In figure B.14, the values of the logistic function ordered by ascending values of its exponent are plotted. Its shape is not very steep and the data points are distributed over all the interval where the logistic is greater than zero, therefore its estimation is supported by the whole sample. Weak stationarity of the model is assured since the roots of the lag polinomial of its linear part lie outside the unit circle. The Lomnicki-Jarque-Bera test does not reject normality in the residuals. The LM test for the hypothesis of serial independence presented in section 3.4.1 does not reject the null hypothesis at any lag, Engle 's $LM$ test for no ARCH does not detect any conditional heteroskedasticity. Parameter constancy is not rejected by the $LM_1$ test at the 10% level, but it is at the 5% by the $LM_2$ and $LM_3$. Results are reported in table B.3. Looking at the approximated correlation matrix of the estimated parameters (table B.4), I find that not only the nonlinear parameters are strongly correlated to each other, but the linear ones as well. This explains the results of the parameter constancy tests: the number of observations is not large enough to yield a stable estimation, and the whole estimated model is affected. For a comparison, see the sunspot example in section 5.2, where $T = 250$. The nonlinear estimated parameters have correlations close or even equal to one, but the linear ones are much less correlated. The uncertainty is mainly concentrated in the two nonlinear units. The larger number of data allows a better estimation, showing a typical property of ANN models: the nonlinear units provide a good fit to the series, but such flexibility is paid by a a more difficult estimation of the parameters. A graphical comparison between the observed data and the fitted ANN model is given in figure B.13. In the next paragraph, I shall compare the ANN, linear and STAR models by their forecasting performances.

### Forecasting one and four steps ahead

The final benchmark in model evaluation is forecasting. For such purpouse, I utilize $t_f = 44$ observations of the series. I want to compare performances of the ANN estimated model to

the estimated linear model and to the STAR model in Teräsvirta (1994) as specified below:

$$y_t = \underset{(49.76)}{1.15}\, y_{t-1} + (\underset{(-5.16)}{-1.35} y_{t-2} \underset{(3.47)}{+1.61}\, y_{t-3} \underset{(-2.91)}{-0.68}\, y_{t-4} \underset{(1.41)}{+0.12}\, y_{t-9} \underset{(-0.51)}{-0.06}\, y_{t-11}) \times \left(1 + \exp\left(\underset{(4.09)}{-1.87} \times 1.8 \left(y_{t-3} - \underset{(22.87)}{3.02}\right)\right)\right)^{-1} + \hat{\varepsilon}_t \tag{2.30}$$

The model has been re-estimated using the first 70 observations but without changing the specification. The t-values of parameter estimates are in parentheses. Following Swanson and White (1997), I base my comparison among the three models on a set of criteria for one-step ahead and four-step ahead out-of-sample forecasts. I also make a direct comparison between forecasts produced by two concurrent models by the Diebold-Mariano statistic (Diebold and Mariano (1995)) with the small-sample correction suggested in Harvey, Leybourne, and Newbold (1997). The MSE (forecast mean square error), MAD (mean absolute forecast error deviation) and MAPE (mean absolute percentage forecast error) are the most popular measures of forecasting accuracy. I compute these indexes for $t_f$ forecasts for the estimated models (not re-estimated after each period): $i = 1$ (linear model), $i = 2$ ( fully estimated ANN model) and $i = 3$ (STAR model); and horizon $h = 1, 4$:

$$MSE(h,i) = \frac{1}{t_f}\sum_{t=1}^{t_f}(\widehat{y}_{t,i}(h) - y_{t+h})^2 \tag{2.31}$$

$$MAD(h,i) = \frac{1}{t_f}\sum_{t=1}^{t_f}|\widehat{y}_{t,i}(h) - y_{t+h}| \tag{2.32}$$

$$MAPE(h,i) = \frac{1}{t_f}\sum_{t=1}^{t_f}\left|\frac{\widehat{y}_{t,i}(h) - y_{t+h}}{y_{t+h}}\right| \tag{2.33}$$

where $\widehat{y}_{t,i}(h)$ is the $h$-step ahead forecast for model $i$ made at time $t$. The forecast error for model $i$ , $h$ steps ahead, is defined as $\widehat{\varepsilon}_{t,i}(h) = \widehat{y}_{t,i}(h) - y_{t+h}$. To illustrate the performance of the models in forecasting the size and the direction of the change in the series, I consider the dispersion matrix $D_i(h)$, which is a $3 \times 3$ contingency table of the ordered actual observations and their forecasts:

| | | predicted data | | |
|---|---|---|---|---|
| | | -1 | 0 | 1 |
| The | -1 | $a_{11}$ | $a_{12}$ | $a_{13}$ |
| actual | 0 | $a_{21}$ | $a_{22}$ | $a_{23}$ |
| data | 1 | $a_{31}$ | $a_{32}$ | $a_{33}$ |

(2.34)

When the predicted or actual change in the series from $t$ to $t+1$ is less than 0.21 in absolute value, such change is classified as 0, when it is greater than 0.21, as 1 and when smaller, as $-1$. In our case, I choose the value 0.21 because such value is the standard deviation of the residuals of the linear model (2.28) and I denote $s = 0.21$ hereafter in this paragraph. For example, the pair $(0,0)$ (corresponding to $a_{22}$ in the table above) represents the cases where both the forecast and actual change are less than $s$ in absolute value. Likewise, $a_{11}$ is the number of cases where a predicted decrease greater than $s$ in absolute value corresponds to an actual decrease greater than $s$. The overall number of correct forecasts,

$a_{11} + a_{22} + a_{33} = trace\,(D_i\,(h))$ is of interest. Comparing it to the total number of forecasts, I get the *confusion rate* $CR_i\,(h)$, $0 \leq CR_i\,(h) \leq 1$:

$$CR_i\,(h) = 1 - \frac{trace\,(D_i\,(h))}{\sum_{l=1}^{3}\sum_{j=1}^{3} a_{lj}} \tag{2.35}$$

As to direct comparisons between the forecasts of two models, the Diebold-Mariano test works as follows. Consider two competing models $m_1$ and $m_2$ which have produced a vector of $h-$step ahead estimated forecast errors $\widehat{\varepsilon}_{m_i}(h) = (\widehat{\varepsilon}_{1,m_i}(h), ..., \widehat{\varepsilon}_{t_f,m_i}(h))'$. Consider the difference

$$d_t = g\left(\widehat{\varepsilon}_{t,m_1}(h)\right) - g\left(\widehat{\varepsilon}_{t,m_2}(h)\right), t = 1, ..., t_f$$

where $g(\cdot)$ is a generic loss function. Here I choose $g\left(\widehat{\varepsilon}_{t,m_i}(h)\right) = \widehat{\varepsilon}^2_{t,m_i}(h)$, so that the null hypothesis of equal performance can be defined as

$$E\left[\widehat{\varepsilon}^2_{t,m_1}(h) - \widehat{\varepsilon}^2_{t,m_2}(h)\right] = 0$$

for all $t$ and $h$. Defining $d_t = \widehat{\varepsilon}^2_{t,m_1}(h) - \widehat{\varepsilon}^2_{t,m_2}(h)$, I denote its mean and variance as follows

$$\overline{d} = \frac{1}{t_f}\sum_{t=1}^{t_f} d_t,\ \hat{V}\left(\overline{d}\right) = \frac{1}{t_f}\left(\hat{\gamma}_0 + 2\sum_{k=1}^{h-1}\hat{\gamma}_k\right)$$

where $\hat{\gamma}_k$ is the estimated $k-$th autocovariance and $\hat{\gamma}_0$ the estimate of the variance. The Diebold-Mariano statistic is

$$S = \left[\hat{V}\left(\overline{d}\right)\right]^{-1/2}\overline{d}$$

which is asymptotically distributed as a $N(0,1)$ under the null hypothesis. The test is oversized in small samples; Harvey, Leybourne, and Newbold (1997) suggested to use:

$$S^* = S \cdot \left(\frac{t_f + 1 - 2h + h(h-1)/t_f}{t_f}\right)^{1/2}$$

where the statistic $S^*$ is compared with the critical values of a student $t$ distribution with $n-1$ degrees of freedom since $\hat{V}\left(\overline{d}\right)$ has to be estimated.

As to one-step ahead (tables B.5-B.7), the ANN model is ranked second after the STAR model for the MSE, MAD, MAPE criteria (table B.5), but it outperforms the other two models as to the confusion rate CR. The modified Diebold-Mariano test cannot find a significant difference between the ANN and linear forecasts, but it clearly shows that the ANN model forecasts worse than the STAR model.

As to the four step-ahead case (tables B.8-B.10), the STAR model outperforms the linear and ANN models in MSE, MAD, MAPE. As to the confusion rate, the ANN performs worse than the other two, albeit only for one datum ($trace\,(D_{ANN}\,(4)) = 37$ against $trace\,(D_{linear}\,(4)) = trace\,(D_{STAR}\,(4)) = 38$). A smaller confusion rate than in the one-step ahead forecast, 0.16 against 0.28, should not mislead the reader; it is still greater than the linear and the STAR, both equal to 0.14. Indeed, MSE, MAD, MAPE are much greater than in the one-step ahead case. From table B.9, one can easily see that in all the dispersion matrices most of the cases are concentrated on the main diagonal. On the other hand, only a small number of them fits the case $(0,0)$, whereas most of them represent $(-1,-1)$ and $(1,1)$. This simply means that, when the horizon $h$ increases, it becomes more difficult to foresee anything but the sign of the variations. The modified Diebold-Mariano test leads to the same conclusions as in the one-step ahead case: ANN does not perform significantly worse than the linear model, but worse than the STAR.

### 2.5.2 Endogenous regressors case 2: the sunspot yearly series

#### Estimation and evaluation

This yearly series spans from 1700 to 1999. The data are kindly provided by the Sunspot Index Data Centre, Brussels, Belgium. I base the estimation of a linear and a neural network model on the period 1700-1949 (250 observations), leaving out the last $t_f = 50$ observations for forecasting evaluation. I tranform the data as in Tong (1990), where the author models $2 \cdot \left(\sqrt{1+S_t} - 1\right)$, where $S_t$ = sunspot number in year $(1699 + t)$. The lags encompassed in the modelling procedure span from 1 to 12, the selection criterion is SBIC since the sample size is not so small to prefer the AIC criterion. As to the linear model, the lags selected by the first order Taylor expansion as in RTT are 1, 2 and 9, and the model is:

$$y_t = \underset{(2.53)}{1.12} + \underset{(26.02)}{1.27}\, y_{t-1} - \underset{(-10.80)}{0.55}\, y_{t-2} + \underset{(6.04)}{0.17} y_{t-9} + \hat{\varepsilon}_t \tag{2.36}$$

| | | | |
|---|---|---|---|
| $R^2 = 0.86$ | $sk = 0.24$ | $AIC = 1.44$ | $LB(10) = 9.58(0.21)$ |
| $s = 2.02$ | $ek = 0.37$ | $SBIC = 1.50$ | $ARCH - LM(4) = 3.67(0.01)$ |
| | $LJB = 3.70(0.16)$ | | |

For the notation, I refer to the lynx series analyzed in the previous paragraph. The residual analysis can reject neither normality nor the hypothesis of no error autocorrelation. The test against ARCH indicates inadequacy of the linear model. Modelling an ANN implies linearity testing first: $LM_{lin} = 2.69\,(6.58 \times 10^{-4})$; therefore linearity is clearly rejected. Using the approximated RTT procedure with the third-order Taylor expansion (see section 3.3.2) up to lag 12, lags 1 and 2 are selected. It is also worth trying with a different order of the expansion, both up to the 4th and 5th, and with the AIC and SBIC selection criteria both. In all cases, lags 1 and 2 are still selected. Nevertheless, the ANN model based on lags 1 and 2 has estimation problems: the nonlinear optimization procedure fails to converge despite feeding it by a large set of starting values for the nonlinear parameters. I decide to include lag 9 as well since it is selected by the first-order expansion and its estimate has t-value 6.04 in the linear model.

The sequence of tests for selecting the number of hidden units rejects the hypothesis of one hidden unit against more than one, with a p-value of 0.015. Being the starting significance level of such sequence 0.10, $LM_{lin}$ (linearity test) is tested at the 10% level, $LM_{hu1}$ at the 5% level, $LM_{hu2}$ at the 2.5% level. A model with two hidden units is specified, but its full estimation turns out to be problematic. The nonlinear optimization procedure cannot find the global maximum; for any set of starting values, either it exceeds the maximum number of iterations or it returns an Hessian whose determinant is positive. As suggested in 3.2, in cases where the estimation of a model with more than one hidden unit turns out to be difficult, the frozen estimation procedure ought to give better results. In this particular case, this means conditioning the estimation of the linear part and the second hidden unit on the first one. The parameter estimates of the linear part and the second hidden unit, and the misspecification tests will be conditioned on the estimates of the first hidden unit. Despite such simplification, the nonlinear optimization procedure still has some numerical problems. Convergence is achieved by drawing a large set of starting values (90.000) for the nonlinear parameters of the two hidden units and by rescaling the exponents of the 2 squashing functions by dividing them by 100. In general, rescaling is necessary to make the gradient of the nonlinear and linear parameters of the same magnitude, but it is not possible

to set the scaling factor in advance, as it is often necessary to try with several of them. The aim is to obtain a balanced Hessian (the terms on the first diagonal should have the same magnitude) and it will be possible to invert it several times without losing accuracy. In this example, I tried by dividing by the standard deviation of the series first, then by 10 and 100 before yielding a balanced Hessian. Estimation troubles become apparent by analyzing the correlation matrix of the parameter estimates of the first hidden unit (table B.13) and of the linear part, and the second hidden unit conditioned on the first one (table B.6). The correlations among the estimates of the nonlinear parameters are all one in absolute value. This implies that their variance is very high, and it is not surprising to find out that their t-values are zero or very close to zero. I shall consider this issue later. To continue, the p-value of $LM_{hu2}$ (0.034) suggests to stop at the second hidden unit as the procedure based on the full estimation did. The estimated model is:

$$\begin{aligned} y_t = & \underset{(-1.73)}{-0.80} + \underset{(11.39)}{0.87}\, y_{t-1} - \underset{(-0.78)}{0.06}\, y_{t-2} + \underset{(5.39)}{0.15} y_{t-9} \\ & + \underset{(5.05)}{2.47} \cdot \left(1 + \exp\left(-\left(\underset{(0.00)}{47.72} y_{t-1} \underset{(-0.00)}{-75.25}\, y_{t-2} \underset{(0.00)}{+7.32}\, y_{t-9} \underset{(0.00)}{-2.58}\right)\right)\right)^{-1} \\ & + \underset{(3.73)}{1.81} \cdot \left(1 + \exp\left(-\left(\underset{(0.012)}{76.47} y_{t-1} \underset{(-0.012)}{-77.60}\, y_{t-2} \underset{(0.00)}{+0.53}\, y_{t-9} \underset{(0.00)}{-9.38}\right)\right)\right)^{-1} + \hat{\varepsilon}_t \end{aligned} \tag{2.37}$$

| | | |
|---|---|---|
| $R^2 = 0.87$ | $sk = 0.49$ | $ek = 0.58$ |
| $LJB = 1.29(1.56 \times 10^{-3})$ | $AIC = 1.27$ | $SBIC = 1.33$ |
| $s = 1.84$ | $ARCH - LM(4) = 1.37(0.24)$ | $\rho_{lin,hu_2} = 0.315$ |
| $LM_{lin} = 2.69\,(6.58 \times 10^{-4})$ | $LM_{hu1} = 2.05\,(0.011)$ | $LM_{hu2} = 1.79\,(0.034)$ |

where the t-values of the parameter estimates of the linear part and second hidden unit, the tests for serial correlation, ARCH and parameter constancy (tables B.11) are to be interpreted as conditioned to the first hidden unit estimates. The Lomnicki-Jarque-Bera test rejects normality (p-value $= 1.56 \times 10^{-3}$), but autocorrelation is not detected at any lag, parameter constancy is never rejected by any of the three $LM$ tests. At the 5% level, Engle's $ARCH - LM$ does not reject the null hypothesis of no ARCH/GARCH in the three tests up to lag 1, 2 and 4 respectively. As to the activity of the units in the model, $\rho_{lin,hu_2} = 0.315$ does detect correlation between the first hidden unit and the other units of the ANN model, but it is not so high such that any unit be redundant. A plot of the first hidden unit (figure B.17) demonstrates that it contributes to the data fitting over the whole sample, but it takes either value zero or one for most of the data (figure B.18, where, graphing its values ordered by ascending values of its exponent shows a transition function close to the step function).

The estimation of the second hidden unit conditioned to the first one turns out to be difficult as well. Again, the transition function (figure B.20) is quite close to the step function. Such steep slopes of the two squashing functions explain the estimation difficulties and the low t-values of the nonlinear parameter estimates. In both cases, there are only three data points in the interval where each logistic shifts from zero to one, as a result the variance of the nonlinear estimates is very high. On the other hand, I know that only the global maximum, corresponding to the lowest sum of squared residuals, provides the correct estimates. In such point of the parameter space the log-likelihood function may be quite flat compared to the points corresponding to local maxima, where the vector of local estimators may have lower variance. I stress the fact that the aim is to find the estimates of the global maximum of the

likelihood, being the reason of a possibly high variability an issue to investigate later. In the ANN fitting this series, such cause is to be found in the steep slopes of both the transition functions.

#### Forecasting one and four steps ahead

As to one-step ahead, the ANN model performs slightly better than the linear for the MSE, MAD, MAPE criteria (table B.14), but the confusion rates CR (table B.15) have both value 0.36. The modified Diebold-Mariano test (table B.16) cannot find a significant difference between the ANN and linear forecasts. In the four step-ahead case (tables B.17-B.19), the difference between the performances of the two models is even less pronounced.

It can be concluded that the sunspot series provides a good example of modelling nonlinearity by an ANN. The specification and estimation procedure this paper is about, has been utilized in all its aspects. It has been demonstrated that, when a model with more than one hidden unit is specified, estimation problems can be overcome by utilizing the frozen estimation procedure. Such method, despite being an approximation, estimates a model which satisfies all the misspecification tests.

The practical conclusion is that the ANN cannot forecast the sunspot series better than a linear model. It confirms the results of the lynx series, but cannot be considered disappointing. In fact, as can be read in the last essay of this thesis, where the forecasting performances of 30 time series are compared, ANN and linear models often perform similarly. The aim of the empirical part of this essay is to carefully apply the methodology developed in order to show that it can be an useful tool of analysis for nonlinear time series. As well as, generally speaking, a better fit is not a guaranteee for a better forecast, a careful application of such a tool cannot guarantee improvements in forecasting perfomances.

### 2.5.3 Exogenous regressors case: the S&P 500 excess return series

Several studies have found that stock returns can be predicted, and a significant portion of this can be done by macroeconomic and financial variables. The economic interpretation of such stylized facts is difficult, neverthless this holds true both across many stock markets and different periods of study. In Pesaran and Timmermann (1995), the authors investigate the predictability of U.S. stock returns by simulating the behaviour of an investor who decides, at each point in time, which economic factors are to be utilized for predicting one-period-ahead stock returns. Qi (1999) extends Pesaran and Timmermann (1995) by making the investor free to choose the functional form by which the selected economic factors predict stock-market returns. She applies a recursive modeling approach to the monthly S&P 500 index return, from January 1954 to December 1992, by using nine economic and financial variables to explain $\rho_t$, the excess return on the S&P 500 index at time $t$. Recursive estimation from 1959(12) to 1992(11) and in-sample forecasting of a linear and an ANN model is carried out, leading to the conclusion that not only the ANN model fits the data better in sample, but it improves out-of-sample forecasts. Racine (2001) could not replicate her published results, despite using the same software and data. Here, my purpose is to apply the ANN modelling procedure developed in this paper to the same data, and 1) check the adequacy of such procedure to the case of exogenous regressors, 2) evaluate my findings in term of accrued final wealth. My results are quite close to Racine's, but Qi's results are also reported.

### Data and methodology

The monthly series span from January 1954 to December 1992 ($T = 468$ observations) and are: $DY_t$ = dividend yield, $EP_t$ = earnings-price ratio, $(I1_t, I1_{t-1})$ = one-month Treasury-bill rate and its lag, $(I12_t, I12_{t-1})$ = 12-months Treasury-bond rate and its lag, $\pi_{t-1}$ = lag of the year-on-year rate of inflation, $\Delta IP_{t-1}$ = lag of the year-on-year rate of change in industrial output, and $\Delta M_{t-1}$ = year-on-year growth rate of narrow money stock. Following the standard practice in the stock-return literature, the excess returns $\rho_t$ is calculated by capital gain plus dividend yield minus the one-month Treasury-bill rate, $\rho_t = [(P_t - P_{t-1} + D_t)/P_{t-1}] - I1_{t-1}$, where $P_t$ is the stock price, and $D_t$ is the dividend. In the following, I group the 9 regressors into the vector

$$\mathbf{x}_t = (DY_t, EP_t, I1_t, I1_{t-1}, I12_t, I12_{t-1}, \pi_{t-1}, \Delta IP_{t-1}, \Delta M_{t-1}, 1)'$$

where the constant 1 has been added for notational convenience. The recursive forecast starts in 1960(1) and ends in 1992(12). Qi and Racine re-estimate each month a NN model with 8 hidden units and one hidden layer by Bayesian regularization, leading to a highly parametrized nonlinear model with 89 parameters. My ANN and linear models are re-specified and estimated each year (34 recursive estimations), as follows. At the beginning of each year, from 1959(12) to 1991(12), the regressors are selected by the third order Taylor expansion for the ANN model and the first order for the linear one; the estimation procedure is carried out and the estimated models are used to forecast step by step the excess return until the end of each year from 1960(12) to 1992(12). For each ANN model, the following results are collected: variables selected by the Taylor expansion procedure (figure B.21); number of hidden units; linear autocorrelation and parameter constancy test for the residuals (figures B.22-B.23); and standard deviation of the residuals (figure B.24). As to forecasting evaluation, MSE, MAD, MAPE and the Confusion Matrix are defined as in (2.31), (2.32), (2.33), (2.34) and (2.35), and are computed and compared to the corresponding indexes of the linear model based on the variables selected by the Taylor expansion of order one (table B.20-B.22).When linearity is not rejected (up to year 1963, the first three estimated models) the ANN models coincide with the linear. Finally, profitability of the linear and ANN models are compared to the results obtained by Racine using a portfolio based on stocks and on bonds only.

### In-sample evaluation

In Qi (1999), excess returns $\rho_t$ are explained by $\mathbf{x}_{t-1}$. Using the notation developed in this paper, the corresponding ANN model can be written:

$$\rho_t = \boldsymbol{\alpha}'\mathbf{x}_{t-1} + \sum_{j=1}^{Q} \beta_j \psi(\widetilde{\boldsymbol{\gamma}}_j'\mathbf{x}_{t-1}) + u_t, \; t = 1, ..., T, \; u_t \sim n.i.d.\left(0, \sigma_u^2\right)$$

where

$$\mathbf{x}_{t-1} = (DY_{t-1}, EP_{t-1}, I1_{t-1}, I1_{t-2}, I12_{t-1}, I12_{t-2}, \pi_{t-2}, \Delta IP_{t-2}, \Delta M_{t-2}, 1)'$$

and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_9, \alpha_0)'$ and $\widetilde{\gamma}_j = \left(\widetilde{\gamma}_{j1}, \widetilde{\gamma}_{j2}, ..., \widetilde{\gamma}_{j,10}\right)'$. I apply the modelling procedure recursively each year, from model 1 (estimation period 1954(1)-1959(12)) to 34 (estimation period 1954(1)-1992(12)). In figure B.21, the selected variables are presented. In the period 1954(1) to 1962(12), corresponding to the first three estimated models, the models are linear

and are based on a single explanatory variable, $I1_{t-2}$, the one-month Treasury-bill rate lagged two months back. In the following years up to 1992, the model is nonlinear but it has a single hidden unit until the end of the sample, $Q = 1$. A first view tells that the regressors which are most frequently included are $EP_{t-1}$ and $I1_{t-1}$, although $I12_{t-2}$ and $\Delta M_{t-2}$ are often included. In general, the models are quite parsimonious; only in one case (model n. 15, corresponding to year 1973) are 5 regressors included, and 4 in another case (model n. 18, year 1976). During the first half of the sample (1959-1973), 2 regressors are included on the average, 3 in the second half (1974-1992). Comparing the 2 half-samples by figure B.24 (standard deviation of the residuals), it is apparent that, while the linear and ANN model perform similarly in the first half except between observation 108 and 192, corresponding to the period 1968-1975, the ANN models have a constantly higher standard deviation in the second one. In figure B.25-B.29 the data, the ANN model and the linear fit are presented in a more detailed manner. Despite such lack of fit, I can affirm that the ANN models are well specified. In figure B.22, the cumulative number of rejections of the null hypothesis of no autocorrelation up to lag 1, 2, 4 and 10 gives an idea of the evolution of such test along time. From model 1 (estimation up to 1959(12)) to model 34 (up to 1992(12)), autocorrelation up to lag 1 and 2 is detected only once, up to lag 4 only twice. Likewise, looking at the 3 LM tests of parameter constancy (figure B.23), the null hypothesis is rejected 11 times ($LM_1$), and 9 times ($LM_2$ and $LM_3$) over 34 estimated models. Such result indicates some uncertainty in the parameter estimates, a problem that I also encountered in the sunspot example in paragraph 5.2.1.

### Forecasting one-step ahead

As to forecasting performances, one-step ahead forecasts are evaluated in tables B.20-B.22. A comparison with the linear model is carried out in all the tables, and results agree with the extensive comparison as in essay III of this thesis. The linear and ANN model perform likewise. As can be seen in table B.20, MSE is equal to 0.0019 for the linear and 0.0022 for the ANN; MAD = 0.0343 and 0.0352, MAPE = 2.577 and 2.259 respectively. The dispersion matrix is slightly better for the linear model, but the confusion rates of the 2 models are the same. The Diebold-Mariano test rejects the hypothesis of the equality of the performances of the one step forecasts at the 5% significance level, but not at the 1% (p-value = 0.0139).

### Profitability

Finally, I compare profitability of the ANN modelling procedure to the linear one as Qi and Racine did, considering three different strategies: 1) the investor always holds equities and reinvests dividends, 2) he always holds bonds reinvesting the accrued interests and 3) the investor follows a switching strategy that has been extensively adopted in the finance literature, holding equities in periods when the forecast returns are greater than the bonds yield and otherwise holding bonds. I start with an investor holding \$100 at the end of 1959(12), considering zero transaction costs for equities and bonds, low transaction costs (0.5% for trading in stocks and 0.1% in trading in bonds), high transaction costs (1% on equities and 0.1% on bonds). Results are reported in table B.23. If I look at the column "Final wealth", I can notice that the bond (panel A) and market portfolios (panel B) are clearly outperformed by the switching portfolios based on the linear model (panel C) and ANN model (panel D). I therefore concentrate on the comparison between the last two, and the wealth produced by Racine's models. I report the results he obtained by using a switching portfolio based on a NN estimated by Bayesian regularization with 100 epochs

and 10 multistarts. At zero transaction costs, my linear model produces a final wealth of \$6780 against \$6132 for the ANN one. The set of ANN models estimated by the procedure developed in this paper produce a final wealth which is very similar results to Racine's: \$6132 against \$6265 for Racine. This occurs despite he restimated the NN model at each period of time (every month), while I do it only every 12 periods (every year). Qi obtained a much higher value, \$13820. At low and high transaction costs, results are quite close too, being again both very far from Qi's. Therefore the final wealth produced by the NN methodology developed in this paper cannot be considered, in my opinion, disappointing. As to risk analysis, represented by the standard deviation of returns, using the linear model turns out to be the safest strategy, while the more risky is the ANN, and the market portfolio approach stands in-between.

## 2.6 Conclusions

In this paper, I discuss a specification technique based on a statistical tools which can be applied to a wide range of nonlinear processes. If linearity is not rejected, a single hidden layer ANN model is fitted to the data and evaluated in a simple but statistically consistent way. The simulation results demonstate the performances of the procedure when the true DGP is an ANN model. It can be applied to any financial and macroeconomic stationary series which satisfies the moment conditions. Both in the two cases with endogenous regressors and in the one with exogenous regressors, out-of sample performances turn out to be acceptable although a subset of the regressors is utilized. This leads to the conclusion that the well-known problem of overfitting, which often arises in NN models estimated by traditional black-box methods, can be overcome by using the methodology developed in this paper.

# Bibliography


AKAIKE, H. (1974): "A New Look at the Statistical Model Identification," *IEEE Transactions of Automatic Control*, AC-19, 716–23.

ANDERS, U., AND O. KORN (1999): "Model Selection in Neural Networks," *Neural Networks*, 12, 309–323.

BOX, G. E. P., AND G. M. JENKINS (1970): *Time Series Analysis, Forecasting and Control.* Holden Day, San Francisco.

CYBENKO, G. (1989): "Approximations by Superimpositions of a Sigmoidal Function," *Mathematics of Control, Signals, and Systems*, 9, 2–54.

DAVIDSON, R., AND J. G. MACKINNON (1993): *Estimation and Inference in Econometrics.* Oxford University Press, Oxford.

——— (1998): "Graphical Methods for Investigating the Size and Power of Tests Statistics," *The Manchester School*, 66, 1–26.

DIEBOLD, F. X., AND R. S. MARIANO (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263.

EITRHEIM, Ø., AND T. TERÄSVIRTA (1996): "Testing the Adequacy of Smooth Transition Autoregressive Models," *Journal of Econometrics*, 74, 59–75.

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

GODFREY, L. (1988): *Misspecification Tests in Econometrics.* University Press, Cambridge.

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

HARVEY, A. C. (1990): *The Econometric Analysis of Time Series.* Philip Allan.

HARVEY, D., S. LEYBOURNE, AND P. NEWBOLD (1997): "Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting*, 13, 281–291.

HORNIK, K., M. STINCHCOMBE, AND H. WHITE (1989): "Multi-Layer Feedforward Networks are Universal Approximators," *Neural Networks*, 2, 359–66.

HWANG, J. T. G., AND A. A. DING (1997): "Prediction Intervals for Artificial Neural Networks," *Journal of the American Statistical Association*, 92, 748–757.

LEISCH, F., A. TRAPLETTI, AND K. HORNIK (1999): "Stationarity and Stability of Autoregressive Neural Network Processes," *Advances in Neural Information Processing Systems*, 11.

LIM, K. S. (1987): "A Comparative Study of Various Univariate Time Series Models for Canadian Lynx Data," *Journal of Time Series Analysis*, 8, 161–176.

LIN, C.-F. J., AND T. TERÄSVIRTA (1994): "Testing the Constancy of Regression Parameters Against Continuous Structural Change," *Journal of Econometrics*, 62, 211–228.

PESARAN, M. H., AND A. TIMMERMANN (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 & Economic Statistics*, 17, 419–429.

RACINE, J. (2001): "On the Nonlinear Predictability of Stock Returns Using Financial and Economic Variables," *Journal of Business & Economic Statistics*, 19, 380–382.

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

REED, R. (1993): "Pruning Algorithms - a Survey," *IEEE Transactions on Neural Networks*, 4, 740–747.

REFENES, A.-P. N., AND A. D. ZAPRANIS (1999): "Neural Model Identification, Variable Selection and Model Adequacy," *Journal of Forecasting*, 18, 299–332.

RISSANEN, J. (1978): "Modeling by Shortest Data Description," *Automatica*, 14, 465–471.

SCHWARZ, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–64.

STINCHCOMBE, M. B. (1995): "Precision and Approximate Flatness in Artificial Neural Networks," *Neural Computation*, 7, 1021–1039.

STONE, M. (1977): "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion Cross Validation," *Journal of the Royal Statistic Society*, B36, 44–47.

SWANSON, N. R., AND H. WHITE (1997): "Forecasting Economic Time Series Using Flexible versus Fixed Specification and Linear versus Nonlinear Econometric Models," *International Journal of Forecasting*, 13, 439–461.

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

TERÄSVIRTA, T. (1998): "Modelling Economic Relationships with Smooth Transition Regressions," in *Handbook of Applied Economic Statistics*, ed. by A. Ullah, and D. Giles, pp. 507–552. Dekker.

TERÄSVIRTA, T., C.-F. LIN, AND C. W. J. GRANGER (1993): "Power of the Neural Network Linearity Test," *Journal of Time Series Analysis*, 14, 209–220.

TERÄSVIRTA, T., AND C.-F. J. LIN (1993): "Determining the Number of Hidden Units in a Single Hidden-Layer Neural Network Model," Discussion paper, Bank of Norway, Research Department.

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

TSAY, R. S. (1989): "Testing and Modeling Threshold Autoregressive Processes," *Journal of the American Statistical Association*, 84, 231–240.

——— (1998): "Testing and Modeling Multivariate Threshold Models," *Journal of the American Statistical Association*, 93, 1188–1202.

VAN DIJK, D. (1999): "Smooth Transition Models: Extensions and Outlier Robust Inference," Ph.D. thesis, Tinbergen Institute, Rotterdam, The Netherlands, www.few.eur.nl/few/people/djvandijk/thesis/thesis.htm.

WHITE, H. (1989a): "An Additional Hidden Unit Test for Neglected Nonlinearity in Multilayer Feedforward Networks," *Proceedings of the international joint conference on neural networks, Washington, DC*, pp. 451–455.

——— (1989b): "Parametric Statistical Estimation with Artificial Neural Networks: A Condensed Discussion," *In Cherkassky, V., Friedman, J.H., Welchsler, H., eds, From Statistics to Neural Networks, NATO ASI series: Series F, Computer and System Sciencies, Vol 136.*

——— (1989c): "Some Asymptotic Results for Learning in Single Hidden-Layer Feedforward Network Models," *Journal of the American Statistical Association*, 84, 1003–1013.

# Appendix A

# Gradient of the ANN model with q hidden units

I need to compute the analytical first derivatives of model (2.8) in order to carry out the tests for the evaluation of the model and to increase the accuracy of the nonlinear optimization procedure. I rewrite model (2.8):

$$y_t = \boldsymbol{\alpha}' \mathbf{w}_t + \sum_{j=1}^{q} \beta_j \psi(\boldsymbol{\gamma}_j, \mathbf{w}_t) + u_t, t = 1, ..., T, u_t \sim n.i.d.(0, \sigma_u^2) \qquad (A.1)$$

where $\mathbf{w}_t = (w_{t1}, ..., w_{tk}, 1)' = (y_{t-1}, y_{t-2}, ..., y_{t-k}, 1)'$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_k, \alpha_0)'$, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_q)'$, $\boldsymbol{\gamma}_j = \left(\gamma_{j1}, \gamma_{j2}, ..., \gamma_{jk}, c_j^*\right)'$, $j = 1, ..., q$.

The first derivatives of $(A.1)$ with respect to the vector of parameters are:

$$\frac{\partial y_t}{\partial \boldsymbol{\alpha}'} = \mathbf{w}_t; \frac{\partial y_t}{\partial \beta_j} = \psi(\boldsymbol{\gamma}_j, \mathbf{w}_t) \ ;$$

As to the first derivatives with respect to each $\boldsymbol{\gamma}_j$, rewriting (2.6):

$$\psi(\boldsymbol{\gamma}_j, \mathbf{w}_t) = \left(1 + c_j^* \exp\left\{-\left(\gamma_{j1} w_{t1} + \gamma_{j2} w_{t2} + ... + \gamma_{jk} w_{tk}\right)\right\}\right)^{-1}$$

$$\begin{aligned} \frac{\partial \psi(\boldsymbol{\gamma}_j, \mathbf{w}_t)}{\partial \boldsymbol{\gamma}_{ji}} &= \left(\psi(\boldsymbol{\gamma}_j, \mathbf{w}_t)\right)^{-2} c_j^* \exp\left\{-\left(\gamma_{j1} w_{t1} + \gamma_{j2} w_{t2} + ... + \gamma_{jk} w_{tk}\right)\right\} w_{ti} \\ i &= 1, ..., k \end{aligned}$$

$$\frac{\partial \psi(\boldsymbol{\gamma}_j, \mathbf{w}_t)}{\partial c_j^*} = -\left(\psi(\boldsymbol{\gamma}_j, \mathbf{w}_t)\right)^{-2} \exp\left\{-\left(\gamma_{j1} w_{t1} + \gamma_{j2} w_{t2} + ... + \gamma_{jk} w_{tk}\right)\right\}$$

# Appendix B

# Tables and figures

| A: Full estimation, sample size T = 200 | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Correct fitting | | | | | Overfitting | Underfitting | | | All | | | | |
| $\alpha_{IN}$ | 0 | 1 | 2 | >2 | Tot | Tot | 2 | >2 | Tot | 0 | 1 | 2 | >2 | Tot |
| 0.05 | 0 | 2 | 182 | 13 | 197 | 0 | 1 | 2 | 3 | 0 | 2 | 183 | 15 | 200 |
| 0.10 | 1 | 2 | 178 | 17 | 197 | 0 | 1 | 2 | 3 | 0 | 2 | 179 | 19 | 200 |
| 0.20 | 0 | 2 | 173 | 22 | 197 | 0 | 1 | 2 | 3 | 0 | 2 | 174 | 24 | 200 |
| B: Frozen estimation, sample size T = 200 | | | | | | | | | | | | | | |
| | Correct fitting | | | | | Overfitting | | Underfitting | | | All | | | |
| $\alpha_{IN}$ | 0 | 1 | 2 | >2 | Tot | 0 | Tot | 2 | >2 | Tot | 0 | 1 | 2 | >2 | Tot |
| 0.05 | 0 | 1 | 78 | 118 | 197 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 79 | 119 | 200 |
| 0.10 | 0 | 0 | 51 | 142 | 193 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 52 | 142 | 194 |
| 0.20 | 0 | 1 | 68 | 128 | 197 | 1 | 1 | 0 | 2 | 2 | 1 | 1 | 68 | 130 | 200 |
| C: Full estimation, sample size T = 1000 | | | | | | | | | | | | | | |
| | Correct fitting | | | | | Overfitting | Underfitting | | | All | | | | |
| $\alpha_{IN}$ | 0 | 1 | 2 | >2 | Tot | Tot | Tot | | | 0 | 1 | 2 | >2 | Tot |
| 0.05 | 0 | 0 | 188 | 12 | 200 | 0 | 3 | | | 0 | 0 | 188 | 12 | 200 |
| 0.10 | 0 | 0 | 193 | 7 | 200 | 0 | 3 | | | 0 | 0 | 193 | 7 | 200 |
| 0.20 | 0 | 0 | 188 | 12 | 200 | 0 | 3 | | | 0 | 0 | 188 | 12 | 200 |
| D: Frozen estimation, sample size T = 1000 | | | | | | | | | | | | | | |
| | Correct fitting | | | | | Overfitting | Underfitting | | | All | | | | |
| $\alpha_{IN}$ | 0 | 1 | 2 | >2 | Tot | Tot | Tot | | | 0 | 1 | 2 | >2 | Tot |
| 0.05 | 0 | 0 | 40 | 160 | 200 | 0 | 0 | | | 0 | 0 | 40 | 160 | 200 |
| 0.10 | 0 | 0 | 38 | 162 | 200 | 0 | 0 | | | 0 | 0 | 38 | 162 | 200 |
| 0.20 | 0 | 0 | 35 | 165 | 200 | 0 | 0 | | | 0 | 0 | 35 | 165 | 200 |

Table B.1: The number of hidden units detected by the sequence of LM tests, conditional on the detection of the correct lags 1, 2, and 4 among the first 6, on overfitting (all correct lags plus redundant ones are detected), and on underfitting (not all correct lag(s) are detected). The initial significance levels of the tests for the number of hidden units are 0.05, 0.10, 0.20. 200 replications of the series generated by process with constant parameters and no autocorrelation in the residuals.

| | | $T = 200$ | | | | $T = 1000$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | true | full estim. | | frozen estim. | | full estim. | | frozen estim. | |
| param. | value | median | MAD | median | MAD | median | MAD | median | MAD |
| $\beta_0$ | 0.07 | −0.30 | 0.019 | −1.88 | 1.58 | −0.30 | 0.00 | −2.13 | 1.38 |
| $\alpha_1$ | 0.20 | 0.19 | 0.034 | −0.58 | 0.78 | 0.20 | 0.01 | −0.74 | 0.71 |
| $\alpha_2$ | −0.10 | −0.10 | 0.047 | 1.38 | 1.54 | −0.10 | 0.02 | 1.53 | 1.11 |
| $\beta_1$ | 0.53 | 0.54 | 0.035 | 2.95 | 2.50 | 0.53 | 0.01 | 3.59 | 2.58 |
| $\gamma_{11}$ | 17 | 16.69 | 2.64 | 1.22 | 0.67 | 17.00 | 1.22 | 1.25 | 0.52 |
| $\gamma_{12}$ | −30.8 | −30.24 | 4.68 | −2.40 | 1.57 | −30.66 | 1.99 | −0.58 | 0.40 |
| $c_1$ | 1.19 | 1.17 | −0.24 | 0.31 | 0.26 | 1.19 | −0.10 | 0.23 | −0.17 |
| $\beta_2$ | −0.37 | 0.37 | 0.02 | 0.27 | 0.14 | 0.37 | 0.06 | 0.15 | 0.03 |
| $\gamma_{21}$ | 20 | −20.67 | 7.40 | 0.10 | 15.14 | −20.07 | 2.65 | 1.49 | 5.55 |
| $\gamma_{22}$ | 48 | −50.57 | 16.18 | −0.21 | 3.79 | −48.40 | 5.74 | −54.36 | 54.64 |
| $c_2$ | −1.4 | −1.31 | 0.57 | −0.07 | −0.93 | −1.38 | 0.24 | 0.32 | 1.08 |

Table B.2: Median and MAD of the NLLS estimates of the parameters of the previous table, where the initial significance level is 0.10

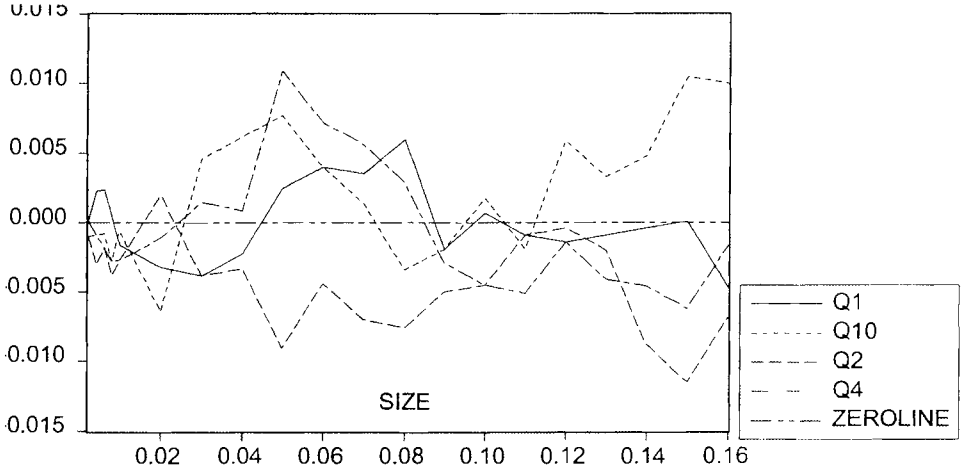


Figure B.1: size discrepancy plot of the test of no error autocorrelation up to lags 1 (Q1), 2 (Q2), 4 (Q4) and 10 (Q10) at the sample size $T = 200$, for 1000 replications of the series generated by process 2.26 with $\rho = 0$ and constant parameters:

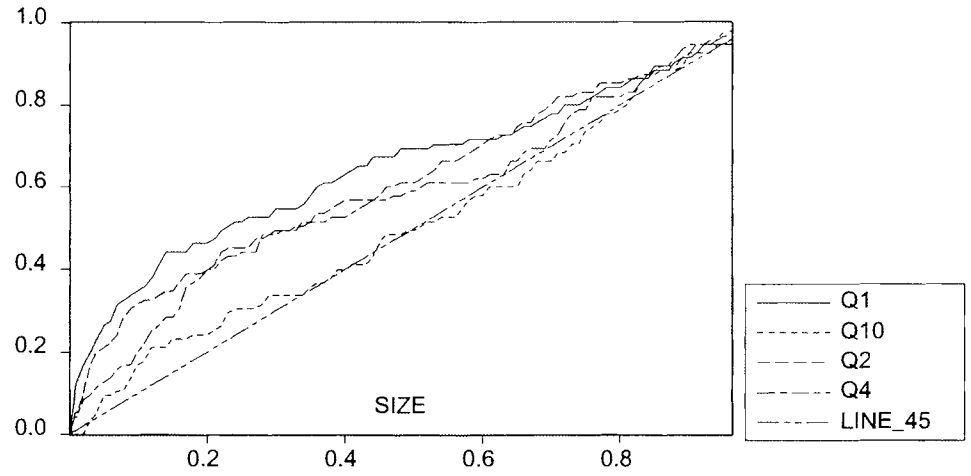


Figure B.2: size-power plot of the test of no error autocorrelation up to lags 1 (Q1), 2 (Q2), 4 (Q4) and 10 (Q10) at the sample size $T = 200$, for 1000 replications of the series generated by process 2.26 with $\rho = 0.1$ and constant parameters.

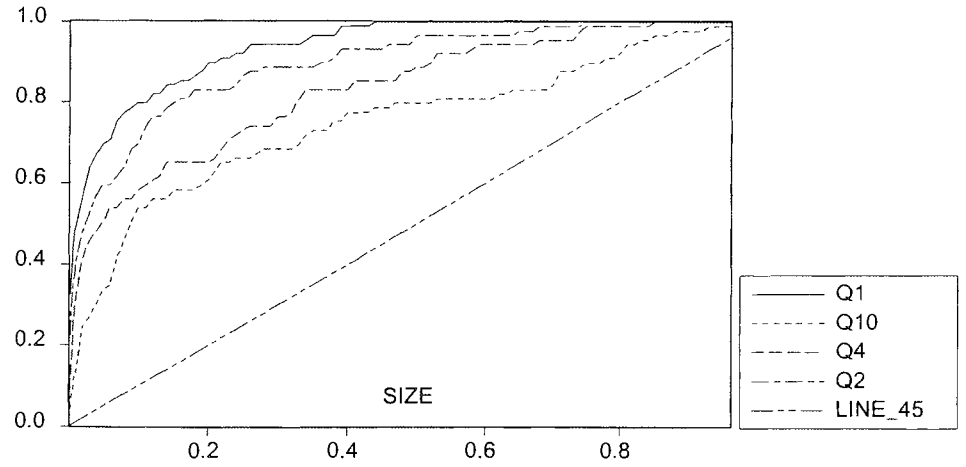


Figure B.3: size-power plot of the test of no error autocorrelation up to lags 1 (Q1), 2 (Q2), 4 (Q4) and 10 (Q10) at the sample size $T = 200$, for 1000 replications of the series generated by process 2.26 with $\rho = 0.2$ and constant parameters.

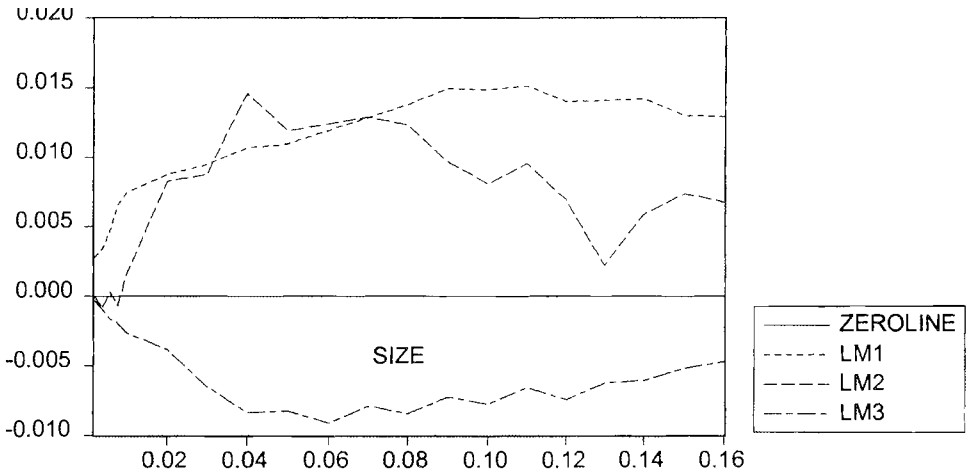


Figure B.4: p-value discrepancy plot of the tests of parameter constancy $LM_1$, $LM_2$, and $LM_3$ at the sample size $T = 200$, for 1000 replications of the series generated by process 2.26 with $\rho = 0$ and constant parameters.

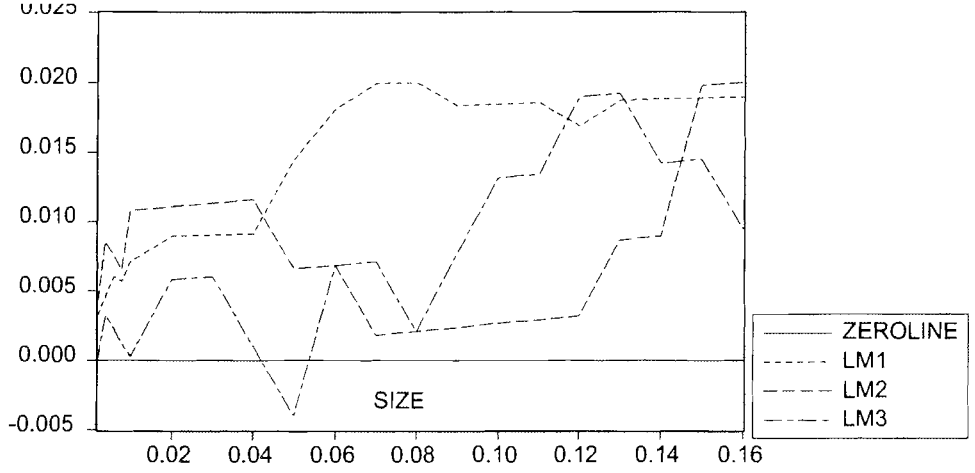


Figure B.5: p-value discrepancy plot of the tests of parameter constancy $LM_1$, $LM_2$, and $LM_3$ at the sample size $T = 200$, for 1000 replications of the series generated by process 2.26 with $\rho = 0.1$ and constant parameters.

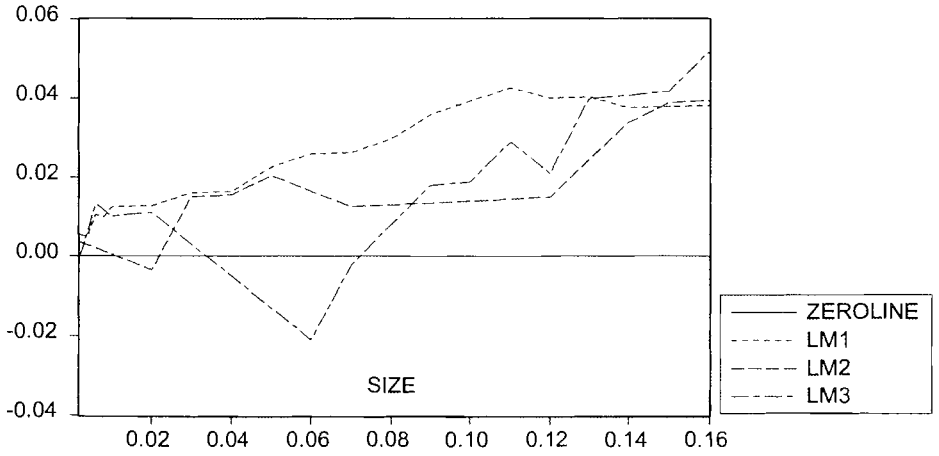


Figure B.6: p-value discrepancy plot of the tests of parameter constancy $LM_1$, $LM_2$, and $LM_3$ at the sample size $T = 200$, for 1000 replications of the series generated by process 2.26 with $\rho = 0.2$ and constant parameters.

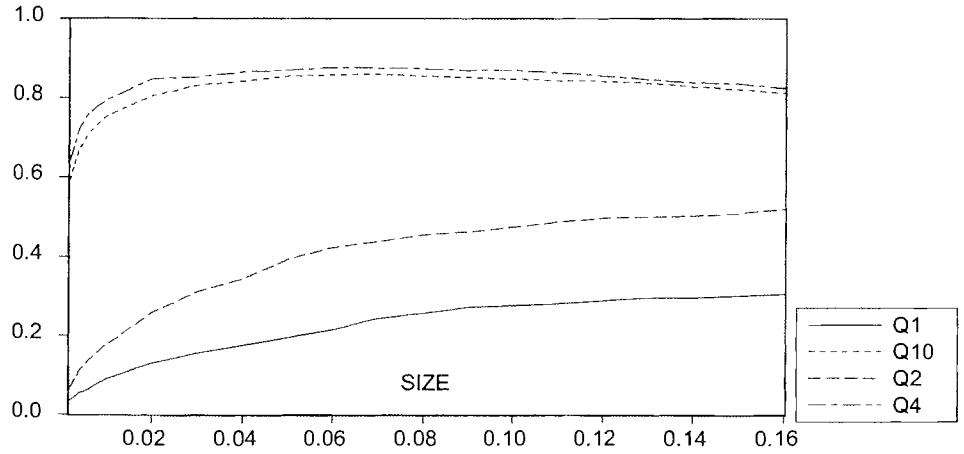


Figure B.7: size discrepancy plots of the test of no error autocorrelation up to lags 1 (Q1), 2 (Q2), 4 (Q4) and 10 (Q10) at the sample size $T = 200$, for 1000 replications of the series generated by process 2.26 with $\rho = 0$ and time-varying parameters.

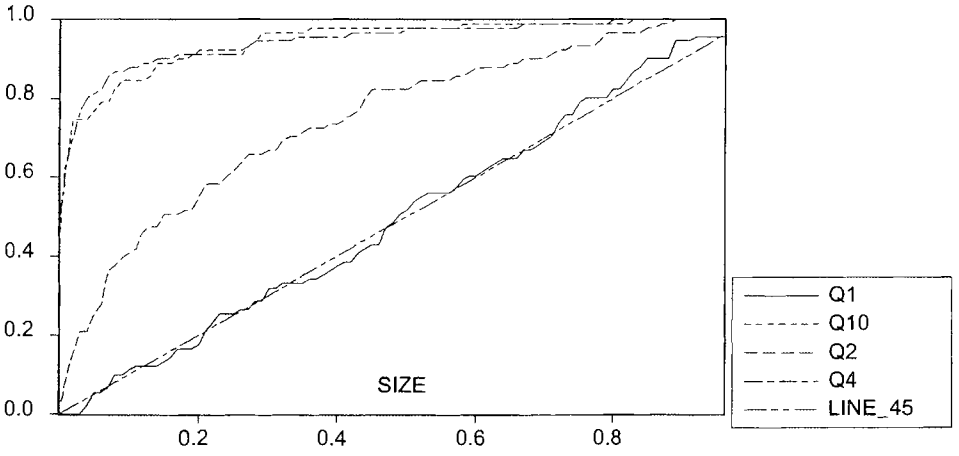


Figure B.8: size-power curves of the test of no error autocorrelation up to lags 1 (Q1), 2 (Q2), 4 (Q4) and 10 (Q10) at the sample size $T = 200$, for 1000 replications of the series generated by process 2.26 with $\rho = 0.1$ and time-varying parameters

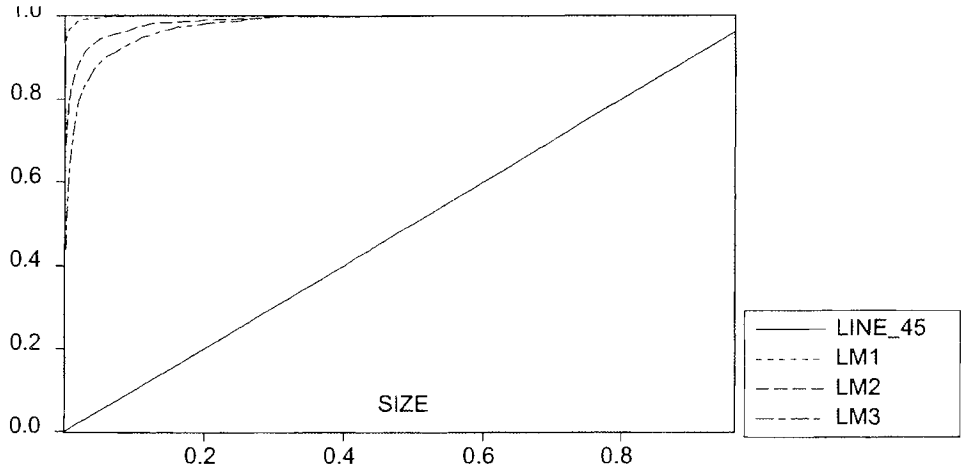


Figure B.9: size-power curves of the three tests of parameter constancy $LM_1, LM_2, LM_3$ against smooth structural change at the sample size $T = 200$, for 1000 replications of the series generated by process 2.26 with $\rho = 0$ and time-varying parameters.

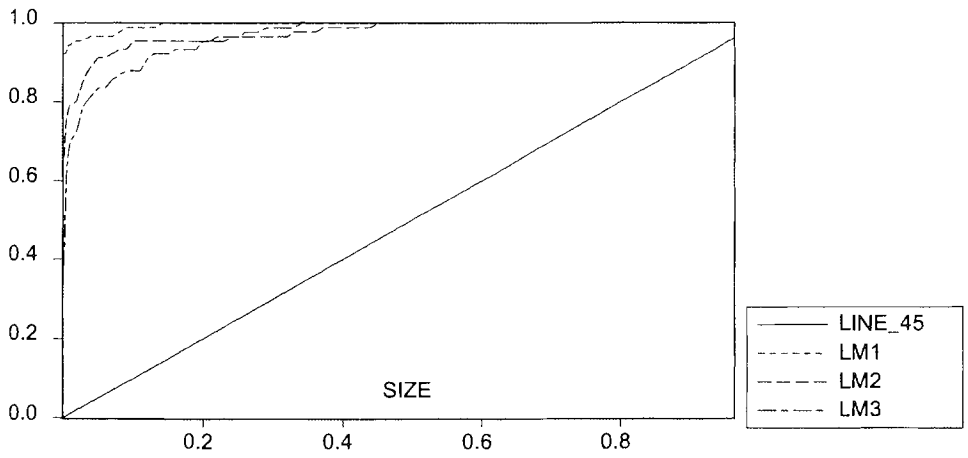


Figure B.10: size-power curves of the three tests of parameter constancy $LM_1, LM_2, LM_3$ against smooth structural change at the sample size $T = 200$, for 1000 replications of the series generated by process 2.26 with $\rho = 0.1$ and time-varying parameters.

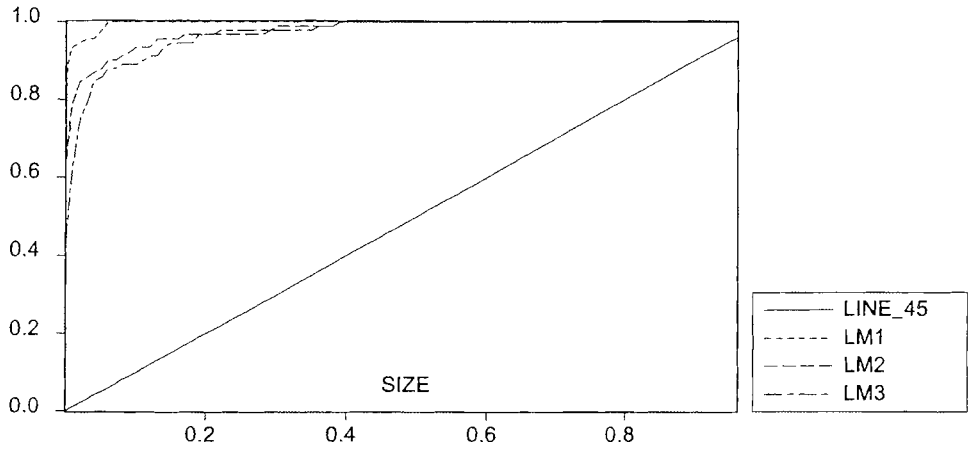


Figure B.11: size-power curves of the three tests of parameter constancy $LM_1, LM_2, LM_3$ against smooth structural change at the sample size $T = 200$, for 1000 replications of the series generated by process 2.26 with $\rho = 0.2$ and time-varying parameters.

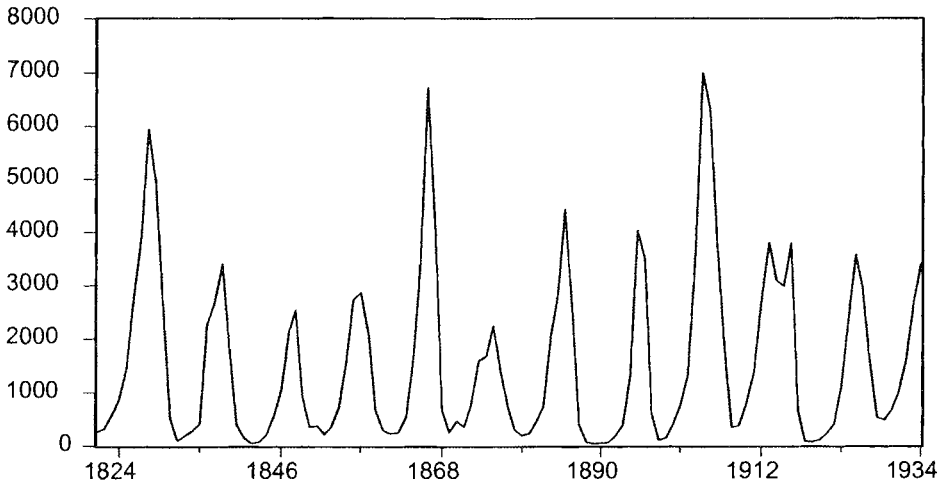


Figure B.12: The annual number of Canadian lynx trappings, 1821-1934. Observations preceding 1891 are used for estimation and the remaining ones for forecasting evaluation.

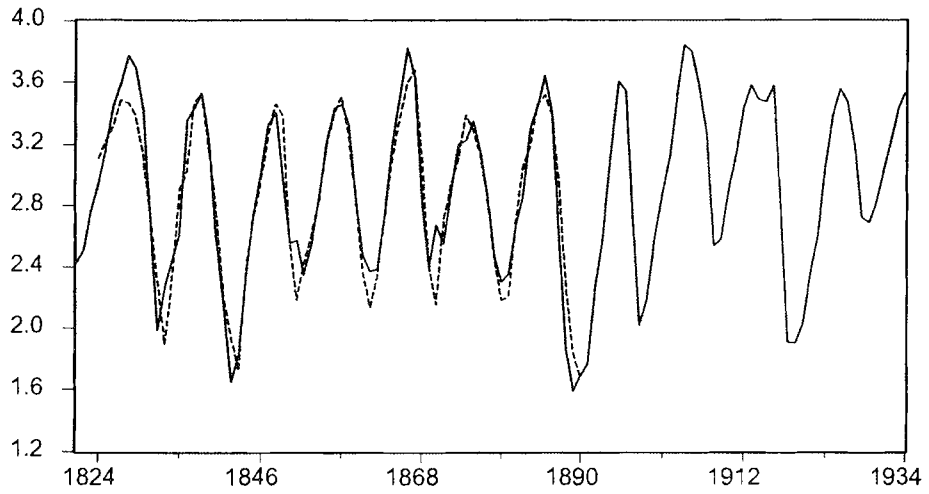


Figure B.13: Solid line: logarithms of the annual number of Canadian lynx trappings, 1821-1934. Observations preceding 1891 are used for estimation and the remaining ones for forecasting evaluation. Dotted line: neural network model fit along the estimation period, 1821-1890

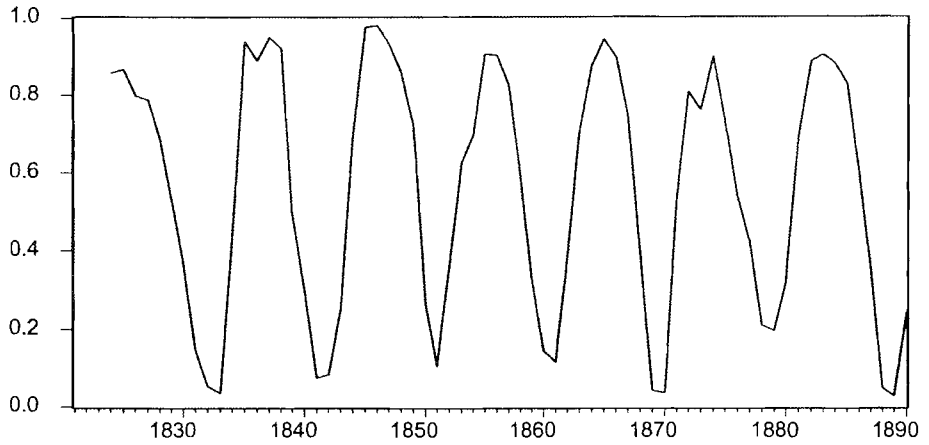


Figure B.14: Values of the first hidden unit for the conditional mean of the model estimated during period 1821-1890 for the logarithm of the Canadian lynx trappings.

| | r=1 | r=2 | r=4 | r=10 |
|---|---|---|---|---|
| serial independence | 0.48 (0.49) | 1.42 (0.25) | 1.40 (0.25) | 0.98 (0.99) |
| no ARCH | 0.20 (0.66) | 0.76 (0.47) | 0.46 (0.77) | 0.22 (0.99) |
| | $LM_1$ | $LM_2$ | $LM_3$ | |
| parameter constancy | 1.10 (0.37) | 1.82 (0.09) | 2.02 (0.04) | |

Table B.3: Canadian lynx series: values of LM tests for serial independence against AR(r) or MA(r) errors, Engle's LM tests for ARCH(r), and parameter constancy tests LM1, LM2, LM3. Corresponding p-values are reported in parenthesis.

| | $\hat{\beta}_0$ | $\hat{\alpha}_1$ | $\hat{\alpha}_2$ | $\hat{\beta}_1$ | $\hat{\gamma}_{11}$ | $\hat{\gamma}_{12}$ | $\hat{c}_1$ |
|---|---|---|---|---|---|---|---|
| $\hat{\beta}_0$ | | 0.923 | -0.989 | -0.982 | 0.964 | -0.952 | -0.870 |
| $\hat{\alpha}_1$ | | | -0.969 | -0.974 | 0.858 | -0.943 | -0.957 |
| $\hat{\alpha}_2$ | | | | 0.995 | -0.937 | 0.961 | 0.914 |
| $\hat{\beta}_1$ | | | | | -0.947 | 0.980 | 0.941 |
| $\hat{\gamma}_{11}$ | | | | | | 0.963 | -0.857 |
| $\hat{\gamma}_{12}$ | | | | | | | 0.964 |
| $\hat{c}_1$ | | | | | | | |

Table B.4: Canadian lynx series: correlation matrix for the parameter estimates.

| | linear model (iii) | ANN fully estimated (ii) | STAR model (i) |
|---|---|---|---|
| MSE | 0.09 | 0.09 | 0.06 |
| MAD | 0.25 | 0.23 | 0.19 |
| MAPE | 0.09 | 0.08 | 0.07 |

Table B.5: Canadian lynx series: one-step ahead measures of forecasting accuracy: MSE, MAD, MAPE. Note: (i), (ii) and (iii) mark the ranking of each measures in the table from best to worse.

| | | Linear model (i) | | | | ANN fully estimated (i) | | | | STAR (iii) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | -1 | 0 | 1 | Tot | -1 | 0 | 1 | Tot | -1 | 0 | 1 | Tot |
| The | -1 | 9 | 0 | 0 | 9 | 8 | 1 | 0 | 9 | 9 | 0 | 0 | 9 |
| actual | 0 | 8 | 9 | 1 | 18 | 5 | 12 | 1 | 18 | 6 | 9 | 3 | 18 |
| data | 1 | 1 | 6 | 10 | 17 | 1 | 4 | 12 | 17 | 0 | 6 | 11 | 17 |
| | Tot | 18 | 15 | 11 | 44 | 14 | 17 | 13 | 44 | 15 | 15 | 14 | 44 |
| | | Linear model (iii) | | | | ANN fully estimated (i) | | | | STAR (iii) | | | |
| Confusion rate | | 0.34 | | | | 0.28 | | | | 0.34 | | | |

Table B.6: Canadian lynx series: one-step ahead measures of forecasting accuracy: dispersion matrix and confusion rate CR. Note: (i), (ii) and (iii) mark the ranking of each measures in the table from best to worse.

| | ANN fully estimated vs. linear model | ANN fully estimated vs. STAR |
|---|---|---|
| value | -0.27 | 1.84 |
| p-value* | 0.61 | 0.04 |
| Reject equality | No | Yes at the 95% level; No at 99% level |

Table B.7: Canadian lynx series: one-step ahead measures of forecasting accuracy: modified Diebold-Mariano test. * 1-side p-values

| | linear model (iii) | ANN fully estimated (ii) | STAR model (i) |
|---|---|---|---|
| MSE | 0.82 | 0.75 | 0.50 |
| MAD | 0.76 | 0.75 | 0.63 |
| MAPE | 0.27 | 0.26 | 0.22 |

Table B.8: Canadian lynx series: four-step ahead measures of forecasting accuracy: MSE, MAD, MAPE. Note: (i), (ii) and (iii) mark the ranking of each measures in the table from best to worse.

| | | Linear model (i) | | | | ANN fully estimated (i) | | | | STAR (iii) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | -1 | 0 | 1 | Tot | -1 | 0 | 1 | Tot | -1 | 0 | 1 | Tot |
| The | -1 | 16 | 0 | 0 | 16 | 14 | 2 | 0 | 16 | 16 | 0 | 0 | 16 |
| actual | 0 | 3 | 2 | 1 | 6 | 1 | 3 | 2 | 6 | 2 | 3 | 1 | 6 |
| data | 1 | 0 | 2 | 20 | 22 | 0 | 2 | 20 | 22 | 0 | 3 | 19 | 22 |
| | Tot | 19 | 4 | 21 | 44 | 15 | 7 | 22 | 44 | 18 | 6 | 20 | 44 |
| | | Linear model (i) | | | | ANN fully estimated (iii) | | | | STAR (i) | | | |
| confusion rate | | 0.14 | | | | 0.16 | | | | 0.14 | | | |

Table B.9: Canadian lynx series: four-step ahead measures of forecasting accuracy: dispersion matrix. Note: (i), (ii) and (iii) mark the ranking of each measures in the table from best to worse.

| | ANN fully estimated vs. linear model | ANN fully estimated vs. STAR |
|---|---|---|
| value | -0.49 | 1.98 |
| p-value* | 0.69 | 0.03 |
| Reject equality | No | Yes |

Table B.10: Canadian lynx series: four-step ahead measures of forecasting accuracy: modified Diebold-Mariano test at the 0.05 significance level. * 1-side p-values

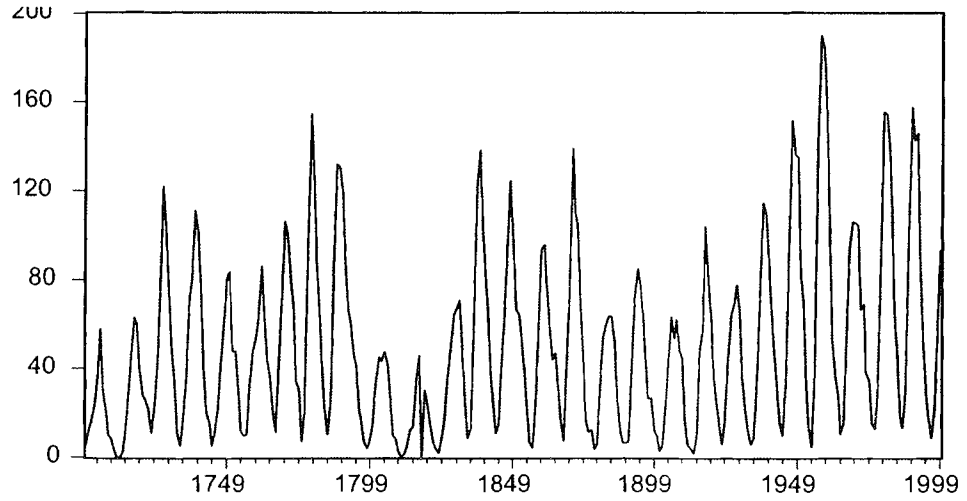


Figure B.15: The annual number of solar sunspots, 1700-1999. Observations preceding 1950 are used for estimation and the remaining ones for forecasting evaluation.

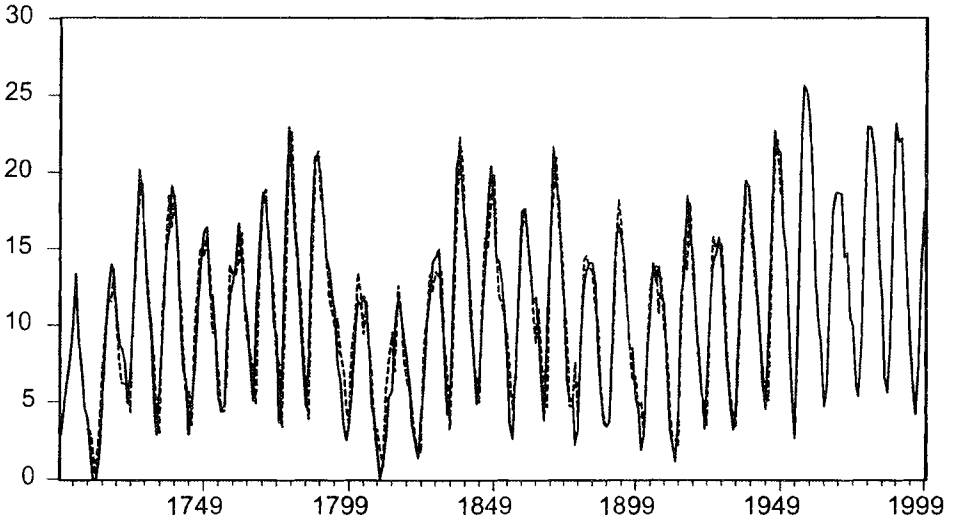


Figure B.16: Solid line: annual sunspot series (transformed by $2 \cdot \left(\sqrt{1+S_t} - 1\right)$), 1700-1999. Observations preceding 1950 are used for estimation and the remaining ones for forecasting evaluation. Dotted line: neural network model fit along the estimation period, 1700-1949.

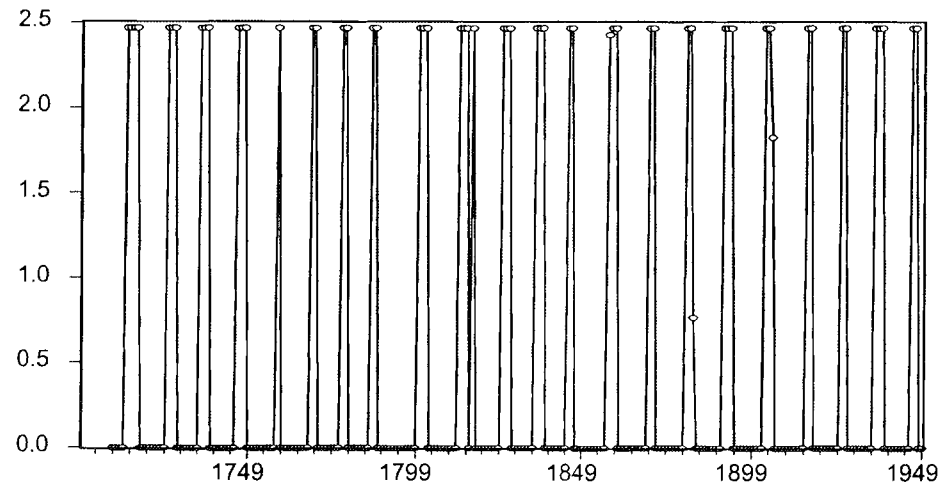


Figure B.17: Values of the first hidden unit for the conditional mean of the model estimated during period 1700-1949 for the transformed series of the annual solar sunspots.

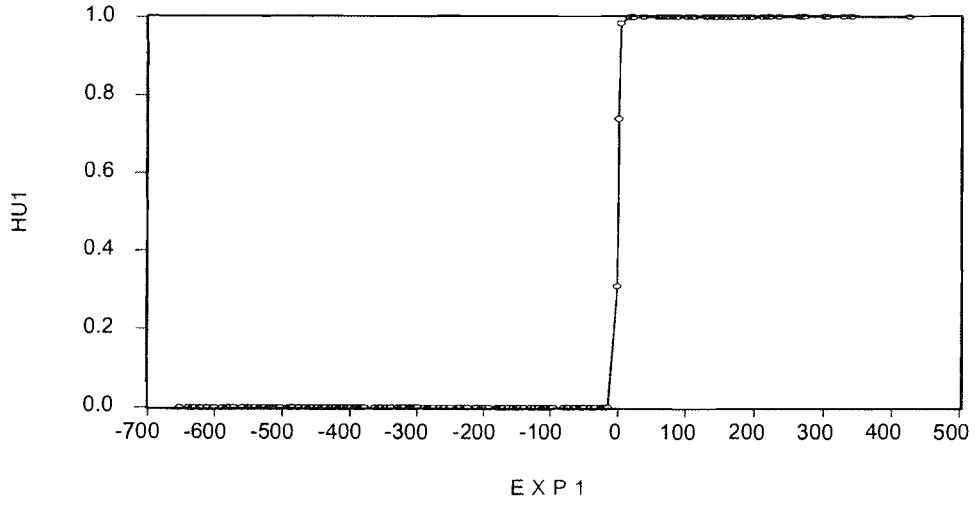


Figure B.18: Values of the estimated logistic function of the first hidden unit for the conditional mean of the model for the transformed series of the annual solar sunspots. Data are ordered by ascending values of its exponent.

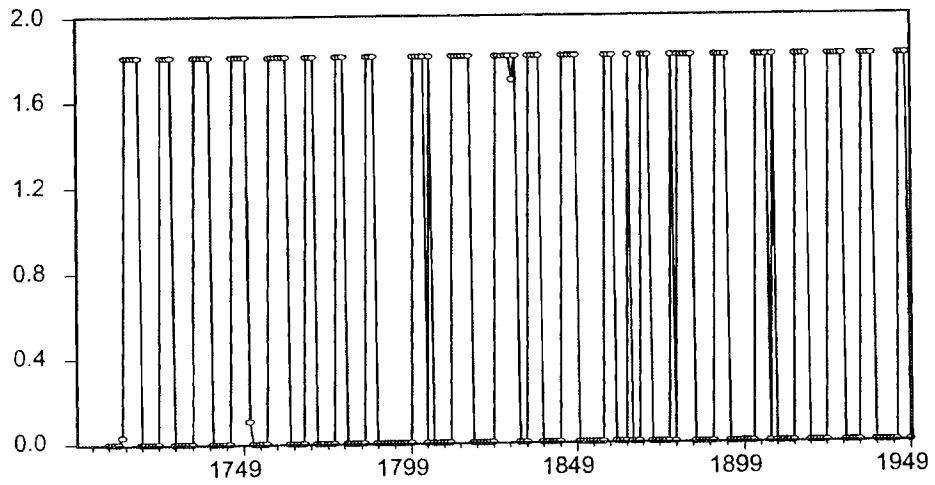


Figure B.19: Values of the second hidden unit for the conditional mean of the model estimated during period 1700-1949 for the transformed series of the annual solar sunspots.

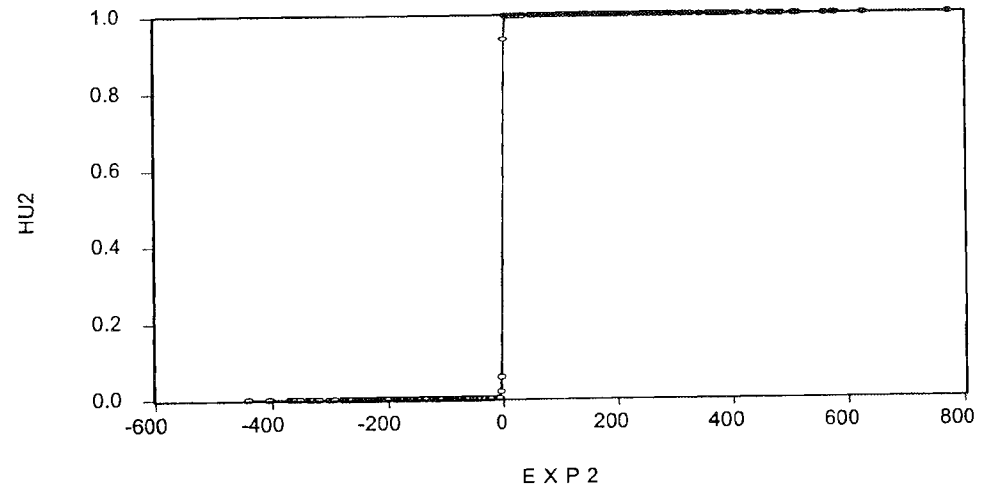


Figure B.20: Values of the estimated logistic function of the second hidden unit for the conditional mean of the model for the transformed series of the annual solar sunspots. Data are ordered by ascending values of its exponent.

| | r=1 | r=2 | r=4 | r=10 |
|---|---|---|---|---|
| serial independence | 0.14 (0.71) | 0.09 (0.92) | 1.04 (0.39) | 1.20 (0.29) |
| no ARCII | 3.27 (0.07) | 1.98 (0.14) | 1.37 (0.24) | 2.23(0.02) |
| | $LM_1$ | $LM_2$ | $LM_3$ | |
| parameter constancy | 1.18 (0.32) | 1.17 (0.31) | 1.07 (0.39) | |

Table B.11: Sunspot series: values of the LM tests for serial independence against AR(r) or MA(r) errors. Corresponding p-values are reported in parenthesis.

| | $\hat{\beta}_1$ | $\hat{\gamma}_{11}$ | $\hat{\gamma}_{12}$ | $\hat{\gamma}_{13}$ | $\widehat{c}_1$ |
|---|---|---|---|---|---|
| $\hat{\beta}_1$ | | -0.150 | 0.150 | -0.150 | -0.150 |
| $\hat{\gamma}_{11}$ | | | -1.000 | 1.000 | 1.000 |
| $\hat{\gamma}_{12}$ | | | | -1.000 | -1.000 |
| $\hat{\gamma}_{13}$ | | | | | 1.000 |
| $\widehat{c}_1$ | | | | | |

Table B.12: Sunspot series: approximated correlation matrix for the parameter estimates of the first hidden unit.

| | $\hat{\alpha}_0$ | $\hat{\alpha}_1$ | $\hat{\alpha}_2$ | $\hat{\alpha}_3$ | $\hat{\beta}_2$ | $\hat{\gamma}_{21}$ | $\hat{\gamma}_{22}$ | $\hat{\gamma}_{23}$ | $\widehat{c}_2$ |
|---|---|---|---|---|---|---|---|---|---|
| $\hat{\alpha}_0$ | | 0.468 | -0.649 | -0.630 | -0.450 | 0.048 | -0.048 | 0.043 | 0.052 |
| $\hat{\alpha}_1$ | | | -0.953 | -0.237 | -0.805 | 0.121 | -0.121 | 0.112 | 0.116 |
| $\hat{\alpha}_2$ | | | | 0.258 | 0.808 | -0.110 | 0.110 | -0.100 | -0.107 |
| $\hat{\alpha}_3$ | | | | | -0.061 | -0.041 | 0.041 | -0.045 | -0.045 |
| $\hat{\beta}_2$ | | | | | | -0.127 | 0.127 | -0.114 | -0.121 |
| $\hat{\gamma}_{21}$ | | | | | | | -1.000 | 0.997 | 0.997 |
| $\hat{\gamma}_{22}$ | | | | | | | | -0.997 | -0.997 |
| $\hat{\gamma}_{23}$ | | | | | | | | | 0.998 |
| $\widehat{c}_2$ | | | | | | | | | |

Table B.13: Sunspot series: approximated correlation matrix for the parameter estimates of the linear part and the second hidden unit, conditioned on the estimates of the first hidden unit.

| | linear model (ii) | ANN fully estimated (i) |
|---|---|---|
| MSE | 5.29 | 5.04 |
| MAD | 1.83 | 1.80 |
| MAPE | 0.18 | 0.17 |

Table B.14: Sunspot series: one-step ahead measures of forecasting accuracy: MSE, MAD, MAPE. Note: (i) and (ii) mark the ranking of each measures in the table from best to worse.

| | | Linear model (i) | | | | ANN fully estimated (i) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | -1 | 0 | 1 | Tot | -1 | 0 | 1 | Tot |
| The | -1 | 12 | 7 | 0 | 19 | 13 | 6 | 0 | 19 |
| actual | 0 | 7 | 9 | 1 | 17 | 7 | 9 | 1 | 17 |
| data | 1 | 0 | 3 | 11 | 14 | 0 | 4 | 10 | 14 |
| | Tot | 19 | 19 | 12 | 50 | 20 | 19 | 11 | 50 |
| | | Linear model (i) | | | | ANN fully estimated.(i) | | | |
| confusion rate | | 0.36 | | | | 0.36 | | | |

Table B.15: Sunspot series: one-step ahead measures of forecasting accuracy: dispersion matrix. Note: (i) and (ii) mark the ranking of each measures in the table from best to worse.

| | ANN fully estimated vs. linear model |
|---|---|
| value | 0.41 |
| p-value* | 0.66 |
| Reject equality | No |

Table B.16: Sunspot series: one-step ahead measures of forecasting accuracy: modified Diebold-Mariano test. * 1-side p-value

| | linear model (ii) | ANN fully estimated (i) |
|---|---|---|
| MSE | 68.15 | 64.56 |
| MAD | 7.23 | 6.99 |
| MAPE | 0.77 | 0.78 |

Table B.17: Sunspot series: four-step ahead measures of forecasting accuracy: MSE, MAD, MAPE. Note: (i) and (ii) mark the ranking of each measures in the table from best to worse.

| | | Linear model (i) | | | | ANN fully estimated (i) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | -1 | 0 | 1 | Tot | -1 | 0 | 1 | Tot |
| The | -1 | 26 | 1 | 0 | 27 | 25 | 2 | 0 | 27 |
| actual | 0 | 1 | 4 | 0 | 5 | 1 | 4 | 0 | 5 |
| data | 1 | 0 | 1 | 17 | 18 | 0 | 0 | 18 | 18 |
| | Tot | 27 | 6 | 17 | 50 | 26 | 6 | 18 | 50 |
| | | Linear model (i) | | | | ANN fully estimated.(i) | | | |
| confusion rate | | 0.06 | | | | 0.06 | | | |

Table B.18: Sunspot series: four-step ahead measures of forecasting accuracy: dispersion matrix. Note: (i) and (ii) mark the ranking of each measures in the table from best to worse.

| | ANN fully estimated vs. linear model |
|---|---|
| value | 0.47 |
| p-value* | 0.68 |
| Reject equality | No |

Table B.19: Sunspot series: four-step ahead measures of forecasting accuracy: modified Diebold-Mariano test. * 1-side p-value.

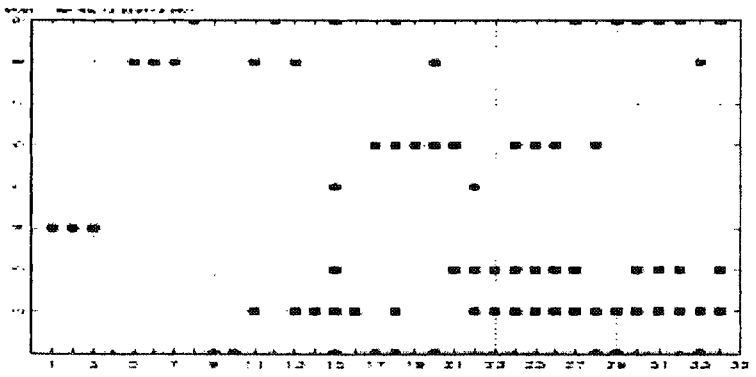

Figure B.21: S&P 500 excess return series: variables selected by the Taylor expansion procedure.

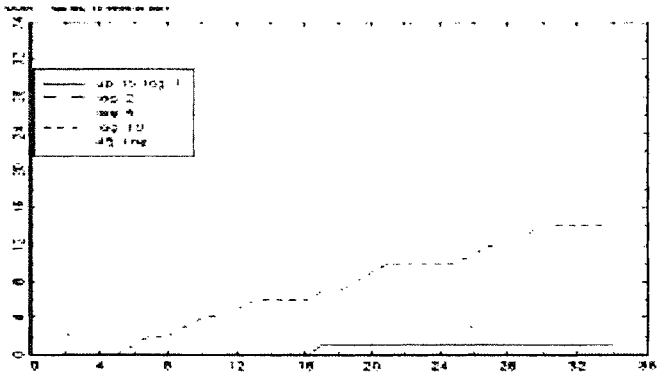

Figure B.22: S&P 500 excess return series: cumulative sum of the number of rejections of the null hypothesis of no linear autocorrelation.

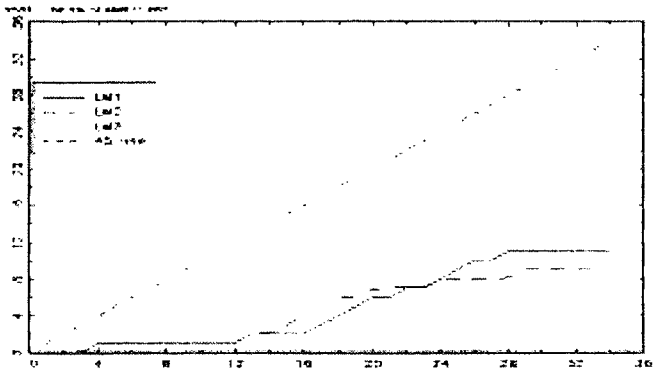

Figure B.23: S&P 500 excess return series: cumulative sum of the number of rejections of the null hypothesis of parameter constancy.

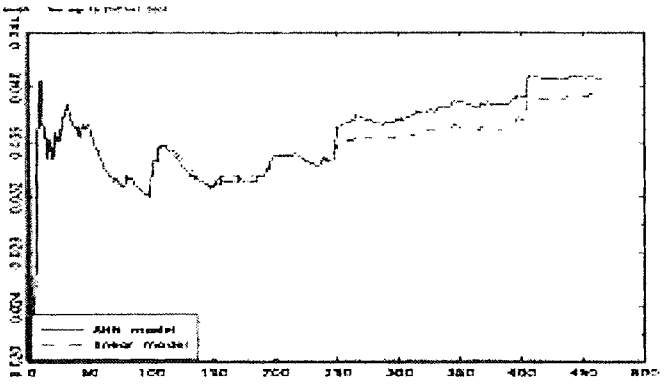

Figure B.24: S&P 500 excess return series: standard deviation of the residuals.

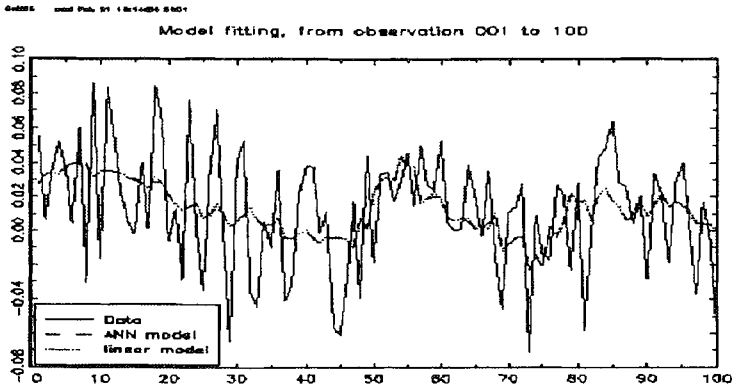


Figure B.25: S&P 500 excess return series: data, ANN model and linear fit: observations 0-100.

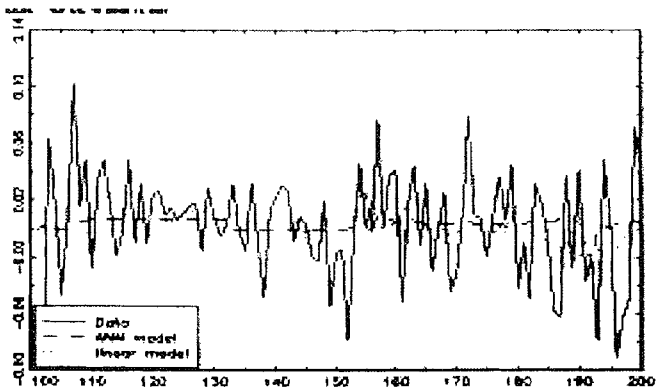


Figure B.26: S&P 500 excess return series: data, ANN model and linear fit: observations 101-200.

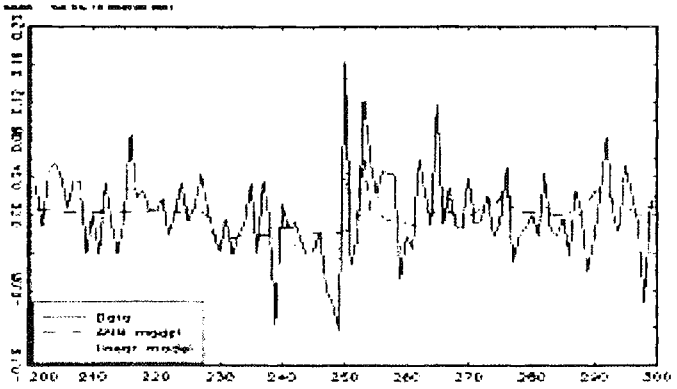


Figure B.27: S&P 500 excess return series: data, ANN model and linear fit: observations 201-300.

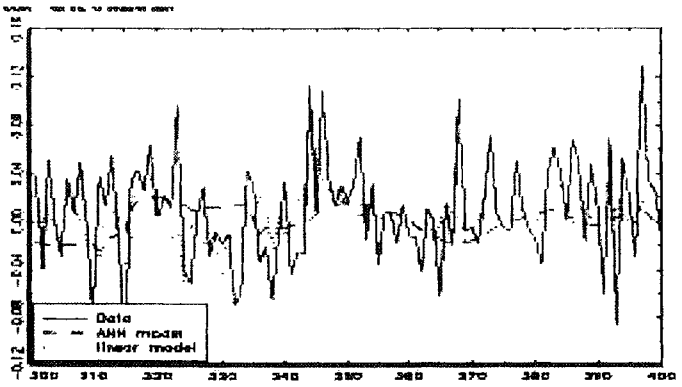


Figure B.28: S&P 500 excess return series: data, ANN model and linear fit: observations 301-400.

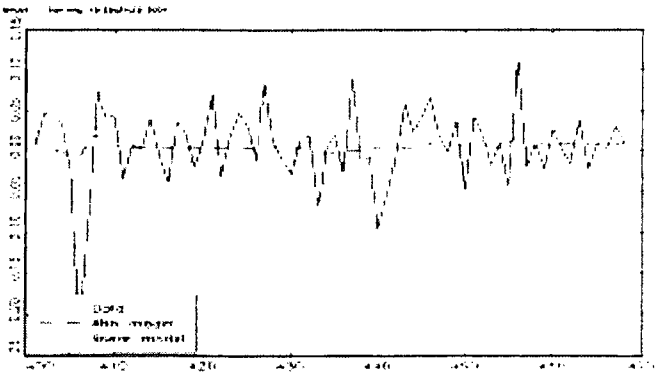

Figure B.29: S&P 500 excess return series: data, ANN model and linear fit: observations 401-468.

| | linear model (i) | ANN model (ii) |
|---|---|---|
| MSE | 0.0019 | 0.0022 |
| MAD | 0.0343 | 0.0352 |
| MAPE | 2.5769 | 2.2590 |

Table B.20: Excess return series: one-step ahead measures of forecasting accuracy: MSE, MAD, MAPE. Note: (i) and (ii) mark the ranking of each measures in the table from best to worse.

| | | Linear model (i) | | | | ANN model (ii) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | -1 | 0 | 1 | Tot | -1 | 0 | 1 | Tot |
| The | -1 | 42 | 40 | 0 | 82 | 27 | 50 | 0 | 77 |
| actual | 0 | 25 | 182 | 24 | 231 | 18 | 199 | 26 | 243 |
| data | 1 | 0 | 50 | 33 | 83 | 0 | 295 | 30 | 76 |
| | Tot | 67 | 272 | 57 | 396 | 45 | 295 | 56 | 396 |
| | | Linear model (i) | | | | ANN model.(i) | | | |
| confusion rate | | 0.35 | | | | 0.35 | | | |

Table B.21: Excess return series: one-step ahead measures of forecasting accuracy: dispersion matrix. Note: (i) and (ii) mark the ranking of each measures in the table from best to worse.

| | ANN model vs. linear model |
|---|---|
| value | -2.2070 |
| p-value* | 0.0139 |
| Reject equality | Yes at 95% No at 99% |

Table B.22: Excess return series: one-step ahead measures of forecasting accuracy: modified Diebold-Mariano test. * 1-side p-value

| Transaction costs | Mean return | St. deviation return | Final wealth ($) |
|---|---|---|---|
| Panel A: market portfolio | | | |
| Zero | 11.15 | 14.90 | 2503 |
| Low | 11.13 | 14.90 | 2463 |
| High | 11.11 | 14.89 | 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 portfolios based on the linear model | | | |
| Zero | 12.91 | 12.73 | 6780 |
| Zero (Racine's) | 13.66 | 10.08 | 7458 |
| Low | 9.97 | 11.31 | 2603 |
| Low (Racine's) | 12.21 | 10.18 | 4631 |
| High | 8.33 | 9.59 | 1527 |
| High (Racine's) | 11.23 | 10.34 | 3346 |
| Panel D: switching portfolios based on the ANN model | | | |
| Zero | 12.66 | 17.62 | 6132 |
| Zero (Qi's) | 15.59 | 10.49 | 13820 |
| Zero (Racine's) | 13.23 | 10.89 | 6265 |
| Low | 9.73 | 16.61 | 2360 |
| Low (Qi's) | 14.08 | 10.52 | 8420 |
| Low (Racine's) | 11.98 | 10.88 | 4205 |
| High | 8.48 | 15.36 | 1576 |
| High (Qi's) | 13.03 | 10.61 | 5963 |
| High (Racine's) | 11.23 | 10.93 | 3292 |

Table B.23: Excess return series: risks and profits of market, bond, and switching portfolios based on the one-step ahead forecast of the linear and ANN model 1960(1) to 1992(12). Results from Qi's and Racine's papers for panel C and D are reported as well.

# Essay 3

# Forecasting with artificial neural network models

## 3.1 Introduction

Over recent years, several nonlinear time series models have been proposed in the literature, see f.i. Tong (1990), Granger and Teräsvirta (1993), van Dijk, Teräsvirta, and Franses (2001). One model that has found a large number of successful applications is the Artificial Neural Network (ANN) model. The ANN model is a mathematical model inspired by the function of the human brain and its use is mainly motivated by its capability of approximating any Borel-measurable function to any degree of accuracy . This is pointed out in several papers and referred to in the introduction of essay II.

In this paper the forecasting performance of a number of artificial neural network (ANN) models is compared with that of linear models. The idea is to find out whether or not different techniques of specifying ANN models, such as pruning and statistical techniques, lead to vastly different models. If they do, then the next question is whether this matters from a forecasting point of view. Values of 30 economic and other time series are predicted using models obtained by a number of different modelling techniques and the forecasts compared using a number of different criteria. The plan of the paper is as follows. Section 3.2 introduces the ANN models. Section 3.3 discusses the the techniques used in this paper to build ANN models. Section 3.4 gives a brief discussion on linear ARMA models. The forecasting experiment is described in Section 3.5. Concluding remarks are made in Section 3.6.

## 3.2 Neural Network Model

The Artificial Neural Network (ANN) time series model is usually defined as

$$y_t = G(\mathbf{w}_t; \mathbf{\Psi}) = \beta_0 + \sum_{j=1}^{q} \beta_j \psi(\boldsymbol{\gamma}_j' \mathbf{w}_t) + u_t, \tag{3.1}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_q)'$, $\boldsymbol{\gamma}_j = (\gamma_{j0}, \gamma_{j1}, ..., \gamma_{j,k-1}, c_j)'$, $j = 1, ..., q$; $\mathbf{w}_t = (w_{1t}, w_{2t}, ..., w_{kt}, 1)' = (y_{t-1}, y_{t-2}, ..., y_{t-k}, 1)'$, $u_t$ is $n.i.d.$ and $\psi(\boldsymbol{\gamma}_i' \mathbf{w}_t)$ is the sigmoid function defined as

$$\psi(\boldsymbol{\gamma}_i' \mathbf{w}_t) = \frac{1}{1 + e^{-\boldsymbol{\gamma}_i' \mathbf{w}_t}}. \tag{3.2}$$

In applications to economic time series it is normally useful to include a linear component in (3.1). The resulting model is defined as in Granger and Teräsvirta (1993)

$$y_t = G(\mathbf{w}_t; \boldsymbol{\Psi}) = \boldsymbol{\alpha}'\mathbf{w}_t + \sum_{i=1}^{q} \beta_i \psi(\boldsymbol{\gamma}_i'\mathbf{w}_t) + u_t, \tag{3.3}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_k, \alpha_0)'$ is a vector of real coefficients and $u_t$ is $n.i.d.$. One of the characteristic features of this model is that it is not globally identified. There are three characteristics of neural networks which cause the non-identifiability. The first one is due to the symmetries in the neural network architecture. The likelihood function of the model will be unchanged if one permutes the hidden units, resulting in $q!$ possibilities for each one of the coefficients of the model. The second reason is caused by the fact that $\psi(x) = 1-\psi(-x)$. The third reason is the mutual dependence of the parameters $\beta_i$ and $\boldsymbol{\gamma}_i$, $i = 1, \ldots, q$. If $\beta_i = 0$, the corresponding $\boldsymbol{\gamma}_i$ can assume any value without affecting the value of the likelihood function. On the other hand, if $\boldsymbol{\gamma}_i = 0$, then $\beta_i$ can take any value. Nevertheless, under certain regularity conditions, the maximum likelihood estimates of the parameters of (3.3) are consistent and asymptotically normal; for a broader discussion, see essay II of this thesis.

## 3.3 Modelling

### 3.3.1 What is Modelling?

In this paper, I consider modelling and forecasting of time series with ANN models. Modelling in this context is taken to mean model specification, parameter estimation, and, finally, evaluation of the estimated model. The idea is to find out whether or not different modelling techniques lead to vastly different ANN models. If they do, the next question is whether this matters from a forecasting point of view. I compare well-known techniques in the neural network literature, such as, *early stopping*, *pruning*, and *regularization*, with a novel one based on statistical tests. It is one of the few techniques making use of in-sample evaluation. The other techniques considered in this work rely on out-of-sample evaluation, i.e., checking the forecasting performance of the specified and estimated model.

In this paper, specification consists of selecting the variables of the ANN model from a set of candidates and determining the number of hidden units. This stage is of crucial importance in order to avoid overfitting and to find an adequate approximation of the true data generating process. In most neural network applications, it is customary to carry out the specification using some "rule of thumb". A vast number of ANN models with different combinations of variables and number of hidden units are estimated, and the one with the best performance according to some known criterion is chosen as the final specification. The statistical approach is based on a sequence of hypothesis tests and gradually increasing the size of the ANN model until the test results suggest one to stop.

With the exception of methods of early stopping and regularization, estimation of parameters in ANN models is carried out by means of nonlinear least squares (NLS) or conditional maximum likelihood (ML). If the errors are normal, independent, and identically distributed, as assumed above, these two methods are equivalent. Hence, the parameter vector $\boldsymbol{\Psi}$ of (3.1) or (3.3) is estimated as

$$\hat{\boldsymbol{\Psi}} = \underset{\boldsymbol{\Psi}}{\operatorname{argmin}} Q_T(\boldsymbol{\Psi}) = \underset{\boldsymbol{\Psi}}{\operatorname{argmin}} \sum_{t=1}^{T} (y_t - G(\mathbf{w}_t; \boldsymbol{\Psi}))^2 . \tag{3.4}$$

In most applications a simple gradient descent algorithm (backpropagation) is used to estimate the parameters. However, the estimation of $\Psi$ is usually not easy (*Hush* (1999)), and in general the optimization algorithm is very sensitive to the choice of the starting-values of the parameters. The use of algorithms such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm or the Levenberg-Marquardt are strongly recommended for the final estimation. See Bertsekas (1995) for details about optimization techniques.

Any estimated model must be evaluated. In econometrics, model misspecification tests play an important role in model evaluation. Most often, these tests concern the estimated residuals of the model. The residuals are used for testing assumed properties of the error process, such as, serial independence and normality. Testing parameter stability is another important model evaluation test. In typical ANN applications, out-of-sample forecasting appears to be the only evaluation tool available. This is, however, not surprising because the asymptotic properties of the parameter estimators are often unknown. The statistical ANN modelling procedure considered here (referred hereafter by SA) is intended to fill the gap between econometric model building and ANN modelling practices by proposing a number of misspecification tests for ANN model building.

### 3.3.2 ANN Modelling Based on a Statistical approach (SA)

The capability of single hidden-layer feedforward neural networks (hereafter NN) to approximating any Borel-measurable function to any degree of accuracy has been pointed out in Hornik, Stinchcombe, and White (1989). Nonlinear features in time series can then be successfully modelled applying statistical tools to the data of interest, since the connection between NN and statistics is generally well accepted. I call the following model an autoregressive NN model of order $k$ with $q$ hidden units and a linear component:

$$y_t = \boldsymbol{\alpha}'\mathbf{w}_t + \sum_{j=1}^{q} \beta_j \psi(\boldsymbol{\gamma}_j'\mathbf{w}_t) + u_t, \ t = 1, ..., T \tag{3.5}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_k, \alpha_0)'$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_q)'$, $\boldsymbol{\gamma}_j = (\gamma_{j0}, \gamma_{j1}, ..., \gamma_{j,k-1}, c_j)'$, $j = 1, ..., q$; $\mathbf{w}_t = (w_{1t}, w_{2t}, ..., w_{kt}, 1)' = (y_{t-1}, y_{t-2}, ..., y_{t-k}, 1)'$ and $u_t$ is $n.i.d.$. Here, I assume for simplicity that all lags from 1 to $k$ enter model (3.5).

Summing up, the overall device works as follows. The set of variables to be included in the model is selected as in Rech, Teräsvirta, and Tschernig (2001). The hypothesis of no nonlinear hidden units (linear model) is tested at a given significance level $\alpha$. If rejected, a model with a linear part and one hidden unit is estimated and the approximate t-values of the parameters computed, approximating the covariance matrix of the parameters by the outer product of the gradient matrix. The lags with low t-values are removed and the model re-estimated. The whole procedure is redone until the hidden unit contains only significant estimates. Subsequently, the hypothesis of no additional hidden units is tested at the significance level $\alpha/2$. If rejected, a model with two hidden units is estimated and the dimension of the model reduced by checking the t-values of its estimates as above. The procedure continues halving the significance level again to $\alpha/4$, $\alpha/8$, ..., stopping the procedure at the first acceptance of the null hypothesis of no additional hidden units. Letting the significance level converge to zero as $q \to \infty$ keeps the dimension of the model under control.

Evaluating a model requires, as in Eitrheim and Teräsvirta (1996), to develop specific LM tests for the hypotesis of no error autocorrelation and parameter constancy, while additional

nonlinearity is already checked when I choose the number of hidden units. The test for error autocorrelation is based on model (3.5), where the residuals $u_t$ follow an autoregressive process of order $r$, $u_t = \sum_{j=1}^{r} a_j u_{t-j} + \varepsilon_t$, $\varepsilon_t \sim n.i.d.(0, \sigma^2)$. The corresponding $LM$ test for the hypotesis $H_0 : \mathbf{a} = 0$ can be carried out in 3 steps as in testing for $q$ against $q+1$ hidden units. As to parameter constancy, I generalize model (3.5) assuming that the hidden units have constant parameters whereas both $\beta$s and $\alpha$s may change smoothly over time. Therefore $\boldsymbol{\alpha} = \boldsymbol{\alpha}(t) = \alpha_0 + \boldsymbol{\lambda}_2 F_1(t, \gamma_1, c_1)$ and $\boldsymbol{\beta}(t) = \beta_0 + \boldsymbol{\lambda}_1 F_j(t, \gamma_j, c_j)$, where the $F_j(\cdot)\, s$ are transitional functions monotonically increasing and bounded between zero and one. The null hypothesis of parameter constancy implies that $F_j(t, \gamma_1, c_1) \equiv$ constant for any $t$. For further details on the modelling procedure, see essay II of this thesis.

### 3.3.3 Early Stopping

In this paper, I compare forecasts from models obtained applying the modelling cycle of Section 3.3.2 to forecasts from ANN models of the form of (3.1) obtained by other means. The first alternative method to be considered is the so-called *early stopping*. It is perhaps one of the most simple and vastly used modelling technique in the ANN framework. The key idea is to split the available data into three subsets. The first subset is used to estimate the parameters. The second subset is called the validation set. The error on the validation set is monitored during the estimation process. When the network begins to overfit the data, the error on the validation set typically begins to rise. When the validation error increases for a specified number of iterations, the estimation process is discontinued, and the parameters estimated at the minimum of the validation error serve as final estimates.

The test set is not used for estimation, but it is saved for comparing different models. A large number of different specifications are estimated and compared by means of the out-of-sample performance. The model with the best forecasting performance is chosen as the final specification.

### 3.3.4 Pruning

Pruning is another popular technique. The objective of pruning is to find the smallest network that fits the data well and produces good forecasts. I consider three pruning techniques. For a general survey on pruning see Reed (1993). The first two, information criterion pruning and cross-validation pruning, are described in detail in Anders and Korn (1999). The third one, interactive pruning, is based on the results of Kaashoek and van Dijk (1998), Siestma and Dow (1991), and Siestma and Dow (1988).

Although there are many other pruning algorithms, the three to be considered here are the one most frequently used in applications.

To estimate the parameters, I used the BFGS optimization procedure with the starting-values given by the Nguyen-Widrow algorithm Nguyen and Widrow (1990).

#### Information Criterion Pruning

The idea of the Information Criterion Pruning (ICP) is to choose between two models with different degrees of complexity. This is done using information criterion (IC) such as, AIC Akaike (1974) or SBIC(Rissanen (1978); Schwarz (1978)). As pointed out in Anders and Korn (1999), such information criteria are derived based on the assumption of asymptotic normality of the maximum likelihood estimators in globally identified models. Hence, their

use is not theoretically justified in the neural network case when the ANN model may be over-parametrized and thus unidentified. To circumvent this problem, Anders and Korn (1999) suggested the following methodology to estimated ANN models based on IC measures.

First, estimate the ANN model with only one hidden unit. Next, compute the residuals and regress them on a third-order Taylor expansion of an additional hidden unit. Calculate the value of information criterion and accept the additional hidden unit if the value of the IC is less than the IC of a white noise. Continue adding hidden units until there is no further improvement. Note that this resembles the strategy presented in Section 3.3.2. However, in determining the number of hidden units, information criteria are used in place of statistical tests.

The second step of the procedure starts from the fully connected network obtained in the first step and tries to detect irrelevant input connections. One thus proceeds in the opposite direction. All submodels with one of the input connections removed are estimated and compared with the full network by means of the IC. If the full network turns out to have the lowest IC value, this model is accepted as the final one. Otherwise the lowest IC submodel is chosen to serve as the baseline for the next round of the specification search. Again all submodels having one input connection less than the baseline model are compared with the latter using an appropriate IC.

It is also possible to start with a network with a large number of hidden units and proceed by removing units one at a time, if necessary. This "general to specific" alternative may involve estimating a considerable amount of unidentified or near-unidentified models and is not applied in this work.

### Cross-Validation Pruning

The idea of the cross-validation pruning (CVP), or, more specifically, leave-$k$-out cross-validation, is the following. In order to find an optimal number of hidden units and the correct set of input variables, it is appealing to compare the mean squared prediction errors (MSPE) of different model specifications. Such prediction errors are obtained by dividing the sample into $M$ subsets, which contain $k$ observations each. The ANN model is repeatedly re-estimated, leaving out a different subset each time. The average MSPE on the $M$ subsets that have been left out defines the cross-validation error, CV:

$$CV = \frac{1}{M}\sum_{m=1}^{M} MSPE_m, \tag{3.6}$$

where

$$MSPE_m = \frac{1}{k}\sum_{i=1}^{k}(\hat{y}_i - y_i)^2. \tag{3.7}$$

The model with the lowest cross-validation error is finally chosen. The CVP is independent of probabilistic assumptions. The main disadvantage of the CVP method is that re-estimating the models for each one of $M$ subsets can be time-consuming. In principle, all combinations of input variables and hidden units can be compared which requires plenty of computational resources.

The CVP is carried out in two stages. The first one consists of estimating the number of hidden units and the second one of removing any irrelevant input variables. The first step is summarized as follows. Start fitting a model with only one hidden unit and compute the value of $CV$. Add another hidden unit to the model and compute the CV again. Compare

the new CV value with the previous one. If the former is smaller than the latter, add another hidden unit. Otherwise discontinue the process and proceed to the second stage, which is the following. Compute the value of the cross-validation error for all submodels containing one input connection less than the original model. If at least one submodel turns out to have an smaller CV than the full model, remove that variable and continue without it. Otherwise terminate the process. The finally chosen ANN model is such that no variable can be removed without increasing the cross-validation error.

As in the ICP method, it is also possible to start with a network with a large number of hidden units and proceed by removing one unit at each round. Again, this is not done in this paper. Several papers have pointed out arguments for and against the CVP.

For example, Stone (1977) presented an insightful discussion of the limitations of cross-validation. See also Zhu and Rohwer (1996), Goutte (1997), and Rivals and Personnaz (1999).

#### Interactive Pruning

Another way of identifying irrelevant units and/or inputs is the so called interactive pruning (IP). The main idea is to start by estimating an ANN model with a large number of both hidden units and variables. After estimating the parameters of the model, the incremental contribution of each hidden unit is considered. This is done in several ways. Siestma and Dow (1991) and Siestma and Dow (1988) tackled the problem by removing hidden units with variance close to zero ($< 10^{-8}$) and combining units with highly correlated responses ($\rho = 0.98$, for example). Kaashoek and van Dijk (1998) suggested a different approach based on the correlation coefficient of the fitted values $\hat{y}_t$, defined as

$$R^2 = 1 - \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{T}(y_t - \overline{y})^2}, \tag{3.8}$$

where $\overline{y} = (1/T)\sum_{t=1}^{T} y_t$. The idea is to remove the $i^{th}$ hidden unit, $i = 1, \ldots, q$, re-estimate the model, and compute the value of the percentage difference in $R^2$, defined as

$$\Delta R^2 = \frac{R^2 - R_i^2}{R^2}, \tag{3.9}$$

where $R_i^2$ is the $R^2$ without the $i^{th}$ hidden unit. However, re-estimation means that only the vector of "connection strengths", $\beta$ in (3.1) is re-estimated each time. If $\Delta R^2$ is small ($\leq 1\%$), I remove the unit from the model. After removing the hidden units deemed unimportant, the procedure is continued by removing one input variable each time and comparing the correlation coefficients. Irrelevant variables are removed from the model following the same procedure as before, based on $\Delta R^2$. Again, only the connection strenght vector $\beta$ is re-estimated each time.

The main drawback of this method is that estimating an ANN model with a large number of inputs and hidden units is very difficult and quite often the estimation algorithm does not converge.

### 3.3.5 Regularization

The last modelling technique considered here is regularization. The idea, familiar from statistics, is to find a balance between the number of parameters and goodness of fit by penalizing large models. The objective function is modified in such a way that the estimation

algorithm effectively prunes the network by driving irrelevant parameter estimates to zero during the estimation process. The parameter vector $\boldsymbol{\Psi}$ is estimated as

$$\hat{\boldsymbol{\Psi}} = \underset{\boldsymbol{\Psi}}{\operatorname{argmin}} \tilde{Q}_T(\boldsymbol{\Psi}) = \underset{\boldsymbol{\Psi}}{\operatorname{argmin}} \left( Q_T(\boldsymbol{\Psi}) + \eta Q_T^*(\boldsymbol{\Psi}) \right), \tag{3.10}$$

where $Q_T(\boldsymbol{\Psi}) = \sum_{t=1}^{T} (y_t - G(\mathbf{w}_t; \boldsymbol{\Psi}))^2$ as in (3.4), $Q_T^*(\boldsymbol{\Psi})$ is the *regularization* or *penalty term*, and $\eta > 0$ is often called the *decay constant*.

To be in agreement with the terminology often used in the ANN literature, in this section "weight" means parameter and "generalization" is a synonym to out-of-sample forecasting.

### Weight Decay

The usual penalty is the sum of squared parameters times a constant. In a linear model, this form of regularization, called *weight decay*, is equivalent to ridge regression. For the ANN model (3.1), the parameters are estimated as

$$\hat{\boldsymbol{\Psi}} = \underset{\boldsymbol{\Psi}}{\operatorname{argmin}} \tilde{Q}_T(\boldsymbol{\Psi}) = \underset{\boldsymbol{\Psi}}{\operatorname{argmin}} \left( \sum_{t=1}^{T} (y_t - G(\mathbf{w}_t; \boldsymbol{\Psi}))^2 + \eta \sum_{j=1}^{n} \Psi_i^2 \right), \tag{3.11}$$

where $\Psi_i$ is a typical element of the parameter vector $\boldsymbol{\Psi}$ and $n$ is the number of parameters of the model.

The penalty term in (3.11), by definition, penalizes large estimates. There exist other regularization methods where the penalty term involves not only the parameter estimates but various partial derivatives of $G(\mathbf{w}_t; \boldsymbol{\Psi})$ as well.

The weight decay penalty term makes the parameter estimates smaller in absolute value than they otherwise would be. Large estimates can affect the forecasting performance in two different ways. Very large parameters estimates of $\gamma_{ij}$, $i = 1, \ldots, q$, $j = 0, \ldots, p$, in (3.1) or (3.3) lead to hidden units that cause $G(\mathbf{w}_t; \boldsymbol{\Psi})$ to be too rough, possibly with near discontinuities. Excessively large estimates of $\boldsymbol{\beta}$ can cause outputs far beyond the range of the data, if the output activation function is not bounded to the same range as the data. To put it another way, large parameter estimates can cause excessive variance of the output (Geman, Bienenstock, and Doursat (1992)). According to Bartlett (1997), the size ($L_1$ norm) of the parameter estimates is more important than the number of parameters in determining the out-of-sample performance of the estimated ANN model.

Other penalty terms besides the sum of squared weights are sometimes used.

### Weight Elimination

Weigend, Rumelhart, and Huberman (1991) proposed the *weight elimination* where

$$\hat{\boldsymbol{\Psi}} = \underset{\boldsymbol{\Psi}}{\operatorname{argmin}} \tilde{Q}_T(\boldsymbol{\Psi}) = \underset{\boldsymbol{\Psi}}{\operatorname{argmin}} \left( \tilde{Q}_T(\boldsymbol{\Psi}) = \sum_{t=1}^{T} (y_t - G(\mathbf{w}_t; \boldsymbol{\Psi}))^2 + \eta \left( \frac{\sum_{j=1}^{n} \Psi_i^2}{1 + \sum_{j=1}^{n} \Psi_i^2} \right) \right). \tag{3.12}$$

While the penalty term using the sum of squared weights tends to shrink the large estimates more than the small ones, weight elimination tends to shrink the small coefficients more, and is therefore more useful for suggesting subset models (pruning).

The forecasting ability of the ANN model can depend crucially on the decay constant $\eta$, especially with small sets of data used for estimation. If $\eta$ is too small, the network may still

overfit, and if it is too large, the ANN model does not have an adequate fit in the estimation period. Usually, different types of parameters in the ANN model will usually require different decay constants for good forecasting ability.

One approach to choosing the decay constant is to estimate several networks with different values of $\eta$ and choose the one that minimizes the sum of squared out-of-sample residuals. Weigend, Rumelhart, and Huberman (1991) iteratively update the decay constant during estimation. Adjusting all these decay constants to produce the model with the best forecasting ability often requires vast amounts of computation.

Fortunately, there is a superior alternative to weight decay and weight elimination: the *Bayesian regularization*. Bayesian regularization makes it possible to estimate efficiently numerous decay constants.

### Bayesian Regularization

One approach to determining the optimal regularization parameter $\eta$ is the Bayesian framework of MacKay (1992). In this context the coefficients of the network are assumed to be random variables with well-specified distributions. The regularization parameters are related to the unknown variances associated with these distributions and can be estimated with statistical techniques. Foresee and Hagan (1997) gave a detailed discussion of the use of Bayesian regularization in combination with the Levenberg-Marquardt optimization algorithm. The main advantage of this method is that even if the ANN model is over-parametrized, the irrelevant parameter estimates are likely to be close to zero and the model behaves like a small network.

## 3.4 AR approximation to Box & Jenkins' ARMA modelling

The ARMA$(p_1, p_2)$ model is defined as

$$(1 - \rho_1 L - \rho_2 L^2 - \ldots - \rho_{p_1} L^{p_1}) y_t = (1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_{p_2} L^{p_2}) u_t, \tag{3.13}$$

where $\rho_1, \ldots, \rho_{p_1}, \theta_1, \ldots, \theta_{p_2}$ are real parameters, $L$ is the backshift operator defined as $L^i y_t = y_{t-i}$, $p_1$ is the autoregressive order, $p_2$ is the moving-average order, and $u_t \sim \text{NID}(0, \sigma_u^2)$. Stationarity and invertibility are basic assumptions. Under such conditions, we may write (3.13) in the autoregressive representation, that is, an $AR(p = \infty)$ process:

$$(1 - \rho_1 L - \rho_2 L^2 - \ldots - \rho_{p_1} L^{p_1})(1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_{p_2} L^{p_2})^{-1} y_t = u_t, \tag{3.14}$$

or, equivalently:

$$(1 - a_1 L - a_2 L^2 - \ldots) y_t = u_t, \tag{3.15}$$

A linear model of the type (3.15), with a finite number of lags, can approximate (3.13) satisfactorily if a sufficient number of lags is encompassed. I choose the maximum lag $p = 25$ because several series are monthly and their need to be investigated at least 2 years back in time. The estimation can be performed by ordinary least squares (herafter OLS).Hereafter, I call such modelling procedure "AR" and I shall refer to it by this abbreviation.

The AR model building procedure is summarized as follows:

1. Specification

(a) If necessary, take logarithm or power transformations.

(b) If necessary, take short and/or long (seasonal) differences of $y_t$ to achieve stationarity in the mean.

(c) Determine the order of the AR model, i. e., choose the value of $p$. This is done by examining the partial autocorrelation (PACF) functions of the series after taking the differences and transforming the data (if it is necessary).

2. Estimate the parameters of model (3.15) by OLS.

3. Model evaluation.

(a) Test whether the errors are white noise (Ljung-Box statistic).

(b) Test whether the errors are normally distributed (skewness, kurtosis and the Jarque-Bera test).

4. Reduce the dimension of the estimated model using some information criterion like SBIC or AIC.

## 3.5 The Experiment

ANN models are mainly used for forecasting. To assess the practical usefulness of different methods of building an ANN model and the linear ARIMA model I conduct a forecasting experiment. The forecasts made by each estimated model are compared according to several statistics. I estimated ANN models for 30 different time series using the following model building procedures: statistical methods, information criterion pruning, cross-validation pruning, and Bayesian regularization. I also tried to estimate ANN models based on interactive pruning. However, in most of the cases considered here I was unable to estimate a satisfactory model. Thus, I discarded these results from the present paper. To compare the performance of the ANN models with the linear model, I also estimate, for each series, a linear autoregressive model.

### 3.5.1 Forecasting

The forecasts were made according to the following procedure.

1. Split the sample into two subsamples: the estimation set ($t = 1, \ldots, t_0$) and the forecasting set ($t = t_0 + 1, \ldots, T$).

2. Estimate the parameters of each model using only the estimation set.

3. For $t = t_0, \ldots, T - 12$, compute the out-of-sample forecasts of one to 12-step-ahead, $\hat{y}_t(h)$, and the associated forecast errors denoted by $\widehat{u}_t(h)$ where $h$ is the forecasting horizon.

4. For each forecasting horizon, compute the following performance measures:

(a) Normalized root mean squared error (nRMSE):

$$nRMSE(h) = \sqrt{\frac{1}{T - t_0 - 11} \frac{\sum_{t=t_0}^{T-12} \widehat{u}_t^2(h)}{\hat{\sigma}_y^2}}, \tag{3.16}$$

where $\hat{\sigma}_y^2$ is the estimated in-sample unconditional variance of the series. The $nRMSE(h)$ is a measure of the relative forecasting efficiency of the estimated model in relation to $\hat{y}_t(h) = \overline{y}$, where $\overline{y}$ is the in-sample mean of the series.

(b) Mean Absolute Error (MAE):

$$MAE(h) = \frac{1}{T - t_0 - 11} \sum_{t=t_0}^{T-12} |\widehat{u}_t(h)| \tag{3.17}$$

(c) Median Absolute Deviation (MAD):

$$MAD(h) = \text{median}(|\widehat{u}_t(h) - \text{median}(\widehat{u}_t(h)|). \tag{3.18}$$

Reporting the MAD was suggested by van Dijk (1999) and can be interpreted as a measure that is robust to outliers.

Another interesting way of comparing the forecasting performance of different models is to compare the sign of $(y_{t+h} - y_t)$ with the sign of $(\hat{y}_{t+h} - y_t)$. This can be done with a $2 \times 2$ contingency table, known as the *confusion matrix* and defined as

| | | actual | |
|---|---|---|---|
| | | *up* | *down* |
| predicted | *up* | $a_{11}$ | $a_{12}$ |
| | *down* | $a_{21}$ | $a_{22}$ |

(3.19)

The columns in (3.19) correspond to actual moves, up or down, while the rows correspond to predicted moves. In this way, the diagonal cells correspond to correct directional predictions, while the off-diagonal cells correspond to incorrect predictions. I measure the overall performance in terms of the *confusion rate*, defined as

$$CR = \frac{a_{12} + a_{21}}{a_{11} + a_{12} + a_{21} + a_{22}}. \tag{3.20}$$

In order to test the hypothesis of equal accuracy in forecast performance between two estimated models, I use the Diebold-Mariano statistic Diebold and Mariano (1995) with the correction proposed by Harvey, Leybourne, and Newbold (1997). Such statistics is described in essay II of this thesis.

### 3.5.2 The Data

The time series used in the forecasting experiment are summarized in Table A.1. For each of the 30 series, the applied transformation, the period of interest, the total number of observations $T$ and the number of observations $t_0$ utilized for estimating the models are given. The first twelve series were obtained from *Economagic* (www.economagic.com). The rest of the series were obtained from Rob J. Hyndman's Time Series Data Library home page (www-personal.buseco.monash.edu.au/ hyndman/TSDL/mhts/). Logarithmic transformations were applied to some of the series to stabilize the variance, and first or twelve-month differences were taken to achieve stationarity and/or to remove seasonal variation. The idea is to perform an investigation on a broad set of data with a different sample size, related to topics from macro and financial economics, to physics, to ecology and health. As to macroeconomic data, I included the M1, M2 and M3 series for US money stock, the unemployment

rate, the industrial production, the consumer and the producer price index for the US. The set of financial series includes 3 daily cross rates (Canadian \$/US \$, German mark/US \$, Japanese yen/US \$) and 2 monthly stock indexes (S&P 500 and Dow Jones). All these series are monthly, seasonally adjusted, and transformed by taking the first difference of logarithms, such that I model their approximated growth rates. One series represents the volcanic dust veil index, 4 series relate to river flows, 2 to blow-flies population, 2 to ozone concentration, 3 to temperatures (1 monthly and 2 yearly), and 3 to cases of chickenpox, measles and mumps. The two classical benchmarks modelled in essay II, the lynx and the annual sunspot series, are included too.

### 3.5.3 Estimation Results

Table A.2 shows the specification results. The set of selected lags widely differ depending on the different method employed. The SA approach is the most parsimonious both in terms of number of hidden units included in the estimated models and number of selected lags. It is also the only one which allows the detection of a linear model. Note that it leads to the detection of a linear model in 17 cases over 30. If I compare such linear models to the AR models estimated by OLS, I can notice that they are much more parsimonious. The AR approach detected even 10 lags in one case, while SA never detected more than 4 lags. Likewise, if I compare the nonlinear models estimated by the SA approach to the ICP and CVP methods (the BRP's specification is fixed), it is seen from table A.2 that SA is the most parsimonious. The question now is if this fact will influence the performances of the SA approach either in the sample fit (table A.3) or in the forecasting performances (tables A.4–A.11) or in both of them.

As to the in-sample fit, table A.3 shows the $nRMSE$ of the estimated models. SA models always have the largest $nRMSE$, but this is expected because the corresponding estimated models are the most parsimonious. As a whole, the BRP approach performs bettern than the others in 11 cases according to the nRMSE criterion, but this is not surprising because it is the less parsimonious one, with 12 lags and 4 hidden units in each estimated model: including more regressors gives, on the average, a better fit, but does not guarantee better forecasts. The ICP has the lowest residual mean square error 7 times, and the CVP and the AR follow, both with 6. Whether or not the larger models forecast better than the more parsimonious ones will be discussed in the next paragraph.

### 3.5.4 Forecasting Results

Forecasting with nonlinear models is far more complicated than in the linear case, especially in the case of multi-steps ahead forecasts. Techniques such as Monte-Carlo and Bootstrap should be used. See Granger and Teräsvirta (1993), p. 130-135 for a discussion. In this paper, I use the Monte-Carlo approach with 500 replications to produce multi-step forecasts from the nonlinear estimated models. Let $\mathbf{y}_t = (y_t, y_{t-1}, ..., y_{t-k})'$ the set of lags entering the estimated ANN function $g(\cdot)$ at time $t+1$, and $\hat{y}_{t+1}$ the forecast one steps ahead, its optimal value is computed as follows:

$$\hat{y}_{t+1} = E[y_{t+1}|I_t] = g(\mathbf{y}_t)$$

For $h > 1$, it still holds that $E[y_{t+2}|I_t] = g(\mathbf{y}_{t+1}), ...E[y_{t+h}|I_t] = g(\mathbf{y}_{t+h-1})$, but we still base our forecast on the information set at time $t$, $I_t$. It follows that f.i. in vector $\mathbf{y}_{t+1}$, we do not know $y_{t+1}$. So forth so on, in vector $\mathbf{y}_{t+h-1}$ we do not know $y_{t+h-1}, y_{t+h-2}, ..., y_{t+1}$,

and they have to be generated by taking their variances into account. The most sensible assumption is that $y_{t+1},...,y_{t+h-1}$ follow the same process than $y_t$, which we assume to be normal: $y_t \sim N\left(\mu,\sigma^2\right)$. It follows that we can generate forecasts at time $h>1$ by using random numbers drawn by the same distribution, $z^{(j)} \sim N\left(0,\sigma^2\right)$. Therefore,

$$\begin{aligned}
\hat{y}_{t+2} &= g(\mathbf{y}_{t+1}) = \frac{1}{500}\sum_{j=1}^{500} g\left(g\left(\mathbf{y}_t\right)+z_{t+2}^{(j)}\right) \\
\hat{y}_{t+3} &= g(\mathbf{y}_{t+2}) = \frac{1}{500}\sum_{j=1}^{500} g\left(g\left(g\left(\mathbf{y}_t\right)+z_{t+2}^{(j)}\right)+z_{t+3}^{(j)}\right) \\
&\cdots \\
\hat{y}_{t+h} &= g(\mathbf{y}_{t+h-1}) = \frac{1}{500}\sum_{j=1}^{500} g\left(g\left(...\left(g\left(\mathbf{y}_t\right)+z_{t+2}^{(j)}\right)...\right)+z_{t+h}^{(j)}\right)
\end{aligned}$$

where the sequences of random numbers $\left\{z_{t+2}^{(j)},...,z_{t+h}^{(j)}\right\}$, $j=1,...,500$, are drawn from the normal distribution, $z_{t+i}^{(j)} \sim N\left(0,\sigma^2\right)$, and are mutually independent for all $i=2,...,h$; $j=1,...,500$. Results are reported in tables A.4–15 and summarized in tables A.4–A.10. In order to examine the forecasting performances for the 5 different models at the 3 horizons h =1, 6 and 12, first I compute a set of indexes which allows to make a comparison in absolute value for nRMSE, MAE (table 12) and MAD and the confusion rate CR (table 13). Albeit this approach does not constitute any sort of valid statistical test, it helps to shed light on the relative merits of the models. Results from tables 12 and 13 are summarized in tables 4 and 5, where the number of times each model performs best are demonstrated for each h and each of the 4 criteria.

The hypothesis of equal accuracy of forecasting performances measured by the nRMSE criterion as a function of the forecasting error is tested, by means of the modified Diebold-Mariano test statistic $S^*$. Results are reported in table 14. The number of times each modelling strategy performs better versus another one is reported in table 6, 7 and 8 for h = 1, 6 and 12 respectively. Likewise, in table 15, I use $S^*$ based on the MAE criterion. Tables 9,10 and 11 summarize the results for the 3 different forecasting horizons.

Let's take the first series, the first difference of the logarithms of the monthly US M1 money stock, seasonally adjusted, for the period 1959.1-1999.7. As can be seen in table 2, the sets of selected lags widely differ depending on the different models. The number of hidden units is 1 for all the NN models but BRP which keeps them fixed to 4 in advance. From tables 12 and 13, first line for model 1 corresponding to h =1, the AR model ranks first as to the nRMSE criterion, MAE and MAD are the same for all the models, while BRP performs best as to the CR criterion. As to h=6 (second line), the BRP performs best as to the nRMSE, MAE and MAD are still the same for all the models, and CVP is the best accroding to CR. At h=12 (third line), AR has the lowest nRMSE, MAE and MAD the same for all the models, BRP has the lowest CR. Summarizing, for series 1 the BRP modelling strategy and the AR are the best for h=1, BRP and CVP for h=6, and BRP and AR at h =12.

In order to perform the analysis by using a statistically tool, the corrected Diebold-Mariano test statistic $S^*$, I look into table 14 first (nRMSE). For h=1, AR and BRP are still the models performing best. AR turns out to be significantly better than ICP at the 5% level and it is better than both SA and CVP slightly over the 10% level , while it cannot

be considered more accurate than BRP. BRP is better than SA at the 10% level, than ICP at the 1%, than CVP at the 5%. Now, performing the same analysis for h = 1 in table 15, the corresponding outcomes for MAE corroborates results in table 14. Proceeding now to the forecasting horizon h = 6, BRP outperforms 2 models and so does AR (nRMSE), while (MAE) BRP turns out to be better over 2 models and AR over 1. At h = 12, neither in table 14 or 15 is there a clear indication of the models perfoming best. Such outcome is expected, because forecasting accuracy obviously diminishes as h increases, and at h = 12 (one year ahead in the M1 series) it becomes difficult, if not impossible, to discriminate among the forecasting performances.

Such analysis is performed for each of the 30 series. A synthetic picture of the forecasting performances can be viewed in table 4 and 5. In the former, the AR methodology outperforms the other 5 ones both for h = 1, 6 and 12, both for the nRMSE criterion (11 ranks best for h =1; 10 for h = 6; 9 for h = 12) and for the MAE (11 times for h = 1; 10 for h = 6; 11 for h = 12). In the same table, neither of the 4 NN clearly outperforms any of the other. In the latter, the AR methodology loses its predominance both as to the MAD and CR criteria, but no clear dominant model can be observed.

If I now take tables 6-12, which summarize tables 14 and 15, I can find the confirmation of the outcomes of table 4. Looking at the column "sum wins - sum losses", for table 6 (nRMSE criterion, h = 1), I can see that the AR model performs best 24 times versus the other models as a whole. In table 9 (MAE, h = 1), the corresponding figure is 20. The AR methodology thus seems superior to the other 2 ones at the forecasting horizon h=1. As to other models, the ICP performs worse both in table 6 (-22) and table 9 (-18), while SA (+2 and +6) ranks second. As to h=6 (tables 7 and 10), results are less clear-cut. BRP is ranked first as to the nRMSE criterion (+16 vs. +15 for AR) but results change for the MAE (+5 vs +15 for AR). ICP is still the worst with -20 and -16, while SA performs quite badly too (-10 and -6). At h=12, it is still clear that AR is the best (+15 for nRMSE and +19 for MAE), but results for the other models are quite difficult to interpret, being all negative or slightly positive (ICP = -2 and -7; CVP = -14 and -14; BRP =-1 and +1; SA = -5 and +1). Therefore, from the analysis of table 6-11 I can conclude that, according to the nRMSE and MAE criteria and on the forecasting comparison based on the DM test on nRMSE and MAE, the AR methodology turns out to outperform all the other ones. Among those methods for NN modeling, none can be considered the best.

Going to table 5, where results for MAD (median absolute deviation) and CR (confusion rate, based on the comparison of the signs of $(y_{t+h} - y_t)$ with the sign of $(\hat{y}_{t+h} - y_t)$) are summarized, I can notice that the predominance of the AR models disappears. This may be due to the fact that that both MAD and the CR criteria are less sensitive to outlayers than nRMSE and MAE. At h = 1, AR ranks last as to MAD together with BRP, and last together with SA as to CR. AR turns out to be the best as to MAD and around average as to CR at h = 6, the second worst in MAD and the second best in CR at h = 12. As to SA, at h = 1 it ranks second best as to MAD and the worst as to CR. Because the CR criterion is a measure for the accuracy in forecasting the sign, and MAD is less sensitive to large forecasting errors than nRMSE and MAE, based on the median of the data and not on the mean, such 2 criteria are more useful to assess the goodness in forecasting the direction of the value of a stochastic process. In such view, the NN models clearly outperform the AR ones in forecasting one period ahead.

It is also of interest to see how the statistical methodology SA derived in this thesis performs in the case of financial series. In this forecasting exercise, I included 3 monthly series of exchange rates: Canadian $ / US $ (series 9), German mark / US $ (series 10), and

Japanese yes / US $. For such series, the SA approach leads to the acceptance of linearity (no neuran network model for any). The corresponding linear autoregressive model selected by the SBIC criterion is based on the first lag (see section 3.2 in this essay or section 3 in essay I for an extensive discussion). All the other approaches produce models which are much less parsimonious for each of the 3 series (see table 2). Furthermore, ranking the models according to their in-sample fit measured by the nRMSE criterion (see table 3), SA is ranked last for series 9,10 and 11. Now, if I look at the comparison for the forecasting performances for series 9, h = 1, corresponding to one month ahead, (table 14 and 15), it is clear that the hypothesis of equal accuracy between SA and ICP cannot be rejected either for nRMSE or MAE. This result holds for SA vs CVP, SA vs BRP, and SA vs AR as well. Series 10 and 11 perform similarly at h =1. In this exercise, the SA methodology leads to parsimonious but very efficient models for financial forecasting.

## 3.6 Conclusion

In order to draw some conclusion on the overall exercise, I summarize the analysis of the 3 monthly financial series first. As to those 3 special cases, I can conclude that 1) the well-known property of parsimonious models of being the best forecasters is confirmed, 2) the SA approach is able to detect a parsimonious model which performs better than the more complicated ones at the forecasting horizon h = 1 and 3) even when SA does not reject linearity and leads to the estimation of a linear model, such model is still competitive to the one estimated by the the AR approximation to Box & Jenkins' methodology More generally, results from table 4 and 5 do not disagree with Swanson and White in (Swanson and White (1997)). In their analysis, the authors found out that each model performs best in at least one case and that there is no clear "winner". Likewise, what happens in this forecasting exercise is the following: while the linear model can be considered the best as the nRMSE and MAE criteria (table 4), it cannot definetely outperform the other models as to the MAD and CR criterion.

An overall conclusion about the results of the SA modelling procedure in this forecasting exercise is that, albeit it does not perform better than the other approaches, it produces similar forecasting performances with the advantage of leading to the estimation of more parsimonious models.

# Bibliography


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

ANDERS, U., AND O. KORN (1999): "Model Selection in Neural Networks," *Neural Networks*, 12, 309–323.

BARTLETT, P. L. (1997): "For Valid Generalization, The Size of The Weights is More Important Than The Size of The Network," in *Advances in Neural Information Processing Systems 9*, ed. by M. C. Mozer, M. I. Jordan, and T. Petsche, pp. 134–140. The MIT Press.

BERTSEKAS, D. P. (1995): *Nonlinear Programming*. Athena Scientific, Belmont, MA.

DIEBOLD, F. X., AND R. S. MARIANO (1995): "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253–263.

EITRHEIM, Ø., AND T. TERÄSVIRTA (1996): "Testing the Adequacy of Smooth Transition Autoregressive Models," *Journal of Econometrics*, 74, 59–75.

FORESEE, F. D., AND M. . T. HAGAN (1997): "Gauss-Newton Approximation to Bayesian Regularization," in *IEEE International Conference on Neural Networks (Vol. 3)*, pp. 1930–1935, New York. IEEE.

GEMAN, S., E. BIENENSTOCK, AND R. DOURSAT (1992): "Neural Networks and the Bias/Variance Dilemma," *Neural Computation*, 4, 1–58.

GOUTTE, C. (1997): "Note on Free Lunches and Cross-Validation," *Neural Computation*, 9, 1211–1215.

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

HARVEY, D., S. LEYBOURNE, AND P. NEWBOLD (1997): "Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting*, 13, 281–291.

HORNIK, K., M. STINCHCOMBE, AND H. WHITE (1989): "Multi-Layer Feedforward Networks are Universal Approximators," *Neural Networks*, 2, 359–66.

HUSH, J. (1999): "Training a Sigmoidal Node is Hard," *Neural Computation*, 11(5), 1249–1260.

KAASHOEK, J. F., AND H. K. VAN DIJK (1998): "A Simple Strategy to Prune Neural Networks with an Application to Economic Time Series," Econometric Institute Report 9854/A, Econometric Institute - Erasmus University.

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

Nguyen, D., and B. Widrow (1990): "Improving the Learning Speed of 2-Layer Neural Networks by Choosing Initial Values of the Adaptive Weights," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 3, pp. 21–26.

Rech, G., T. Teräsvirta, and R. Tschernig (2001): "A Simple Variable Selection Method for Nonlinear Models," *Communications in Statistics, Theory and Methods*, 30, 1227–1241.

Reed, R. (1993): "Pruning Algorithms – A Survey," *IEEE Transactions on Neural Networks*, 4, 740–747.

Rissanen, J. (1978): "Modeling by Shortest Data Description," *Automatica*, 14, 465–471.

Rivals, I., and L. Personnaz (1999): "On Cross-Validation for Model Selection," *Neural Computation*, 11, 871–901.

Schwarz, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–64.

Siestma, J., and R. J. F. Dow (1988): "Neural Net Pruning – Why and How," in *IEEE International Conference on Neural Networks*, vol. 1, pp. 325–333.

——— (1991): "Creating Artificial Neural Networks that Generalize," *Neural Networks*, 4(1), 67–69.

Stone, M. (1977): "Asymptotics For and Against Cross-Validation," *Biometrika*, 64, 29–35.

Swanson, N. R., and H. White (1997): "Forecasting Economic Time Series Using Flexible versus Fixed Specification and Linear versus Nonlinear Econometric Models," *International Journal of Forecasting*, 13, 439–461.

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

van Dijk, D. (1999): "Smooth Transition Models: Extensions and Outlier Robust Inference," Ph.D. thesis, Tinbergen Institute, Rotterdam, The Netherlands, www.few.eur.nl/few/people/djvandijk/thesis/thesis.htm.

van Dijk, D., T. Teräsvirta, and P. H. Franses (2001): "Smooth Transition Autoregressive Models - A Survey of Recent Developments," *Econometric Reviews*, forthcoming.

Weigend, A. S., D. E. Rumelhart, and B. A. Huberman (1991): "Generalization by Weight-Elimination with Application to Forecasting," in *Advances in Neural Information Processing Systems 9*, ed. by R. P. Lippmann, J. Moody, and D. S. Touretzky. Morgan Kaufmann.

Zhu, H., and R. Rohwer (1996): "No Free-Lunch for Cross-Validation," *Neural Computation*, 8, 1421–1426.

# Appendix A

# Tables

| | Series | Description | Transformation | Period | $T$ | $t_0$ |
|---|---|---|---|---|---|---|
| 1 | M1 | US M1 money stock, seasonally adjusted | first diff of log | 1959.1-1999.7 | 486 | 365 |
| 2 | M2 | US M2 money stock, seasonally adjusted | first diff of log | 1959.1-1999.7 | 486 | 365 |
| 3 | M3 | US M3 money stock, seasonally adjusted | first diff of log | 1959.1-1999.7 | 486 | 365 |
| 4 | CUR | US civilian unemployment rate, seasonally adj. | first diff of log | 1959.1-1999.7 | 486 | 365 |
| 5 | IP | US industrial production, seasonally adjusted | first diff of log | 1959.1-1999.7 | 486 | 365 |
| 6 | CPI | US consumer price index, seasonally adjusted | first diff of log | 1959.1-1999.7 | 486 | 365 |
| 7 | PPI | US producer price index, seasonally adjusted | first diff of log | 1959.1-1999.7 | 486 | 365 |
| 8 | EPU | US electric power use, seasonally adjusted | first diff of log | 1972.1-1999.6 | 329 | 240 |
| 9 | EXCAUS | Exchange rate: Canadian $ to one US $ | first diff of log | 1871.1-1999.8 | 331 | 240 |
| 10 | EXGEUS | Exchange rate: German marks to one US $ | first diff of log | 1871.1-1999.8 | 331 | 240 |
| 11 | EXJPUS | Exchange rate: Japanese yens to one US $ | first diff of log | 1871.1-1999.8 | 331 | 240 |
| 12 | SP500 | S&P 500 monthly close | first diff of log | 1950.1-1999.9 | 596 | 480 |
| 13 | DJIA | Dow Jones Industrial Average monthly close | first diff of log | 1950.1-1999.9 | 596 | 480 |
| 14 | DVI | Volcanic dust veil index, northern hemisphere | $\log(\text{obs.} + 1)$ | 1500-1969 | 470 | 300 |
| 15 | ASKEW1 | M.ly riverflow: Sacramento R. at Keswick, CA | 12th diff. of the log | 1939.10-1960.7 | 252 | 170 |
| 16 | ASKEW2 | M.ly r.flow: Madison R. near Yellowstone, MT | 12th diff of the log | 1923.1-1960.1 | 444 | 300 |
| 17 | FISHER | Mean daily flow, cms. Fisher River , Dallas | logarithm | 01.01 1988 - 31.12.1991 | 1461 | 1000 |
| 18 | OLDMAN | Mean daily flow, cms. Oldman River, Brocket | logarithm | 01.01. 1988 - 31.121991 | 1461 | 1000 |
| 19 | SUNSP | Annual sunspot | $2\left[\sqrt{(1+\text{obse.}} - 1\right]$ | 1700-1998 | 299 | 200 |
| 20 | DEATHBF | Deaths in adult population of sheep blow-flies | logarithm | - | 318 | 200 |
| 21 | TOTALBF | Total population of sheep blow-flies | logarithm | - | 319 | 200 |
| 22 | AZUSA | Ozone concentration, AZUSA | logarithm | 1956.1 - 1970.12 | 180 | 100 |
| 23 | OZONELA | Ozone concentraion, downtown L.A | logarithm | 1955.1 - 1972.12 | 216 | 100 |
| 24 | TPMON | Monthly temperatures in England | 12th difference | 1723.1-1970.12 | 2964 | 2000 |
| 25 | SUMMER | Mean summer temperature in Munich | logarithm | 1781 - 1988 | 208 | 100 |
| 26 | GLOBTP | Changes in global temperature (annual) | first difference | 1880 - 1985 | 105 | 70 |
| 27 | LYNX | Annual n of Lynx trapped, MacKenzie River | logarithm base 10 | 1821 - 1934 | 114 | 70 |
| 28 | CHICKNY | Rept. n. of cases of Chickenpox, m.thly, NYC | 12th diff. of the log | 1931.1 - 1972.12 | 486 | 300 |
| 29 | MEASL | Rept. n. of cases of Measles, monthly, NYC | 12th diff. of the log | 1928.1 - 1972.12 | 522 | 300 |
| 30 | MUMPS | Rept. n. of cases of Mumps, monthly, NYC | 12th diff. of the log | 1928.1 - 1972.12 | 522 | 300 |

Table A.1: Data sets.

| | ICP | | CVP | | BRP | | SA | | AR |
|---|---|---|---|---|---|---|---|---|---|
| Series | Lags | $q$ | Lags | $q$ | Lags | $q$ | Lags | $q$ | AR Lags |
| 1 | 1,3,9,12 | 1 | 1–3,6–12 | 1 | 1–12 | 4 | 1,6 | 1 | 1,3,5 |
| 2 | 1–3,9,10 | 1 | 1–3,5,7–10 | 1 | 1–12 | 4 | 1–3,9 | 0 | 1–3 |
| 3 | 1,3 | 1 | 1,2,4,5,7,9–12 | 1 | 1–12 | 4 | 1,6 | 1 | 1–4 |
| 4 | 1–6,8,10,12 | 1 | 1,2,4–6,8,9,11,12 | 1 | 1–12 | 4 | 1,4,12 | 1 | 2–4,12 |
| 5 | 1,5–7,9,11,12 | 1 | 2-12 | 1 | 1–12 | 4 | 1,2,5,12 | 2 | 1 |
| 6 | 1,8,9 | 1 | 1–12 | 1 | 1–12 | 4 | 1,10 | 0 | 1,5,9 |
| 7 | 2–4,7,11,12 | 1 | 1–9,11,12 | 1 | 1–12 | 4 | 2,4,12 | 1 | 1–3,5,6 |
| 8 | 7,9–12 | 1 | 2–7,9–11 | 1 | 1–12 | 4 | 1 | 0 | 1,6,12,24 |
| 9 | 1,3,8–10 | 1 | 1–8,10–12 | 1 | 1–12 | 4 | 1 | 0 | 1–3,8,10,13 |
| 10 | 1,4,6–11 | 1 | 1–9,11,12 | 1 | 1–12 | 4 | 1 | 0 | 1 |
| 11 | 1–4 | 1 | 1–4,6,8–11 | 1 | 1–12 | 4 | 1 | 0 | 1–3 |
| 12 | 1–3,5–12 | 1 | 1–12 | 1 | 1–12 | 4 | 1,9,12 | 1 | 1,5,14 |
| 13 | 1,3,5–10 | 1 | 2,4–10,12 | 1 | 1–12 | 4 | 1,8 | 1 | 1,5,14 |
| 14 | 1,2,5,6 | 1 | 1–8,10–12 | 1 | 1–12 | 4 | 1,4,5 | 0 | 1,4,5 |
| 15 | 1,12 | 1 | 1–12 | 1 | 1–12 | 4 | 1,12 | 0 | 1–3,11-13 |
| 16 | 1,12 | 1 | 1–12 | 1 | 1–12 | 4 | 1,5,12 | 0 | 1,12,13 |
| 17 | 1,2,3 | 1 | 1–8,10–12 | 1 | 1–12 | 4 | 1,2,4 | 2 | 1–3 |
| 18 | 1,2,3 | 1 | 1–3,5–12 | 2 | 1–12 | 4 | 1,2 | 1 | 1–3 |
| 19 | 1,2,9 | 1 | 1–5,7–9,11 | 1 | 1–12 | 4 | 1,2,11 | 1 | 1,2,9 |
| 20 | 1,4,11 | 1 | 1,5,6,8,10,11 | 1 | 1–12 | 4 | 1,9,12 | 1 | 1,2,8,11 |
| 21 | 1,2,8 | 1 | 1,3–5,7,10–12 | 1 | 1–12 | 4 | 1,3,4,12 | 1 | 1,2,8 |
| 22 | 1,4,9,10,12 | 1 | 1–5,7–12 | 1 | 1–12 | 4 | 1,4,6,12 | 0 | 1,4,12 |
| 23 | 1,4,5,10,11 | 1 | 1,4–9,12 | 1 | 1–12 | 4 | 1,10,12 | 0 | 1,10,12 |
| 24 | 1,2,4 | 1 | 1–5,7–12 | 1 | 1–12 | 4 | 1,2,12 | 0 | 1,2,4,6,8,12,13,15,24,25 |
| 25 | 1–7,10–12 | 1 | 1–4,6,8–12 | 1 | 1–12 | 4 | 1,6 | 0 | 1 |
| 26 | 1–6,8–11 | 2 | 1–3,7–9,11,12 | 1 | 1–12 | 4 | 2 | 0 | 1–4 |
| 27 | 1–12 | 4 | 1,4,5,10,12 | 1 | 1–12 | 4 | 1,3 | 1 | 1,2,4,7 |
| 28 | 1,9 | 1 | 1–9,11,12 | 1 | 1–12 | 4 | 1,7,12 | 0 | 1,4,9,12,13,20,23–25 |
| 29 | 1,2,11 | 1 | 1–7,9,11 | 1 | 1–12 | 4 | 1,2,9,12 | 0 | 1,2,4,11–13,21,24,25 |
| 30 | 1,11 | 1 | 1,2,4–12 | 1 | 1–12 | 4 | 1,12 | 0 | 1,11–13,24,25 |

Table A.2: Specification results.

| Series | ICP | CVP | BRP | SA | AR |
|---|---|---|---|---|---|
| 1 | 0.859 | 0.849 | 0.855 | 0.883 | 0.871 |
| 2 | 0.682 | 0.679 | 0.681 | 0.693 | 0.698 |
| 3 | 0.695 | 0.702 | 0.609 | 0.693 | 0.692 |
| 4 | 0.908 | 0.918 | 0.911 | 0.917 | 0.923 |
| 5 | 0.874 | 0.936 | 0.904 | 0.912 | 0.919 |
| 6 | 0.589 | 0.582 | 0.415 | 0.714 | 0.587 |
| 7 | 0.795 | 0.801 | 0.809 | 0.854 | 0.835 |
| 8 | 0.951 | 0.946 | 0.747 | 0.990 | 0.940 |
| 9 | 0.924 | 0.912 | 0.931 | 0.985 | 0.931 |
| 10 | 0.879 | 0.922 | 0.944 | 0.954 | 0.954 |
| 11 | 0.906 | 0.896 | 0.918 | 0.938 | 0.924 |
| 12 | 0.975 | 0.970 | 0.988 | 0.986 | 0.988 |
| 13 | 0.982 | 0.979 | 0.997 | 0.995 | 0.986 |
| 14 | 0.663 | 0.651 | 0.591 | 0.656 | 0.607 |
| 15 | 0.723 | 0.701 | 0.700 | 0.721 | 0.642 |
| 16 | 0.684 | 0.676 | 0.678 | 0.686 | 0.663 |
| 17 | 0.124 | 0.124 | 0.118 | 0.141 | 0.124 |
| 18 | 0.106 | 0.105 | 0.098 | 0.100 | 0.106 |
| 19 | 0.367 | 0.363 | 0.321 | 0.424 | 0.377 |
| 20 | 0.419 | 0.399 | 0.301 | 0.400 | 0.423 |
| 21 | 0.362 | 0.337 | 0.154 | 0.374 | 0.389 |
| 22 | 0.532 | 0.524 | 0.489 | 0.539 | 0.538 |
| 23 | 0.575 | 0.570 | 0.469 | 0.593 | 0.593 |
| 24 | 0.962 | 0.840 | 0.821 | 0.843 | 0.791 |
| 25 | 0.874 | 0.882 | 0.998 | 0.959 | 0.972 |
| 26 | 0.745 | 0.832 | 0.995 | 0.964 | 0.873 |
| 27 | 0.050 | 0.326 | 0.180 | 0.332 | 0.351 |
| 28 | 0.678 | 0.590 | 0.573 | 0.600 | 0.537 |
| 29 | 0.250 | 0.248 | 0.215 | 0.245 | 0.207 |
| 30 | 0.333 | 0.329 | 0.322 | 0.332 | 0.286 |

Table A.3: $nRMSE$ of estimated models.

| Horizon | ICP | CVP | BRP | SA | AR |
|---|---|---|---|---|---|
| 1 | 2 | 4 | 7 | 6 | 11 |
| | 4 | 6 | 5 | 4 | 11 |
| 6 | 4 | 5 | 4 | 7 | 10 |
| | 5 | 4 | 6 | 5 | 10 |
| 12 | 5 | 3 | 5 | 8 | 9 |
| | 4 | 3 | 8 | 4 | 11 |

Table A.4: Forecasting results: number of series where each model is the best model according to the nRMSE (first line) and MAE (second line).

| Horizon | ICP | CVP | BRP | SA | AR |
|---|---|---|---|---|---|
| | 5 | 8 | 5 | 7 | 5 |
| 1 | 9 | 7 | 9 | 5 | 5 |
| | 6 | 7 | 6 | 3 | 8 |
| 6 | 4 | 11 | 10 | 7 | 7 |
| | 6 | 4 | 6 | 9 | 5 |
| 12 | 6 | 6 | 10 | 11 | 10 |

Table A.5: Forecasting results: number of series where each model is the best model according to the MAD (first line) and CR (second line).

| | ICP | CVP | BRP | SA | AR | sum of losses | sum of wins - sum of losses |
|---|---|---|---|---|---|---|---|
| ICP | – | 6 | 11 | 8 | 11 | 36 | -22 |
| CVP | 2 | – | 8 | 7 | 8 | 25 | -4 |
| BRP | 4 | 6 | – | 5 | 8 | 23 | 1 |
| SA | 4 | 5 | 3 | – | 7 | 19 | 2 |
| AR | 2 | 4 | 2 | 2 | – | 10 | 24 |
| sum of wins | 12 | 21 | 24 | 22 | 34 | – | – |

Table A.6: Forecasting results: number of series where model A (column) is better than model B (line) according to the modified Diebold-Mariano statistic at a 0.10 level for 1-step-ahead (nRMSE test).

| | ICP | CVP | BRP | SA | AR | sum of losses | sum of wins - sum of losses |
|---|---|---|---|---|---|---|---|
| ICP | – | 7 | 10 | 8 | 9 | 34 | -20 |
| CVP | 4 | – | 8 | 6 | 7 | 25 | -1 |
| BRP | 2 | 6 | – | 3 | 5 | 16 | 16 |
| SA | 7 | 7 | 8 | – | 8 | 30 | -10 |
| AR | 1 | 4 | 6 | 3 | – | 14 | 15 |
| sum of wins | 14 | 24 | 32 | 20 | 29 | – | – |

Table A.7: Forecasting results: number of series where model A (column) is better than model B (line) according to the modified Diebold-Mariano statistic at a 0.10 level for 6-steps-ahead (nRMSE test).

| | ICP | CVP | BRP | SA | AR | sum of losses | sum of wins - sum of losses |
|---|---|---|---|---|---|---|---|
| ICP | – | 6 | 8 | 6 | 9 | 29 | -2 |
| CVP | 8 | – | 10 | 8 | 9 | 35 | -14 |
| BRP | 7 | 5 | – | 6 | 10 | 28 | -1 |
| SA | 9 | 7 | 6 | – | 10 | 32 | -5 |
| AR | 3 | 3 | 3 | 4 | – | 13 | 15 |
| sum of wins | 27 | 21 | 27 | 24 | 38 | – | – |

Table A.8: Forecasting results: number of series where model A (column) is better than model B (line) according to the modified Diebold-Mariano statistic at a 0.10 level for 12-steps-ahead (nRMSE test).

| | ICP | CVP | BRP | SA | AR | sum of losses | sum of wins - sum of losses |
|---|---|---|---|---|---|---|---|
| ICP | – | 6 | 9 | 7 | 10 | 32 | -18 |
| CVP | 2 | – | 7 | 5 | 8 | 20 | 1 |
| BRP | 4 | 7 | – | 7 | 10 | 28 | -7 |
| SA | 4 | 3 | 2 | – | 7 | 16 | 6 |
| AR | 4 | 5 | 3 | 3 | – | 15 | 20 |
| sum of wins | 14 | 21 | 21 | 22 | 35 | – | – |

Table A.9: Forecasting results: number of series where model A (column) is better than model B (line) according to the modified Diebold-Mariano statistic at a 0.10 level for 1-step-ahead (MAE test).

| | ICP | CVP | BRP | SA | AR | sum of losses | sum of wins - sum of losses |
|---|---|---|---|---|---|---|---|
| ICP | – | 6 | 9 | 9 | 10 | 34 | -16 |
| CVP | 3 | – | 8 | 7 | 7 | 25 | -4 |
| BRP | 6 | 7 | – | 4 | 7 | 24 | 5 |
| SA | 7 | 5 | 8 | – | 8 | 28 | -6 |
| AR | 2 | 3 | 5 | 2 | – | 12 | 20 |
| sum of wins | 18 | 21 | 29 | 22 | 32 | – | – |

Table A.10: Forecasting results: number of series where model A (column) is better than model B (line) according to the modified Diebold-Mariano statistic at a 0.10 level for 6-steps-ahead (MAE test).

| | ICP | CVP | BRP | SA | AR | sum of losses | sum of wins - sum of losses |
|---|---|---|---|---|---|---|---|
| ICP | – | 6 | 8 | 6 | 10 | 30 | -7 |
| CVP | 8 | – | 8 | 7 | 7 | 30 | -14 |
| BRP | 7 | 4 | – | 6 | 8 | 25 | 1 |
| SA | 6 | 4 | 6 | – | 7 | 23 | 1 |
| AR | 2 | 2 | 4 | 5 | – | 13 | 19 |
| sum of wins | 23 | 16 | 26 | 24 | 32 | – | – |

Table A.11: Forecasting results: number of series where model A (column) is better than model B (line) according to the modified Diebold-Mariano statistic at a 0.10 level for 12-steps-ahead (MAE test).

Table A.12: Forecasting results (nRMSE and MAE): 1 (first line), 6 (second line) and 12 (third line) steps ahead.

| | ICP | | CVP | | BRP | | SA | | AR | |
|---|---|---|---|---|---|---|---|---|---|---|
| Series | nRMSE | MAE | nRMSE | MAE | nRMSE | MAE | nRMSE | MAE | nRMSE | MAE |
| | 0.919 | 0.003 | 0.893 | 0.003 | 0.848 | 0.003 | 0.878 | 0.003 | 0.834 | 0.003 |
| 1 | 1.161 | 0.004 | 1.119 | 0.004 | 1.027 | 0.004 | 1.124 | 0.004 | 1.057 | 0.004 |
| | 1.374 | 0.005 | 1.397 | 0.005 | 1.277 | 0.005 | 1.290 | 0.005 | 1.226 | 0.005 |
| | 0.686 | 0.002 | 0.686 | 0.002 | 0.676 | 0.002 | 0.640 | 0.002 | 0.666 | 0.002 |
| 2 | 1.254 | 0.004 | 1.274 | 0.004 | 1.137 | 0.003 | 0.973 | 0.003 | 1.136 | 0.003 |
| | 1.531 | 0.004 | 1.589 | 0.005 | 1.397 | 0.004 | 1.124 | 0.003 | 1.281 | 0.004 |
| | 0.726 | 0.002 | 0.750 | 0.002 | 1.337 | 0.003 | 0.722 | 0.002 | 0.711 | 0.002 |
| 3 | 1.212 | 0.003 | 1.265 | 0.003 | 1.516 | 0.004 | 1.108 | 0.003 | 1.061 | 0.003 |
| | 1.448 | 0.004 | 1.554 | 0.004 | 1.824 | 0.005 | 1.311 | 0.003 | 1.330 | 0.003 |
| | 0.792 | 0.020 | 0.797 | 0.019 | 0.777 | 0.019 | 0.789 | 0.019 | 0.775 | 0.019 |
| 4 | 0.799 | 0.020 | 0.779 | 0.019 | 0.788 | 0.019 | 0.774 | 0.019 | 0.776 | 0.019 |
| | 0.813 | 0.020 | 0.797 | 0.020 | 0.799 | 0.020 | 0.791 | 0.020 | 0.795 | 0.020 |
| | 0.583 | 0.004 | 0.602 | 0.004 | 0.559 | 0.004 | 0.513 | 0.004 | 0.513 | 0.004 |
| 5 | 0.552 | 0.004 | 0.567 | 0.004 | 0.563 | 0.004 | 0.515 | 0.004 | 0.505 | 0.004 |
| | 0.551 | 0.004 | 0.563 | 0.004 | 0.549 | 0.004 | 0.522 | 0.004 | 0.494 | 0.004 |
| | 0.780 | 0.153 | 0.755 | 0.144 | 1.056 | 0.187 | 0.863 | 0.171 | 0.781 | 0.153 |
| 6 | 0.755 | 0.147 | 0.747 | 0.141 | 0.841 | 0.164 | 1.012 | 0.197 | 0.754 | 0.144 |
| | 0.849 | 0.166 | 0.854 | 0.166 | 1.102 | 0.199 | 1.154 | 0.228 | 0.851 | 0.166 |
| | 0.788 | 0.003 | 0.710 | 0.003 | 0.709 | 0.003 | 1.065 | 0.005 | 0.748 | 0.003 |
| 7 | 0.745 | 0.003 | 0.715 | 0.003 | 0.723 | 0.003 | 0.758 | 0.003 | 0.749 | 0.003 |
| | 0.798 | 0.003 | 0.804 | 0.003 | 0.761 | 0.003 | 0.735 | 0.003 | 0.707 | 0.003 |
| | 0.583 | 0.006 | 0.674 | 0.007 | 0.627 | 0.007 | 0.623 | 0.006 | 0.585 | 0.006 |
| 8 | 0.601 | 0.006 | 0.616 | 0.007 | 0.618 | 0.006 | 0.587 | 0.006 | 0.561 | 0.006 |
| | 0.602 | 0.006 | 0.613 | 0.007 | 0.693 | 0.007 | 0.590 | 0.006 | 0.551 | 0.006 |
| | 1.109 | 0.009 | 1.079 | 0.009 | 1.006 | 0.008 | 1.045 | 0.008 | 1.008 | 0.008 |
| 9 | 1.037 | 0.009 | 1.016 | 0.008 | 1.031 | 0.008 | 1.101 | 0.009 | 0.990 | 0.008 |
| | 1.158 | 0.009 | 1.164 | 0.009 | 1.151 | 0.009 | 1.182 | 0.009 | 1.108 | 0.009 |
| | 1.013 | 0.023 | 0.916 | 0.021 | 0.904 | 0.008 | 0.924 | 0.021 | 0.922 | 0.021 |
| 10 | 0.882 | 0.021 | 0.866 | 0.020 | 0.856 | 0.008 | 0.895 | 0.020 | 0.893 | 0.020 |
| | 0.849 | 0.019 | 0.857 | 0.020 | 0.840 | 0.009 | 0.868 | 0.019 | 0.866 | 0.194 |

continued on next page

continued from previous page

| | ICP | | CVP | | BRP | | SA | | AR | |
|---|---|---|---|---|---|---|---|---|---|---|
| Series | nRMSE | MAE | nRMSE | MAE | nRMSE | MAE | nRMSE | MAE | nRMSE | MAE |
| | 1.040 | 0.022 | 1.040 | 0.023 | 0.929 | 0.020 | 0.957 | 0.020 | 0.938 | 0.021 |
| 11 | 1.021 | 0.022 | 0.998 | 0.022 | 0.958 | 0.021 | 0.990 | 0.021 | 0.987 | 0.021 |
| | 1.030 | 0.023 | 1.034 | 0.023 | 1.013 | 0.022 | 1.021 | 0.022 | 1.022 | 0.022 |
| | 0.953 | 0.030 | 0.969 | 0.031 | 0.939 | 0.030 | 0.927 | 0.029 | 0.939 | 0.030 |
| 12 | 0.934 | 0.030 | 0.945 | 0.031 | 0.936 | 0.030 | 0.928 | 0.029 | 0.932 | 0.030 |
| | 0.887 | 0.028 | 0.917 | 0.030 | 0.897 | 0.029 | 0.893 | 0.029 | 0.905 | 0.030 |
| | 0.987 | 0.031 | 0.981 | 0.032 | 0.964 | 0.030 | 0.974 | 0.030 | 0.978 | 0.031 |
| 13 | 0.979 | 0.031 | 0.968 | 0.031 | 0.952 | 0.030 | 0.966 | 0.030 | 0.965 | 0.031 |
| | 0.982 | 0.032 | 0.935 | 0.030 | 0.936 | 0.030 | 0.946 | 0.030 | 0.962 | 0.031 |
| | 0.655 | 0.974 | 0.647 | 1.000 | 0.668 | 0.975 | 0.671 | 0.987 | 0.659 | 0.971 |
| 14 | 1.020 | 2.094 | 0.980 | 2.066 | 0.981 | 1.960 | 0.989 | 2.010 | 0.990 | 2.098 |
| | 1.031 | 2.170 | 0.988 | 2.100 | 0.996 | 1.922 | 1.017 | 2.192 | 1.019 | 2.193 |
| | 0.969 | 0.373 | 1.042 | 0.395 | 1.005 | 0.390 | 0.971 | 0.371 | 0.943 | 0.356 |
| 15 | 1.102 | 0.444 | 1.162 | 0.476 | 1.121 | 0.455 | 1.056 | 0.435 | 1.099 | 0.427 |
| | 1.113 | 0.446 | 1.182 | 0.460 | 1.133 | 0.447 | 1.071 | 0.446 | 1.111 | 0.436 |
| | 0.553 | 0.081 | 0.551 | 0.081 | 0.544 | 0.081 | 0.549 | 0.082 | 0.507 | 0.073 |
| 16 | 0.690 | 0.116 | 0.688 | 0.117 | 0.699 | 0.119 | 0.683 | 0.116 | 0.651 | 0.107 |
| | 0.704 | 0.121 | 0.719 | 0.125 | 0.742 | 0.129 | 0.709 | 0.124 | 0.661 | 0.111 |
| | 0.135 | 0.082 | 0.135 | 0.082 | 0.132 | 0.084 | 0.138 | 0.084 | 0.137 | 0.083 |
| 17 | 0.490 | 0.352 | 0.493 | 0.345 | 0.539 | 0.376 | 0.576 | 0.526 | 0.507 | 0.356 |
| | 0.606 | 0.487 | 0.610 | 0.474 | 1.060 | 1.078 | 0.744 | 0.752 | 0.619 | 0.478 |
| | 0.099 | 0.050 | 0.099 | 0.050 | 0.097 | 0.053 | 0.101 | 0.050 | 0.099 | 0.052 |
| 18 | 0.328 | 0.227 | 0.324 | 0.228 | 0.338 | 0.270 | 0.334 | 0.242 | 0.328 | 0.238 |
| | 0.450 | 0.361 | 0.462 | 0.386 | 0.449 | 0.359 | 0.448 | 0.421 | 0.444 | 0.361 |
| | 0.423 | 1.856 | 0.417 | 1.766 | 0.492 | 2.028 | 0.479 | 2.009 | 0.407 | 1.755 |
| 19 | 0.775 | 3.215 | 0.739 | 3.029 | 0.980 | 3.746 | 1.162 | 4.209 | 0.690 | 2.803 |
| | 0.890 | 3.817 | 0.846 | 3.627 | 1.148 | 4.780 | 1.147 | 4.893 | 0.834 | 3.560 |
| | 0.525 | 0.639 | 0.537 | 0.669 | 0.458 | 0.536 | 0.411 | 0.452 | 0.599 | 0.761 |
| 20 | | | | | | | | | | |

continued on next page

continued from previous page

| | ICP | | CVP | | BRP | | SA | | AR | |
|---|---|---|---|---|---|---|---|---|---|---|
| Series | nRMSE | MAE | nRMSE | MAE | nRMSE | MAE | nRMSE | MAE | nRMSE | MAE |
| | 1.787 | 2.289 | 2.641 | 2.499 | 1.439 | 1.880 | 0.799 | 0.875 | 1.536 | 2.003 |
| | 0.532 | 0.573 | 2.970 | 3.651 | 0.944 | 1.103 | 0.803 | 0.839 | 0.913 | 1.122 |
| | 0.270 | 0.236 | 0.268 | 0.232 | 0.420 | 0.375 | 0.373 | 0.347 | 0.357 | 0.335 |
| 21 | 1.101 | 1.106 | 1.166 | 1.158 | 1.437 | 1.350 | 1.663 | 1.712 | 1.298 | 1.323 |
| | 1.043 | 1.053 | 1.107 | 1.113 | 0.612 | 0.543 | 1.157 | 1.165 | 0.944 | 0.938 |
| | 0.624 | 0.193 | 0.629 | 0.198 | 0.573 | 0.180 | 0.556 | 0.183 | 0.610 | 0.188 |
| 22 | 0.655 | 0.190 | 0.647 | 0.187 | 0.620 | 0.181 | 0.584 | 0.188 | 0.650 | 0.196 |
| | 0.553 | 0.171 | 0.545 | 0.168 | 0.511 | 0.155 | 0.486 | 0.160 | 0.753 | 0.180 |
| | 0.657 | 0.186 | 0.691 | 0.202 | 0.623 | 0.018 | 0.606 | 0.178 | 0.606 | 0.178 |
| 23 | 0.656 | 0.189 | 0.738 | 0.211 | 0.674 | 0.189 | 0.649 | 0.185 | 0.649 | 0.185 |
| | 0.653 | 0.190 | 0.736 | 0.207 | 0.637 | 0.179 | 0.664 | 0.192 | 0.664 | 0.192 |
| | 0.968 | 1.476 | 0.815 | 1.247 | 0.827 | 1.260 | 0.815 | 1.240 | 0.778 | 1.195 |
| 24 | 0.997 | 1.510 | 0.828 | 1.262 | 0.836 | 1.274 | 0.830 | 1.262 | 0.795 | 1.211 |
| | 1.006 | 1.524 | 0.828 | 1.263 | 0.830 | 1.266 | 0.830 | 1.263 | 0.794 | 1.210 |
| | 0.848 | 0.039 | 0.856 | 0.037 | 0.928 | 0.043 | 0.855 | 0.038 | 0.906 | 0.040 |
| 25 | 0.846 | 0.038 | 0.862 | 0.039 | 0.887 | 0.041 | 0.850 | 0.038 | 0.902 | 0.041 |
| | 0.870 | 0.040 | 0.874 | 0.040 | 0.898 | 0.041 | 0.880 | 0.040 | 0.913 | 0.041 |
| | 1.059 | 0.106 | 1.083 | 0.119 | 1.109 | 0.116 | 1.080 | 0.109 | 0.990 | 0.099 |
| 26 | 1.022 | 0.104 | 1.142 | 0.117 | 1.107 | 0.114 | 1.106 | 0.115 | 1.077 | 0.113 |
| | 1.532 | 0.158 | 1.368 | 0.144 | 1.339 | 0.141 | 1.328 | 0.141 | 1.333 | 0.141 |
| | 17.404 | 4.735 | 0.507 | 0.224 | 0.708 | 0.327 | 0.595 | 0.262 | 0.532 | 0.246 |
| 27 | 6.662 | 1.407 | 0.649 | 0.305 | 1.183 | 0.543 | 1.024 | 0.477 | 0.820 | 0.406 |
| | 9.621 | 2.302 | 0.746 | 0.332 | 0.955 | 0.444 | 0.886 | 0.415 | 0.758 | 0.342 |
| | 0.697 | 0.221 | 0.612 | 0.196 | 0.611 | 0.196 | 0.613 | 0.195 | 0.586 | 0.188 |
| 28 | 0.911 | 0.302 | 0.808 | 0.265 | 0.770 | 0.255 | 0.778 | 0.253 | 0.769 | 0.251 |
| | 0.886 | 0.294 | 0.768 | 0.249 | 0.759 | 0.246 | 0.747 | 0.242 | 0.759 | 0.243 |
| | 0.239 | 0.344 | 0.240 | 0.341 | 0.229 | 0.334 | 0.235 | 0.340 | 0.213 | 0.307 |
| 29 | 0.603 | 0.885 | 0.605 | 0.897 | 0.558 | 0.856 | 0.590 | 0.892 | 0.543 | 0.824 |
| | 0.667 | 0.982 | 0.652 | 0.970 | 0.582 | 0.887 | 0.634 | 0.967 | 0.622 | 0.933 |

continued on next page

continued from previous page

| | ICP | | CVP | | BRP | | SA | | AR | |
|---|---|---|---|---|---|---|---|---|---|---|
| Series | nRMSE | MAE | nRMSE | MAE | nRMSE | MAE | nRMSE | MAE | nRMSE | MAE |
| | 0.331 | 0.162 | 0.323 | 0.160 | 0.321 | 0.160 | 0.296 | 0.158 | 0.308 | 0.155 |
| 30 | 0.508 | 0.285 | 0.617 | 0.314 | 0.588 | 0.298 | 0.560 | 0.303 | 0.569 | 0.290 |
| | 0.638 | 0.331 | 0.716 | 0.371 | 0.700 | 0.364 | 0.637 | 0.350 | 0.673 | 0.347 |

Table A.13: Forecasting results (MAD and CR): 1 (first line), 6 (second line) and 12 (third line) steps ahead.

| | ICP | | CVP | | BRP | | SA | | AR | |
|---|---|---|---|---|---|---|---|---|---|---|
| Series | MAD | CR | MAD | CR | MAD | CR | MAD | CR | MAD | CR |
| | 0.003 | 0.382 | 0.003 | 0.346 | 0.003 | 0.309 | 0.003 | 0.382 | 0.003 | 0.355 |
| 1 | 0.004 | 0.318 | 0.004 | 0.282 | 0.003 | 0.336 | 0.004 | 0.336 | 0.003 | 0.355 |
| | 0.005 | 0.391 | 0.005 | 0.409 | 0.005 | 0.382 | 0.005 | 0.400 | 0.004 | 0.391 |
| | 0.001 | 0.404 | 0.001 | 0.464 | 0.001 | 0.446 | 0.001 | 0.409 | 0.001 | 0.436 |
| 2 | 0.002 | 0.436 | 0.002 | 0.446 | 0.002 | 0.418 | 0.002 | 0.391 | 0.002 | 0.427 |
| | 0.002 | 0.446 | 0.002 | 0.446 | 0.002 | 0.418 | 0.002 | 0.400 | 0.002 | 0.409 |
| | 0.002 | 0.382 | 0.001 | 0.373 | 0.002 | 0.446 | 0.001 | 0.409 | 0.001 | 0.355 |
| 3 | 0.002 | 0.373 | 0.002 | 0.391 | 0.003 | 0.446 | 0.002 | 0.391 | 0.002 | 0.346 |
| | 0.003 | 0.391 | 0.003 | 0.391 | 0.003 | 0.381 | 0.003 | 0.400 | 0.003 | 0.382 |
| | 0.015 | 0.300 | 0.015 | 0.318 | 0.014 | 0.291 | 0.014 | 0.300 | 0.015 | 0.291 |
| 4 | 0.014 | 0.418 | 0.014 | 0.409 | 0.015 | 0.418 | 0.014 | 0.373 | 0.015 | 0.391 |
| | 0.016 | 0.364 | 0.015 | 0.364 | 0.015 | 0.364 | 0.015 | 0.346 | 0.015 | 0.373 |
| | 0.003 | 0.327 | 0.004 | 0.363 | 0.003 | 0.309 | 0.003 | 0.327 | 0.003 | 0.355 |
| 5 | 0.003 | 0.200 | 0.003 | 0.246 | 0.003 | 0.191 | 0.003 | 0.282 | 0.003 | 0.200 |
| | 0.003 | 0.246 | 0.003 | 0.273 | 0.003 | 0.218 | 0.003 | 0.264 | 0.003 | 0.218 |
| | 0.111 | 0.291 | 0.108 | 0.273 | 0.115 | 0.355 | 0.106 | 0.345 | 0.115 | 0.346 |
| 6 | 0.112 | 0.346 | 0.108 | 0.346 | 0.123 | 0.400 | 0.111 | 0.361 | 0.107 | 0.355 |
| | 0.131 | 0.300 | 0.120 | 0.309 | 0.128 | 0.409 | 0.122 | 0.319 | 0.119 | 0.309 |

continued on next page

continued from previous page

| | ICP | | CVP | | BRP | | SA | | AR | |
|---|---|---|---|---|---|---|---|---|---|---|
| Series | MAD | CR | MAD | CR | MAD | CR | MAD | CR | MAD | CR |
| 7 | 0.002 | 0.336 | 0.002 | 0.309 | 0.002 | 0.327 | 0.002 | 0.455 | 0.002 | 0.373 |
| | 0.002 | 0.300 | 0.002 | 0.263 | 0.002 | 0.309 | 0.002 | 0.300 | 0.002 | 0.318 |
| | 0.002 | 0.364 | 0.002 | 0.382 | 0.002 | 0.355 | 0.002 | 0.364 | 0.002 | 0.336 |
| 8 | 0.006 | 0.192 | 0.006 | 0.256 | 0.006 | 0.308 | 0.005 | 0.218 | 0.006 | 0.218 |
| | 0.006 | 0.333 | 0.006 | 0.321 | 0.006 | 0.269 | 0.006 | 0.282 | 0.006 | 0.295 |
| | 0.006 | 0.284 | 0.005 | 0.333 | 0.006 | 0.372 | 0.006 | 0.321 | 0.005 | 0.282 |
| 9 | 0.008 | 0.250 | 0.007 | 0.225 | 0.007 | 0.200 | 0.006 | 0.213 | 0.007 | 0.263 |
| | 0.008 | 0.263 | 0.008 | 0.250 | 0.007 | 0.263 | 0.007 | 0.275 | 0.008 | 0.238 |
| | 0.007 | 0.213 | 0.008 | 0.213 | 0.008 | 0.225 | 0.008 | 0.225 | 0.008 | 0.213 |
| 10 | 0.019 | 0.400 | 0.018 | 0.325 | 0.016 | 0.375 | 0.006 | 0.350 | 0.017 | 0.350 |
| | 0.016 | 0.200 | 0.016 | 0.188 | 0.018 | 0.175 | 0.007 | 0.188 | 0.018 | 0.188 |
| | 0.015 | 0.200 | 0.016 | 0.213 | 0.016 | 0.225 | 0.008 | 0.250 | 0.016 | 0.250 |
| 11 | 0.018 | 0.363 | 0.018 | 0.325 | 0.017 | 0.300 | 0.016 | 0.350 | 0.016 | 0.363 |
| | 0.017 | 0.250 | 0.017 | 0.238 | 0.015 | 0.275 | 0.016 | 0.238 | 0.016 | 0.250 |
| | 0.016 | 0.313 | 0.018 | 0.300 | 0.016 | 0.300 | 0.017 | 0.263 | 0.017 | 0.288 |
| 12 | 0.022 | 0.229 | 0.025 | 0.267 | 0.022 | 0.286 | 0.023 | 0.257 | 0.023 | 0.276 |
| | 0.023 | 0.210 | 0.022 | 0.229 | 0.021 | 0.238 | 0.020 | 0.248 | 0.021 | 0.248 |
| | 0.022 | 0.295 | 0.022 | 0.295 | 0.022 | 0.295 | 0.021 | 0.286 | 0.023 | 0.295 |
| 13 | 0.017 | 0.238 | 0.022 | 0.219 | 0.023 | 0.210 | 0.023 | 0.210 | 0.023 | 0.219 |
| | 0.023 | 0.295 | 0.022 | 0.276 | 0.022 | 0.276 | 0.022 | 0.286 | 0.021 | 0.295 |
| | 0.022 | 0.267 | 0.022 | 0.248 | 0.022 | 0.248 | 0.022 | 0.248 | 0.023 | 0.267 |
| 14 | 0.458 | 0.503 | 0.318 | 0.554 | 0.448 | 0.547 | 0.412 | 0.506 | 0.416 | 0.509 |
| | 1.839 | 0.509 | 1.852 | 0.491 | 1.603 | 0.503 | 1.918 | 0.488 | 1.844 | 0.503 |
| | 2.391 | 0.516 | 2.283 | 0.503 | 1.794 | 0.516 | 2.077 | 0.494 | 2.605 | 0.503 |
| 15 | 0.205 | 0.254 | 0.214 | 0.324 | 0.223 | 0.282 | 0.219 | 0.282 | 0.196 | 0.268 |
| | 0.282 | 0.225 | 0.292 | 0.282 | 0.257 | 0.254 | 0.294 | 0.197 | 0.216 | 0.211 |
| | 0.260 | 0.183 | 0.219 | 0.197 | 0.250 | 0.169 | 0.304 | 0.183 | 0.229 | 0.225 |
| 16 | 0.048 | 0.301 | 0.044 | 0.308 | 0.045 | 0.293 | 0.047 | 0.293 | 0.042 | 0.256 |
| | 0.090 | 0.203 | 0.093 | 0.196 | 0.096 | 0.195 | 0.090 | 0.173 | 0.076 | 0.158 |
| | 0.090 | 0.203 | 0.092 | 0.218 | 0.094 | 0.248 | 0.092 | 0.196 | 0.084 | 0.165 |

continued on next page

continued from previous page

| Series | ICP | | CVP | | BRP | | SA | | AR | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAD | CR | MAD | CR | MAD | CR | MAD | CR | MAD | CR |
| | 0.040 | 0.378 | 0.039 | 0.387 | 0.041 | 0.389 | 0.041 | 0.384 | 0.041 | 0.402 |
| 17 | 0.178 | 0.447 | 0.180 | 0.420 | 0.191 | 0.442 | 0.185 | 0.533 | 0.188 | 0.427 |
| | 0.273 | 0.453 | 0.252 | 0.429 | 0.266 | 0.560 | 0.236 | 0.538 | 0.278 | 0.424 |
| | 0.016 | 0.360 | 0.019 | 0.404 | 0.020 | 0.424 | 0.015 | 0.362 | 0.019 | 0.409 |
| 18 | 0.109 | 0.398 | 0.119 | 0.409 | 0.113 | 0.447 | 0.123 | 0.422 | 0.122 | 0.462 |
| | 0.225 | 0.420 | 0.259 | 0.418 | 0.218 | 0.398 | 0.277 | 0.398 | 0.255 | 0.416 |
| | 1.635 | 0.080 | 1.526 | 0.090 | 1.788 | 0.125 | 1.939 | 0.182 | 1.436 | 0.136 |
| 19 | 1.889 | 1.890 | 1.890 | 0.068 | 2.628 | 0.091 | 3.037 | 0.125 | 2.019 | 0.068 |
| | 3.142 | 2.985 | 2.985 | 0.250 | 4.054 | 0.352 | 4.180 | 0.398 | 3.075 | 0.261 |
| | 0.261 | 0.523 | 0.229 | 0.533 | 0.260 | 0.551 | 0.317 | 0.411 | 0.250 | 0.561 |
| 20 | 0.005 | 0.439 | 1.235 | 0.402 | 0.454 | 0.430 | 0.694 | 0.336 | 0.589 | 0.439 |
| | 0.352 | 0.280 | 1.409 | 0.486 | 0.539 | 0.421 | 0.533 | 0.327 | 0.364 | 0.430 |
| | 0.155 | 0.411 | 0.149 | 0.398 | 0.207 | 0.407 | 0.176 | 0.463 | 0.165 | 0.426 |
| 21 | 0.306 | 0.435 | 0.339 | 0.444 | 0.617 | 0.398 | 0.402 | 0.472 | 0.385 | 0.435 |
| | 0.293 | 0.509 | 0.292 | 0.537 | 0.295 | 0.370 | 0.308 | 0.519 | 0.294 | 0.482 |
| | 0.192 | 0.261 | 0.186 | 0.290 | 0.157 | 0.290 | 0.163 | 0.304 | 0.170 | 0.290 |
| 22 | 0.186 | 0.044 | 0.159 | 0.044 | 0.169 | 0.044 | 0.159 | 0.058 | 0.175 | 0.058 |
| | 0.177 | 0.362 | 0.166 | 0.333 | 0.163 | 0.333 | 0.149 | 0.348 | 0.189 | 0.377 |
| | 0.149 | 0.267 | 0.160 | 0.295 | 0.151 | 0.238 | 0.152 | 0.248 | 0.152 | 0.248 |
| 23 | 0.147 | 0.038 | 0.164 | 0.057 | 0.136 | 0.028 | 0.161 | 0.057 | 0.161 | 0.057 |
| | 0.141 | 0.429 | 0.185 | 0.467 | 0.143 | 0.419 | 0.177 | 0.438 | 0.177 | 0.438 |
| | 1.203 | 0.267 | 1.011 | 0.210 | 1.027 | 0.218 | 0.988 | 0.209 | 0.938 | 0.200 |
| 24 | 1.195 | 0.258 | 1.037 | 0.208 | 1.053 | 0.212 | 1.017 | 0.208 | 0.979 | 0.198 |
| | 1.200 | 0.172 | 1.029 | 0.155 | 1.041 | 0.156 | 1.017 | 0.158 | 0.980 | 0.155 |
| | 0.034 | 0.299 | 0.030 | 0.247 | 0.032 | 0.289 | 0.031 | 0.227 | 0.030 | 0.299 |
| 25 | 0.033 | 0.330 | 0.033 | 0.320 | 0.033 | 0.299 | 0.034 | 0.351 | 0.032 | 0.309 |
| | 0.036 | 0.289 | 0.035 | 0.289 | 0.033 | 0.258 | 0.033 | 0.299 | 0.033 | 0.247 |
| | 0.070 | 0.208 | 0.098 | 0.208 | 0.075 | 0.208 | 0.084 | 0.208 | 0.075 | 0.250 |
| 26 | | | | | | | | | | |

continued on next page

continued from previous page

| | ICP | | CVP | | BRP | | SA | | AR | |
|---|---|---|---|---|---|---|---|---|---|---|
| Series | MAD | CR | MAD | CR | MAD | CR | MAD | CR | MAD | CR |
| | 0.099 | 0.333 | 0.072 | 0.375 | 0.080 | 0.292 | 0.083 | 0.292 | 0.073 | 0.292 |
| | 0.137 | 0.417 | 0.117 | 0.250 | 0.165 | 0.208 | 0.105 | 0.208 | 0.107 | 0.208 |
| | 0.367 | 0.242 | 0.231 | 0.091 | 0.288 | 0.182 | 0.216 | 0.242 | 0.251 | 0.152 |
| 27 | 0.579 | 0.273 | 0.249 | 0.121 | 0.470 | 0.242 | 0.291 | 0.152 | 0.346 | 0.152 |
| | 0.618 | 0.424 | 0.296 | 0.091 | 0.245 | 0.091 | 0.304 | 0.061 | 0.258 | 0.121 |
| | 0.169 | 0.354 | 0.143 | 0.274 | 0.146 | 0.286 | 0.138 | 0.286 | 0.139 | 0.286 |
| 28 | 0.240 | 0.229 | 0.220 | 0.189 | 0.213 | 0.183 | 0.206 | 0.189 | 0.195 | 0.200 |
| | 0.229 | 0.131 | 0.196 | 0.109 | 0.195 | 0.103 | 0.189 | 0.097 | 0.190 | 0.126 |
| | 0.302 | 0.228 | 0.292 | 0.228 | 0.280 | 0.213 | 0.308 | 0.218 | 0.257 | 0.232 |
| 29 | 0.725 | 0.071 | 0.722 | 0.062 | 0.786 | 0.076 | 0.798 | 0.062 | 0.765 | 0.114 |
| | 0.847 | 0.119 | 0.821 | 0.109 | 0.807 | 0.085 | 0.863 | 0.114 | 0.840 | 0.095 |
| | 0.132 | 0.308 | 0.135 | 0.313 | 0.125 | 0.303 | 0.123 | 0.322 | 0.136 | 0.294 |
| 30 | 0.224 | 0.147 | 0.225 | 0.147 | 0.234 | 0.161 | 0.222 | 0.151 | 0.199 | 0.119 |
| | 0.273 | 0.128 | 0.263 | 0.147 | 0.294 | 0.147 | 0.260 | 0.142 | 0.265 | 0.137 |

Table A.14: forecasting results ($S^*$ for the $nRMSE$ test): 1 (first line), 6 (second line) and 12 (third line) steps ahead. Note: the apex 1 denotes rejection of the null hypothesis of equal accuracy in forecast performance at the 0.10 significance level, 2 at 0.05, 3 at 0.01. nc = not computed.

| Ser. | SA vs ICP | SA vs CVP | SA vs BRP | SA vs AR | ICP vs CVP | ICP vs BRP | ICP vs AR | CVP vs BRP | CVP vs AR | BRP vs AR |
|---|---|---|---|---|---|---|---|---|---|---|
| | -1.457 | -0.384 | $1.846^{1}$ | 1.629 | $2.572^{2}$ | $3.005^{3}$ | $2.253^{2}$ | $2.122^{2}$ | 1.493 | 0.261 |
| 1 | -1.316 | 0.390 | $2.309^{2}$ | $2.070^{2}$ | $1.910^{1}$ | $2.369^{2}$ | $1.890^{1}$ | 1.627 | 0.976 | $-1.790^{1}$ |
| | -1.417 | -1.307 | 0.470 | 1.185 | -0.699 | 1.076 | 1.333 | 1.020 | 1.265 | 1.596 |
| | $-3.100^{3}$ | $-2.972^{3}$ | $-2.542^{2}$ | $-4.086^{3}$ | -0.151 | $1.794^{1}$ | 1.442 | $2.046^{2}$ | 1.401 | 0.382 |
| 2 | $-4.268^{3}$ | $-4.396^{3}$ | $-3.651^{3}$ | $-3.909^{3}$ | $-6.615^{3}$ | $5.079^{3}$ | $4.875^{3}$ | $5.307^{3}$ | $5.081^{3}$ | $-3.255^{3}$ |
| | $-3.323^{3}$ | $-3.460^{3}$ | $-2.929^{3}$ | $-2.577^{2}$ | $-4.509^{3}$ | $4.183^{3}$ | $3.893^{3}$ | $4.281^{3}$ | $4.008^{3}$ | $3.522^{3}$ |
| | -0.333 | -0.727 | $-5.325^{3}$ | 0.654 | -0.627 | $-5.323^{3}$ | $2.155^{2}$ | $-5.286^{3}$ | $1.918^{1}$ | $5.410^{3}$ |
| 3 | $-3.357^{3}$ | $-3.315^{3}$ | $-3.645^{3}$ | $3.362^{3}$ | $-2.291^{2}$ | $-3.412^{3}$ | $3.539^{3}$ | $-3.499^{3}$ | $3.540^{3}$ | $3.773^{3}$ |

continued on next page

continued from previous page

| Ser. | SA vs ICP | SA vs CVP | SA vs BRP | SA vs AR | ICP vs CVP | ICP vs BRP | ICP vs AR | CVP vs BRP | CVP vs AR | BRP vs AR |
|---|---|---|---|---|---|---|---|---|---|---|
| | -2.104[2] | -2.179[2] | -2.221[2] | -1.244 | -2.315[2] | -2.250[2] | 2.064[2] | -2.224[2] | 2.169[2] | 2.208[2] |
| | 0.291 | 0.108 | 0.943 | 1.151 | -0.448 | 1.555 | 1.160 | 1.703[1] | 1.543 | 0.183 |
| 4 | -0.720 | 0.351 | -0.187 | 0.531 | 3.845[3] | 2.248[2] | 2.022[2] | -1.874[1] | 0.271 | 0.940 |
| | -0.323 | 0.157 | 0.105 | 0.596 | 2.561[2] | 1.965[2] | 0.666 | -0.479 | 0.107 | 0.189 |
| | -0.708 | -1.340 | 0.533 | 0.302 | -0.491 | 0.819 | 0.704 | 1.947[1] | 1.443 | -0.456 |
| 5 | 0.602 | 0.006 | 0.132 | 0.647 | -0.507 | -1.093 | -0.282 | 0.116 | 0.519 | 0.756 |
| | 1.211 | 2.400[2] | 0.568 | 0.998 | -0.793 | 0.078 | 0.683 | 0.344 | 0.742 | 0.597 |
| | 2.017[2] | 2.404[2] | -1.348 | 1.872[1] | 1.352 | -1.872[1] | -0.121 | -2.036[2] | -2.448[2] | 1.876[1] |
| 6 | 1.870[1] | 1.875[1] | 1.474 | 1.960[2] | 0.437 | nc | 0.117 | nc | -0.639 | nc |
| | 2.445[2] | 2.487[2] | 0.270 | 2.547[2] | -0.550 | -1.209 | -0.409 | -1.221 | 0.423 | 1.216 |
| | 4.861[3] | 5.363[3] | 5.914[3] | 5.246[3] | 1.275 | 1.507 | 0.673 | 0.067 | -1.799[1] | -2.905[3] |
| 7 | 0.744 | 3.157[3] | 2.799[3] | 0.358 | 1.596 | 1.222 | -0.361 | -0.873 | -1.846[1] | -2.138[2] |
| | -1.239 | -3.631[3] | -1.140 | 1.789[1] | -0.194 | 1.084 | 2.106[2] | 6.494[3] | 5.183[3] | 3.340[3] |
| | 1.213 | -1.882[2] | -0.445 | 1.025 | -2.259[2] | -1.263 | -0.274 | 1.460 | 2.046[2] | 1.058 |
| 8 | -0.614 | -2.180[2] | -1.020 | 0.535 | -0.528 | -0.416 | 0.678 | -0.026 | 1.423 | 1.096 |
| | -1.250 | -7.042[3] | -3.688[3] | 1.769[1] | -1.072 | -3.206[3] | 5.359[3] | -3.017[3] | 4.062[3] | 5.412[3] |
| | -1.539 | -0.890 | 0.729 | 0.452 | 0.350 | 1.838[1] | 1.747[1] | 1.319 | 1.200 | -0.170 |
| 9 | 1.881[1] | 1.641 | 1.676[1] | 1.581 | 1.132 | 0.366 | 1.204 | -1.098 | 0.993 | 1.223 |
| | 0.104 | -0.011 | 1.704[1] | 2.019[2] | -0.314 | 0.046 | 0.556 | 0.152 | 0.631 | 1.675[1] |
| | -2.653[3] | -0.386 | -0.053 | -0.829 | 2.583[3] | 2.727[3] | 2.652[3] | 0.430 | 0.385 | 0.051 |
| 10 | -1.313 | 1.710[1] | 1.690[1] | -1.213 | 2.024[2] | 2.990[3] | 1.305 | 2.709[3] | -1.711[1] | -1.692[1] |
| | 0.059 | -0.572 | 1.149 | -1.436 | -4.609[3] | 0.540 | -0.063 | 1.253 | 0.567 | -1.151 |
| | -2.001[2] | -2.144[2] | 0.748 | 0.349 | -0.400 | 2.835[3] | 2.620[3] | 3.019[3] | 2.765[3] | -0.566 |
| 11 | -1.889[1] | -1.229 | 1.470 | -0.634 | 1.953[1] | 2.993[3] | 2.150[2] | 2.855[3] | 1.335 | -2.318[2] |
| | -0.600 | -1.459 | -0.660 | -0.908 | -0.585 | 0.450 | 0.460 | 2.746[3] | 1.729[1] | 0.272 |
| | -1.245 | -1.723[1] | -1.186 | -1.176 | -0.527 | 0.857 | 0.589 | 1.379 | 1.097 | -0.220 |
| 12 | -0.527 | -0.796 | -0.786 | -0.584 | -0.897 | -0.320 | -0.014 | 0.704 | 0.663 | 0.252 |
| | 0.792 | -2.802[3] | -1.977[2] | -3.232[3] | -2.490[2] | -1.675[1] | -2.548[2] | 3.240[3] | 1.100 | -1.799[1] |
| | -0.980 | -0.360 | 0.892 | -0.695 | 0.192 | 1.464 | 0.381 | 0.595 | 0.053 | -1.222 |
| 13 | -8.783[3] | -0.276 | 2.681[3] | -0.399 | 0.462 | 6.912[3] | 0.880 | 0.705 | 0.041 | -1.199 |

continued on next page

continued from previous page

| Ser. | SA vs ICP | SA vs CVP | SA vs BRP | SA vs AR | ICP vs CVP | ICP vs BRP | ICP vs AR | CVP vs BRP | CVP vs AR | BRP vs AR |
|---|---|---|---|---|---|---|---|---|---|---|
| | -4.620[3] | 3.605[3] | 2.840[3] | -2.169[2] | 5.069[3] | 4.259[3] | 1.033 | -0.462 | -4.352[3] | -3.141[3] |
| | 0.258 | 1.101 | -0.477 | -1.106 | 0.643 | -0.529 | -0.265 | -0.946 | -1.111 | 0.473 |
| 14 | -0.770 | 0.512 | 0.238 | -0.941 | 1.071 | 1.477 | 0.768 | -0.019 | -0.537 | -0.249 |
| | -0.295 | 1.083 | 0.299 | -0.528 | 1.487 | 0.548 | 0.292 | -0.158 | -1.105 | -0.303 |
| | 0.448 | -0.856 | -0.387 | 0.797 | -1.678[1] | -0.878 | 0.468 | 1.181 | 1.769[1] | 1.272 |
| 15 | -2.721[3] | -0.999 | -0.638 | -0.713 | -0.727 | -0.284 | -0.029 | 2.210[2] | 0.672 | 0.252 |
| | -0.764 | -0.564 | -0.352 | -0.370 | -0.522 | -0.228 | -0.026 | 1.100 | 0.394 | 0.144 |
| | -0.307 | -0.234 | 0.949 | 2.173[2] | 0.292 | 0.770 | 1.998[2] | 0.915 | 2.134[2] | 1.971[2] |
| 16 | -0.763 | -0.952 | -0.871 | 1.748[1] | 0.305 | -0.400 | 1.597 | -0.646 | 1.695[1] | 1.655[1] |
| | 0.581 | -0.656 | -0.936 | 4.733[3] | -0.983 | -1.099 | 2.627[3] | -1.119 | 2.632[3] | 1.935[1] |
| | 1.091 | 0.784 | 0.942 | -0.431 | -1.473 | 0.597 | -1.417 | 0.721 | -1.212 | -1.064 |
| 17 | 3.569[3] | 2.646[3] | 0.873 | 2.525[2] | -0.366 | -1.493 | -1.130 | -1.327 | -0.827 | 1.138 |
| | 3.058[3] | 2.368[2] | -5.440[3] | 2.325[2] | -0.268 | -5.693[3] | -0.791 | -5.393[3] | -0.561 | 5.821[3] |
| | 0.608 | 0.789 | 1.695[1] | 0.571 | 0.919 | 1.077 | -0.231 | 0.857 | -0.849 | -1.073 |
| 18 | 0.559 | 1.036 | -0.143 | 0.727 | 1.048 | -0.571 | 0.032 | -0.919 | -0.883 | 0.592 |
| | 2.214[2] | 1.925[1] | 2.153[2] | 1.739[1] | -1.052 | 0.135 | 0.441 | 0.576 | 0.955 | 0.085 |
| | 1.756[1] | 1.953[1] | -0.307 | 2.316[2] | 1.111 | -2.095[2] | 1.549 | -2.366[2] | 0.672 | 2.511[2] |
| 19 | 2.930[3] | 3.031[3] | 1.617 | 3.107[3] | 3.102[3] | -1.613 | 3.927[3] | -1.889[1] | 2.987[3] | 2.057[2] |
| | 2.731[3] | 2.652[3] | 0.150 | 2.710[3] | 1.874[1] | -2.634[3] | 2.222[2] | -2.572[2] | 0.342 | 2.602[3] |
| | -3.089[3] | -3.264[3] | -1.177 | -5.053[3] | -0.908 | 3.777[3] | -5.480[3] | 3.798[3] | -7.211[3] | -6.925[3] |
| 20 | -4.475[3] | -3.366[3] | -5.093[3] | -4.869[3] | -1.541 | 2.416[2] | 3.525[3] | 2.012[2] | 2.201[2] | -1.186 |
| | 2.200[2] | -2.943[3] | -5.521[3] | -1.506 | -2.939[3] | -3.483[3] | -4.373[3] | 2.875[3] | 2.813[3] | 0.612 |
| | 7.554[3] | 8.268[3] | -2.110[2] | 1.722[1] | 0.720 | -5.669[3] | -8.973[3] | -5.582[3] | -7.986[3] | 2.956[3] |
| 21 | 5.718[3] | 5.482[3] | 3.251[3] | 5.818[3] | -6.804[3] | -2.396[2] | -4.974[3] | -2.161[2] | -4.307[3] | 1.238 |
| | 22.22[3] | nc | 4.403[3] | 6.764[3] | -10.293[3] | 3.832[3] | 5.132[3] | 4.292[3] | 7.260[3] | -3.555[3] |
| | -1.084 | -1.020 | 0.795 | -1.283 | -0.367 | 2.939[3] | 0.605 | 3.770[3] | 0.615 | -1.139 |
| 22 | -1.037 | -0.743 | 0.301 | -2.051[2] | 0.620 | 1.834[1] | 0.214 | 1.863[1] | -0.111 | -1.050 |
| | -2.902[3] | -1.665[1] | 0.589 | -4.045[3] | 0.935 | 3.132[3] | -2.930[3] | 4.724[3] | -2.223[2] | -3.005[3] |
| | 0.186 | -0.920 | 0.930 | nc | -0.921 | 1.093 | 1.402 | 2.050[2] | 2.489[2] | 0.661 |
| 23 | 1.877[1] | 2.036[2] | 2.070[2] | nc | -1.633 | -0.983 | 0.092 | 2.027[2] | 1.873[1] | 0.700 |

continued on next page

continued from previous page

| Ser. | SA vs ICP | SA vs CVP | SA vs BRP | SA vs AR | ICP vs CVP | ICP vs BRP | ICP vs AR | CVP vs BRP | CVP vs AR | BRP vs AR |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2.456[2] | 2.659[3] | 5.573[3] | nc | -1.675[1] | 0.347 | -0.304 | 5.886[3] | 1.141 | -0.458 |
| 24 | -8.720[3] | 0.425 | -2.290[2] | 3.748[3] | 8.804[3] | 8.146[3] | 9.662[3] | -2.811[3] | 3.790[3] | 4.658[3] |
| | -8.200[3] | 0.885 | -1.462 | 4.156[3] | 8.157[3] | 7.942[3] | 9.196[3] | -2.013[2] | 4.016[3] | 4.521[3] |
| | -9.287[3] | 0.672 | -0.158 | 4.397[3] | 8.865[3] | 9.221[3] | 9.263[3] | -0.683 | 4.306[3] | 4.245[3] |
| 25 | 0.312 | 0.139 | -1.816[1] | -1.626 | -0.203 | -1.638 | -1.108 | -1.497 | -0.956 | 1.003 |
| | 0.491 | -0.248 | -0.822 | -0.885 | -1.100 | -0.895 | -0.950 | -0.577 | -0.644 | -1.809[1] |
| | 1.703[1] | 1.586 | -0.597 | -0.680 | -0.313 | -0.968 | -1.016 | -1.186 | -1.226 | -1.570 |
| 26 | 0.451 | 0.037 | -0.417 | 1.418 | -0.375 | -0.761 | 0.570 | -0.245 | 0.827 | 1.175 |
| | nc | -0.283 | 1.325 | 1.524 | -1.599 | nc | nc | 0.362 | 0.665 | 1.647 |
| | nc | nc | -0.429 | nc | nc | nc | nc | nc | nc | -0.216 |
| 27 | -2.893[3] | 1.697[1] | -1.420 | 0.879 | 2.895[3] | 2.892[3] | 2.895[3] | -3.231[3] | -1.264 | 2.173[2] |
| | -1.092 | 1.669[1] | -0.872 | 1.316 | 1.107 | 1.078 | 1.102 | -3.830[3] | -1.854[1] | 2.057[2] |
| | -1.116 | 0.785 | -0.287 | 0.617 | 1.123 | 1.115 | 1.122 | -1.798[1] | -0.719 | 3.598[3] |
| 28 | -3.341[3] | 0.397 | 0.403 | 1.425 | 3.263[3] | 3.442[3] | 3.941[3] | 0.139 | 1.245 | 1.150 |
| | -2.617[3] | -2.177[2] | 0.708 | 0.497 | 1.999[2] | 2.721[3] | 2.692[3] | 2.279[2] | 1.405 | 0.021 |
| | -2.502[2] | -1.284 | -0.601 | -0.864 | 2.201[2] | 2.553[2] | 2.773[3] | 1.802[1] | 0.311 | -0.034 |
| 29 | -1.094 | -1.073 | 0.885 | 2.781[3] | -0.241 | 1.206 | 2.912[3] | 1.349 | 2.965[3] | 1.848[1] |
| | -0.715 | -1.152 | 0.907 | 1.438 | -0.315 | 0.986 | 1.391 | 1.040 | 1.519 | 0.315 |
| | -1.575 | -1.129 | 1.941[1] | 0.427 | 2.213[2] | 1.852[1] | 1.969[2] | 1.781[1] | 1.070 | -0.630 |
| 30 | -2.588[3] | -1.132 | -0.565 | 1.253 | 2.563[2] | 2.086[2] | 2.303[2] | 0.459 | 1.586 | 1.479 |
| | 2.026[2] | -2.899[3] | 0.609 | 1.121 | -2.405[2] | -1.324 | -0.074 | 1.114 | 1.628 | 0.596 |
| | 1.883[1] | -2.345[2] | -0.936 | 0.528 | -2.761[3] | -2.270[2] | -0.898 | 1.187 | 1.519 | 1.104 |

Table A.15: forecasting results ($S^*$ for the $MAE$): 1 (first line), 6 (second line) and 12 (third line) steps ahead. Note: the apex 1 denotes rejection of the null hypothesis of equal accuracy in forecast performance at the 0.10 significance level, 2 at 0.05, 3 at 0.01. nc = not computed.

| Ser. | SA vs ICP | SA vs CVP | SA vs BRP | SA vs AR | ICP vs CVP | ICP vs BRP | ICP vs AR | CVP vs BRP | CVP vs AR | BRP vs AR |
|---|---|---|---|---|---|---|---|---|---|---|
| | -0.924 | -0.437 | 1.778[1] | 1.416 | 1.289 | 2.776[3] | 1.958[1] | 2.624[3] | 1.562 | -0.390 |
| 1 | -0.457 | 0.762 | 2.097[2] | 1.790[1] | 1.407 | 1.980[2] | 1.332 | 1.117 | 0.414 | -1.689[1] |
| | -1.254 | -1.184 | 0.465 | 1.121 | -0.638 | 0.989 | 1.233 | 0.932 | 1.169 | 1.083 |
| | -4.270[3] | -4.084[3] | -3.223[3] | -3.650[3] | -0.513 | 1.866[1] | 2.297[2] | 2.486[2] | 2.321[2] | 1.050 |
| 2 | -6.003[3] | -6.121[3] | -4.569[3] | -4.587[3] | -6.854[3] | 8.212[3] | 8.943[3] | 8.349[3] | 8.748[3] | -1.556 |
| | -5.635[3] | -6.066[3] | -4.776[3] | -3.856[3] | -12.46[3] | 7.548[3] | 7.217[3] | 8.492[3] | 7.821[3] | 6.659[3] |
| | -0.452 | -0.715 | -6.519[3] | 0.871 | -0.318 | -6.415[3] | 2.823[3] | -6.479[3] | 1.922[1] | 6.737[3] |
| 3 | -3.357[3] | -3.592[3] | -5.053[3] | 2.972[3] | -3.466[3] | -4.813[3] | 3.693[3] | -4.672[3] | 3.947[3] | 5.102[3] |
| | -1.972[1] | -2.086[2] | -2.533[2] | -0.574 | -2.312[2] | -2.721[3] | 2.270[2] | -2.713[3] | 2.330[2] | 2.599[3] |
| | -0.190 | -0.043 | 0.534 | 0.310 | 0.330 | 1.580 | 0.806 | 1.190 | 0.557 | -0.551 |
| 4 | -1.307 | -0.044 | -0.664 | -0.094 | 3.668[3] | 2.530[2] | 2.006[2] | -2.832[3] | -0.051 | 0.895 |
| | -0.553 | -0.077 | -0.141 | 0.009 | 3.849[3] | 3.239[3] | 0.735 | -0.769 | 0.111 | 0.203 |
| | 0.217 | -1.712[1] | -0.353 | -0.917 | -1.708[1] | -0.362 | -0.380 | 1.934[1] | 1.606 | -0.051 |
| 5 | 1.330 | 0.280 | 0.931 | 1.252 | -1.225 | -0.503 | -1.231 | 0.873 | 1.114 | -0.249 |
| | 1.655[1] | 2.215[2] | 0.918 | 1.270 | -1.366 | 0.322 | 0.704 | 0.702 | 1.045 | 0.247 |
| | 2.116[2] | 3.108[3] | -0.832 | 2.012[2] | 2.752[3] | -1.861[1] | -0.065 | -2.418[2] | -3.281[3] | 1.874[1] |
| 6 | 1.972[2] | 2.273[2] | 1.601 | 2.335[2] | 1.217 | -1.976[2] | 0.625 | -2.576[3] | -0.892 | 2.631[3] |
| | 2.182[2] | 2.358[2] | 0.778 | 2.376[2] | 0.144 | -1.322 | 0.302 | -1.395 | 0.331 | 1.363 |
| | 7.061[3] | 7.422[3] | 7.526[3] | 6.815[3] | 0.949 | 0.690 | -0.297 | -0.920 | -2.216[2] | -2.495[3] |
| 7 | 1.390 | 3.395[3] | 2.230[2] | 0.947 | 0.750 | 0.118 | -0.707 | -1.510 | -3.003[3] | -4.721[3] |
| | -1.374 | -4.803[3] | -1.483 | 4.417 | -0.784 | 1.031 | 2.563[3] | 34.01[3] | 7.375[3] | 3.927[3] |
| | 0.535 | -1.099 | -0.235 | 0.518 | -1.278 | -0.607 | -0.093 | 0.918 | 1.158 | 0.564 |
| 8 | -0.720 | -3.208[3] | -0.702 | 0.191 | -0.661 | -0.092 | 0.577 | 0.417 | 1.342 | 0.814 |
| | -0.526 | -5.681[3] | -5.168[3] | 2.395[2] | -1.150 | -2.723[3] | 2.079[2] | -2.219[2] | 3.417[3] | nc |
| | -1.227 | -0.760 | 0.935 | 0.170 | 0.275 | 1.732[1] | 1.194 | 1.364 | 0.860 | -0.747 |
| 9 | 0.652 | 0.869 | 1.923[1] | 1.133 | 1.078 | 1.467 | 1.064 | 0.163 | 0.425 | 0.206 |
| | 0.224 | -0.034 | 0.890 | 1.229 | -1.502 | -0.098 | 0.219 | 0.167 | 0.504 | 0.738 |
| | -1.689[1] | 0.235 | 0.215 | -0.711 | 2.178[2] | 1.737[1] | 1.689[1] | -0.177 | -0.236 | -0.217 |
| 10 | -0.868 | nc | 0.564 | -1.288 | 0.689 | 1.062 | 0.864 | 0.637 | nc | -0.571 |
| | 0.978 | -0.403 | 1.374 | -1.535 | -2.224[2] | -0.450 | -0.982 | 1.161 | 0.395 | -1.375 |

continued on next page

continued from previous page

| Ser | SA vs ICP | SA vs CVP | SA vs BRP | SA vs AR | ICP vs CVP | ICP vs BRP | ICP vs AR | CVP vs BRP | CVP vs AR | BRP vs AR |
|---|---|---|---|---|---|---|---|---|---|---|
| | -2.137[2] | -2.362[2] | -0.377 | -0.762 | -0.300 | 2.049[2] | 1.786[1] | 2.285[2] | 1.938[1] | -0.987 |
| 11 | -2.080[2] | -1.397 | 0.385 | -0.780 | 1.844[1] | 3.036[3] | 2.304[2] | 2.169[2] | 1.466 | -0.915 |
| | -0.736 | -0.947 | -0.865 | -1.201 | -0.028 | 0.553 | 0.546 | 1.525 | 0.772 | 0.370 |
| | -1.267 | -1.598 | -1.260 | -1.094 | -0.547 | 0.760 | 0.605 | 1.183 | 0.983 | -0.034 |
| 12 | -1.215 | -1.648 | -1.235 | -1.199 | -1.588 | 0.288 | 0.251 | 1.579[1] | 1.461 | 0.153 |
| | 1.207 | -3.263[3] | -3.499[3] | -7.454[3] | -2.906[3] | -2.612[3] | -4.999[3] | 3.207[3] | 1.211 | -2.333[2] |
| | -1.068 | -1.472 | 0.699 | -1.718[1] | -0.532 | 1.372 | -0.156 | 1.814[1] | 0.411 | -2.162[2] |
| 13 | -12.46[3] | -2.285[2] | 2.704[3] | -1.356 | 0.268 | 9.644[3] | 0.655 | 3.815[3] | 0.376 | -3.567[3] |
| | -4.688[3] | 1.452 | 3.200[3] | -3.384[3] | 3.892[3] | 4.402[3] | 1.355 | 0.707 | -4.169[3] | -4.958[3] |
| | -0.100 | -1.237 | -0.096 | 1.705[1] | -0.990 | -0.018 | 0.109 | 0.651 | 1.248 | 0.102 |
| 14 | 0.033 | 0.708 | 1.773[1] | -1.195 | 0.335 | 2.522[2] | -0.042 | 2.186[2] | -0.735 | -1.797[1] |
| | 0.219 | 1.310 | 1.453[1] | -0.391 | 1.234 | 1.739[1] | -0.226 | 1.406 | -1.332 | -1.460 |
| | -0.158 | -0.972 | -0.789 | 0.645 | -1.138 | -0.646 | 0.745 | 0.824 | 1.7592[1] | 1.432 |
| 15 | -1.676[1] | -0.840 | -0.479 | 0.488 | -0.668 | -0.276 | 1.124 | 2.594[3] | 0.988 | 0.637 |
| | -0.033 | -0.155 | -0.015 | 0.261 | -0.176 | -0.010 | 0.322 | 0.624 | 0.260 | 0.147 |
| | 0.224 | 0.292 | 0.770 | 2.646[3] | 0.056 | 0.242 | 2.271[2] | 0.318 | 2.256[2] | 2.319[2] |
| 16 | -0.416 | -0.462 | -0.591 | 1.730[1] | 0.196 | -0.305 | 1.703[1] | -0.547 | 1.653[1] | 1.575[1] |
| | 3.258[3] | -0.108 | -0.559 | 3.385[3] | -0.818 | -0.910 | 2.515[2] | -0.942 | 2.300[2] | 1.811[1] |
| | 1.917[1] | 1.327 | 0.193 | 0.949 | -1.257 | -1.119 | -1.866[2] | -0.791 | -0.803 | 0.493 |
| 17 | 7.816[3] | 7.388[3] | 5.815[3] | 6.552[3] | 1.112 | -1.803[1] | -0.431 | -2.105[2] | -0.951 | 1.320 |
| | 5.985[3] | 5.830[3] | -10.02[3] | 5.330[3] | 1.384 | -8.006[3] | 0.634 | -7.852[3] | -0.400 | 7.779[3] |
| | 0.299 | -0.256 | -1.878[1] | -1.206 | -1.588 | -2.365[2] | -5.773[3] | -1.830[1] | -2.075[2] | 0.940 |
| 18 | 1.735[1] | 1.625 | -1.972[2] | 0.425 | -0.438 | -2.749[3] | -1.738[1] | -2.670[3] | -1.240 | 2.055[2] |
| | 2.312[2] | 1.648 | 2.232[2] | 1.566 | -1.897[1] | 0.102 | 0.012 | 0.937 | 1.013 | -0.077 |
| | 1.244 | 1.833[1] | -0.112 | 2.047[2] | 2.233[2] | -1.240 | 1.560 | -1.929[1] | 0.143 | 1.770[1] |
| 19 | 2.941[3] | 3.500[3] | 1.124 | 3.666[3] | 5.256[3] | -1.092 | 6.458[3] | -1.545 | 3.004[3] | 1.793[1] |
| | 3.210[3] | 3.255[3] | 0.521 | 3.285[3] | 1.916[1] | -3.333[3] | 2.111[2] | -3.255[3] | 0.506 | 3.104[3] |
| | -3.937[3] | -4.745[3] | -1.875[1] | -6.533[3] | -1.772[1] | 4.810[3] | -6.592[3] | 5.222[3] | -7.271[3] | -9.178[3] |
| 20 | -5.866[3] | -4.815[3] | -5.989[3] | -6.255[3] | -1.274 | 2.037[2] | 3.457[3] | 1.775[1] | 2.110[2] | -0.970 |
| | 1.589 | -4.875[3] | -15.12[3] | -2.651[3] | -4.213[3] | -3.187[3] | -5.170[3] | 4.448[3] | 3.870[2] | -0.213 |

continued on next page

continued from previous page

| Ser. | SA vs ICP | SA vs CVP | SA vs BRP | SA vs AR | ICP vs CVP | ICP vs BRP | ICP vs AR | CVP vs BRP | CVP vs AR | BRP vs AR |
|---|---|---|---|---|---|---|---|---|---|---|
| | 8.268[3] | 8.527[3] | -1.555 | 1.344 | 0.501 | -7.057[3] | -8.869[3] | -6.829[3] | -7.560[3] | 2.504[2] |
| 21 | 11.25[3] | 10.12[3] | 4.138[3] | 10.78[3] | -7.263[3] | -2.064[2] | -7.493[3] | -1.674[1] | -6.176[3] | 0.287 |
| | nc | nc | 5.841[3] | nc | -10.33[3] | 4.997[3] | 23.61[3] | 5.782[3] | nc | -4.215[3] |
| | -1.250 | -1.468 | 0.284 | -1.547 | -0.900 | 1.998[2] | 0.652 | 3.102[3] | 0.963 | -0.770 |
| 22 | -0.251 | 0.141 | 0.759 | -1.022 | 0.656 | 1.420 | -0.608 | 0.761 | -0.752 | -1.493 |
| | -2.919[3] | -1.397 | 0.893 | -7.030[3] | 0.916 | 3.354[3] | -1.588 | 3.427[3] | -1.456 | -2.801[3] |
| | 0.568 | -0.771 | 1.213 | nc | -1.386 | 0.815 | 0.764 | 2.319[2] | 2.185[2] | 0.036 |
| 23 | 1.633 | 1.332 | 2.420[2] | nc | -1.393 | 0.033 | 0.300 | 4.014[3] | 1.640 | 0.300 |
| | 1.522 | 1.249 | 4.887[3] | nc | -1.018 | 0.797 | -0.077 | 5.977[3] | 0.793 | -0.559 |
| | -9.171[3] | -1.942[1] | -2.148[2] | 2.782[3] | 8.854[3] | 8.154[3] | 9.691[3] | -1.537 | 3.336[3] | 3.685[3] |
| 24 | -8.586[3] | 0.018 | -1.353 | 3.288[3] | 8.825[3] | 8.698[3] | 9.249[3] | -1.480 | 3.228[3] | 3.508[3] |
| | -9.384[3] | -0.137 | -0.437 | 3.591[3] | 9.152[3] | 9.735[3] | 9.157[3] | -0.354 | 3.628[3] | 3.479[3] |
| | -0.152 | 0.664 | -2.014[2] | -1.407 | 0.801 | -1.400 | -0.603 | -2.017[2] | -1.300 | 1.653[1] |
| 25 | -0.001 | -0.634 | -1.399 | -1.473 | -1.621 | -1.127 | -1.220 | -0.825 | -0.941 | -2.099[2] |
| | 0.556 | 0.000 | -0.582 | -0.667 | 0.511 | -0.570 | -0.643 | -0.864 | -0.918 | -1.347 |
| | 0.157 | -0.699 | -1.104 | 1.382 | -0.902 | -0.787 | 0.594 | 0.218 | 1.517 | 1.778[1] |
| 26 | 7.448[3] | -0.136 | 1.178 | 1.229 | -1.203 | -3.267[1] | nc | 0.213 | 0.315 | 0.742 |
| | nc | -0.689 | 0.711 | 0.145 | nc | nc | nc | 0.496 | 0.434 | -0.767 |
| | -3.010[3] | 1.536 | -1.657[1] | 0.570 | 3.062[3] | 2.960[3] | 3.046[3] | -2.969[3] | -1.201 | 1.917[1] |
| 27 | -1.762[1] | 1.236 | -0.586 | 0.771 | 1.861[1] | 1.492 | 1.820[1] | -3.264[3] | -2.051[2] | 1.765[1] |
| | -1.552 | 0.746 | -0.255 | 0.614 | 1.766[1] | 1.536 | 1.717[1] | -1.532 | -0.353 | 3.035[3] |
| | -3.172[3] | -0.279 | -0.148 | 1.009 | 2.834[3] | 2.747[3] | 3.374[3] | 0.135 | 1.079 | 0.984 |
| 28 | -2.860[3] | -2.604[3] | -0.299 | 0.219 | 1.993[1] | 2.771[3] | 2.763[3] | 1.295 | 1.293 | 0.335 |
| | -3.323[3] | -1.412 | -0.614 | -0.300 | 2.967[3] | 3.389[3] | 3.937[3] | 0.867 | 0.598 | 0.255 |
| | -0.769 | -0.191 | 0.522 | 2.324[2] | 0.724 | 0.912 | 2.593[3] | 0.710 | 2.440[2] | 1.853[1] |
| 29 | 0.313 | -0.366 | 0.802 | 1.135 | -1.563 | 0.626 | 0.896 | 0.832 | 1.131 | 0.289 |
| | -0.577 | -0.150 | 2.138[2] | 0.505 | 1.250 | 1.632 | 0.961 | 1.708[1] | 0.646 | -0.461 |
| | -1.565 | -0.786 | -0.762 | 0.534 | 1.224 | 0.736 | 1.258 | -0.184 | 0.829 | 0.917 |
| 30 | 1.869[1] | -3.000[3] | 0.366 | 0.919 | -2.401[2] | -1.502 | -0.310 | 1.062 | 1.726[1] | 0.557 |
| | 1.036 | -2.482[2] | -1.641 | 0.234 | -2.130[2] | -1.968[2] | -0.670 | 1.312 | 1.677[1] | 1.245 |

continued on next page

continued from previous page

| Ser. | SA vs ICP | SA vs CVP | SA vs BRP | SA vs AR | ICP vs CVP | ICP vs BRP | ICP vs AR | CVP vs BRP | CVP vs AR | BRP vs AR |
|---|---|---|---|---|---|---|---|---|---|---|

# EFI
# The Economic Research Institute

**Reports since 1995**
A complete publication list can be found at www.hhs.se/efi

*Published in the language indicated by the title*

*2001*

**Andersson, P.,** Expertise in Credit Granting: Studies on Judgment and Decision-Making behavior.
**Björklund, C.,** Work Motivation - Studies of its Determinants and Outcomes.
**Center for Management and Organization 50 (1951-2001).**
**Charpentier, C.,** Uppföljning av kultur- och fritidsförvaltningen efter stadsdelsnämnds-reformen.
**Dahlén, M.,** Marketing on the Web - Empirical Studies of Advertising and Promotion Effectiveness.
**Ekelund, M.,** Competition and Innovation in the Swedish Pharmaceutical Market.
**Ericsson, D.,** Kreativitetsmysteriet – Ledtrådar till arbetslivets kreativisering och skrivandets metafysik.
**Eriksson, R.,** Price Responses to Changes in Costs and Demand.
**Frisell, L.,** Information and Politics.
**Giordani, P.,** Essays in Monetary Economics and Applied Econometrics.
**Gustavsson, P.,** Essays on Trade, Growth and Applied Econometrics.
**Hedlund, A.,** Konsumentens erfarenhet – och dess inverkan på livsmedelsinköp på Internet.
**Hill, M.,** Essays on Environmental Policy Analysis: Computable General Equilibrium Approaches Applied to Sweden.
**Hvenmark, J.,**Varför slocknar elden? Om utbrändhet bland chefer i ideella organisationer.
**Hägglund, P.B.,** Företaget som investeringsobjekt – Hur placerare och analytiker arbetar med att ta fram ett investeringsobjekt.
**Höök, P.,** Stridspiloter i vida kjolar, om ledarutveckling och jämställdhet.
**Johansson, C.,** Styrning för samordning.
**Kjellberg, H.,** Organising Distribution - Hakonbolaget and the efforts to rationalise food distribution, 1940-1960.
**Lange, F. och Wahlund, R.,** Category Management – När konsumenten är manager.
**Liljenberg, A.,** Customer-geared competition – A socio-Austrian explanation of Tertius Gaudens.
**Lindkvist, B.,** Kunskapsöverföring mellan produktutvecklingsprojekt.
**Ljunggren, U.,** Nyckeltal i grundskolan i Stockholms stad före och efter stadsdels-nämndsreformen.
**Läkemedel – Kostnad eller resurs för sjukvården?** (red. Jönsson, B.,).
**Löf, M.,** On Seasonality and Cointegration.
**Martensen, K.,** Essays on Entry Externalities and Market Segmentation.
**Matros, A.,** Stochastic Stability and Equilibrium Selection in Games.
**Mårtensson, P.,** Management Processes – An Information Perspective on Managerial Work.
**Nilsson, A.,** Market Transparency.
**Persson, B.,** Essays on Altruism and Health Care Markets.

**Skoglund, J.,** Essays on Random Effects Models and GARCH.
**Strand, N.,** Empirical Studies of Pricing.
**Thorén, B.,** Stadsdelsnämndsreformen och det ekonomiska styrsystemet - Om budget-avvikelser.

*2000*

**Berg-Suurwee, U.,** Styrning före och efter stadsdelsnämndsreform inom kultur och fritid – Resultat från intervjuer och enkät.
**Bergkvist, L.,** Advertising Effectiveness Measurement: Intermediate Constructs and Measures.
**Brodin, B., Lundkvist, L., Sjöstrand, S-E., Östman, L.,** Koncernchefen och ägarna.
**Bornefalk, A.,** Essays on Social Conflict and Reform.
**Charpentier, C., Samuelson, L.A.,** Effekter av en sjukvårdsreform.
**Edman, J.,** Information Use and Decision Making in Groups.
**Emling, E.,** Svenskt familjeföretagande.
**Ericson, M.,** Strategi, kalkyl, känsla.
**Gunnarsson, J., Wahlund, R., Flink, H.,** Finansiella strategier i förändring: segment och beteenden bland svenska hushåll.
**Hellman, N.,** Investor Behaviour – An Empirical Study of How Large Swedish Institutional Investors Make Equity Investment Decisions.
**Hyll, M.,** Essays on the Term Structure of Interest Rates.
**Håkansson, P.,** Beyond Private Label – The Strategic View on Distributor Own Brands.
**I huvudet på kunden.** Söderlund, M., (red).
**Karlsson Stider, A.,** Familjen och firman.
**Ljunggren, U.,** Styrning av grundskolan i Stockholms stad före och efter stadsdelsnämndsreformen – Resultat från intervjuer och enkät.
**Ludvigsen, J.,** The International Networking between European Logistical Operators.
**Nittmar, H.,** Produktutveckling i samarbete – Strukturförändring vid införande av nya Informationssystem.
**Robertsson, G.,** International Portfolio Choice and Trading Behavior.
**Schwarz, B., Weinberg, S.,** Serviceproduktion och kostnader – att söka orsaker till kommunala skillnader.
**Stenström, E.,** Konstiga företag.
**Styrning av team och processer – Teoretiska perspektiv och fallstudier.** Bengtsson, L., Lind, J., Samuelson, L.A., (red).
**Sweet, S.,** Industrial Change Towards Environmental Sustainability - The Case of Replacing Chloroflouorocarbons.
**Tamm Hallström, K.,** Kampen för auktoritet – Standardiseringsorganisationer i arbete.

*1999*

**Adler, N.,** Managing Complex Product Development.
**Allgulin, M.,** Supervision and Monetary Incentives.
**Andersson, P.,** Experto Credite: Three Papers on Experienced Decision Makers.
**Ekman, G.,** Från text till batong – Om poliser, busar och svennar.
**Eliasson, A-C.,** Smooth Transitions in Macroeconomic Relationships.
**Flink, H., Gunnarsson, J., Wahlund, R.,** Svenska hushållens sparande och skuldsättning– ett konsumentbeteende-perspektiv.
**Gunnarsson, J.,** Portfolio-Based Segmentation and Consumer Behavior: Empirical Evidence and Methodological Issues.
**Hamrefors, S.,** Spontaneous Environmental Scanning.

**Helgesson, C-F.,** Making a Natural Monopoly: The Configuration of a Techno-Economic Order in Swedish Telecommunications.
**Japanese Production Management in Sunrise or Sunset.** Karlsson, C., (red).
**Jönsson, B., Jönsson, L., Kobelt, G.,** Modelling Disease Progression And the Effect of Treatment in Secondary Progressive MS. Research Report.
**Lindé, J.,** Essays on the Effects of Fiscal and Monetary Policy.
**Ljunggren, U.,** Indikatorer i grundskolan i Stockholms stad före stadsdels-nämndsreformen – en kartläggning.
**Ljunggren, U.,** En utvärdering av metoder för att mäta produktivitet och effektivitet i skolan – Med tillämpning i Stockholms stads grundskolor.
**Lundbergh, S.,** Modelling Economic High-Frequency Time Series.
**Mägi, A.,** Store Loyalty? An Empirical Study of Grocery Shopping.
**Mölleryd, B.G.,** Entrepreneurship in Technological Systems – the Development of Mobile Telephony in Sweden.
**Nilsson, K.,** Ledtider för ledningsinformation.
**Osynlig Företagsledning.** Sjöstrand, S-E., Sandberg, J., Tyrstrup, M., (red).
**Rognes, J.,** Telecommuting – Organisational Impact of Home Based – Telecommuting.
**Sandström, M.,** Evaluating the Benefits and Effectiveness of Public Policy.
**Skalin, J.,** Modelling Macroeconomic Time Series with Smooth Transition Autoregressions.
**Spagnolo, G.,** Essays on Managerial Incentives and Product-Market Competition.
**Strauss, T.,** Governance and Structural Adjustment Programs: Effects on Investment, Growth and Income Distribution.
**Svedberg Nilsson, K.,** Effektiva företag? En studie av hur privatiserade organisationer konstrueras.
**Söderström, U.,** Monetary Policy under Uncertainty.
**Werr, A.,** The Language of Change The Roles of Methods in the Work of Management Consultants.
**Wijkström, F.,** Svenskt organisationsliv – Framväxten av en ideell sektor.

*1998*

**Andersson, M.,** On Testing and Forecasting in Fractionally Integrated Time Series Models.
**Berg-Suurwee, U.,** Styrning av kultur- och fritidsförvaltning innan stadsdelsnämnds-reformen.
**Berg-Suurwee, U.,** Nyckeltal avseende kultur- och fritidsförvaltning innan stadsdels-nämndsreformen.
**Bergström, F.,** Essays on the Political Economy of Industrial Policy.
**Bild, M.,** Valuation of Takeovers.
**Charpentier, C., Samuelson, L.A.,** Effekter av en sjukvårdsreform – en analys av Stockholmsmodellen.
**Eriksson-Skoog, G.,** The Soft Budget Constraint: The Emergence, Persistence and Logic of an Institution. The Case of Tanzania 1967-1992.
**Gredenhoff, M.,** Bootstrap Inference in Time Series Econometrics.
**Ioannidis, D.,** I nationens tjänst? Strategisk handling i politisk miljö – en nationell teleoperatörs interorganisatoriska, strategiska utveckling.
**Johansson, S.,** Savings Investment, and Economic Reforms in Developing Countries.
**Levin, J.,** Essays in Company Valuation.
**Ljunggren, U.,** Styrning av grundskolan i Stockholms stad innan stadsdelsnämnds-reformen.
**Mattsson, S.,** Från stat till marknad – effekter på nätverksrelationer vid en bolagise-ringsreform.
**Nyberg, A.,** Innovation in Distribution Channels – An Evolutionary Approach.

**Olsson, P.,** Studies in Company Valuation.
**Reneby, J.,** Pricing Corporate Securities.
**Roszbach, K.,** Essays on Banking Credit and Interest Rates.
**Runsten, M.,** The Association Between Accounting Information and Stock Prices. Model development and empirical tests based on Swedish Data.
**Segendorff, B.,** Essays on Bargaining and Delegation.
**Sjöberg, L., Bagozzi, R., Ingvar, D.H.,** Will and Economic Behavior.
**Sjögren, A.,** Perspectives on Human Capital: Economic Growth, Occupational Choice and Intergenerational Mobility.
**Studier i kostnadsintäktsanalys.** Jennergren, P., (red)
**Söderholm, J.,** Målstyrning av decentraliserade organisationer. Styrning mot finansiella och icke-finansiella mål.
**Thorburn, K.,** Cash Auction Bankruptcy and Corporate Restructuring
**Wijkström, F.,** Different Faces of Civil Society.
**Zethraeus, N.,** Essays on Economic Evaluation in Health Care. Evaluation of Hormone Replacement Therapy and Uncertainty in Economic Evaluations.

*1997*

**Alexius, A.,** Essays on Exchange Rates, Prices and Interest Rates.
**Andersson, B.,** Essays on the Swedish Electricity Market.
**Berggren, N.,** Essays in Constitutional Economics.
**Ericsson, J.,** Credit Risk in Corporate Securities and Derivatives. Valuation and Optimal Capital Structure Choice.
**Charpentier, C.,** Budgeteringens roller, aktörer och effekter. En studie av budget-processerna i en offentlig organisation.
**De Geer, H., Silfverberg, G.,** Citizens'Trust and Authorities' Choices, A Report from The Fourth International Conference on Ethics in the Public Service, Stockholm June 15-18, 1994.
**Friberg, R.,** Prices, Profits and Exchange Rates.
**Från optionsprissättning till konkurslagstiftning.** Bergström, C., Björk, T., (red)
**Hagerud, G.E.,** A New Non-Linear GARCH Model.
**Haksar, A.,** Environmental Effects of Economywide Policies: Case Studies of Costa Rica and Sri Lanka.
**He, C.,** Statistical Properties of Garch Processes.
**Holmgren, M.,** Datorbaserat kontrollrum inom processindustrin; erfarenheter i ett tidsperspektiv.
**Jennergren, P.,** Tutorial on the McKinsey Model for Valuation of Companies.
**Lagerlöf, J.,** Essays on Political Economy, Information, and Welfare.
**Lange, F., Wahlund, R.,** Planerade och oplanerade köp - Konsumenternas planering och köp av dagligvaror.
**Löthgren, M.,** Essays on Efficiency and Productivity; Contributions on Bootstrap, DEA and Stochastic Frontier Models.
**Nilsson, H.E.,** Using Conceptual Data Modelling in Management Accounting: A Case Study.
**Sjöberg, L., Ramsberg, J.,** En analys av en samhällsekonomisk bedömning av ändrade säkerhetsföreskrifter rörande heta arbeten.
**Säfvenblad, P.,** Price Formation in Multi-Asset Securities Markets.
**Sällström, S.,** On the Dynamics of Price Quality.
**Södergren, B.,** På väg mot en horisontell organisation? Erfarenheter från näringslivet av decentralisering och därefter.
**Tambour, M.,** Essays on Performance Measurement in Health Care.
**Thorén, B., Berg-Surwee, U.,** Områdesarbete i Östra Hökarängen - ett försök att studera effekter av decentralisering.

**Zhang Gang.,** Chinese Rural Enterprises Between Plan and Market.
**Åhlström, P.,** Sequences in the Profess of Adopting Lean Production.
**Åkesson, G.,** Företagsledning i strategiskt vakuum. Om aktörer och förändrings-processer.
**Åsbrink, S.,** Nonlinearities and Regime Shifts in Financial Time Series.

*1996*

**Advancing your Business. People and Information Systems in Concert.**
Lundeberg, M., Sundgren, B (red).
**Att föra verksamheten framåt. Människor och informationssystem i samverkan.**
red. Lundeberg, M., Sundgren, B.
**Andersson, P.,** Concurrence, Transition and Evolution - Perspectives of Industrial Marketing Change Processes.
**Andersson, P.,** The Emergence and Change of Pharmacia Biotech 1959-1995. The Power of the Slow Flow and the Drama of Great Events.
**Asplund, M.,** Essays in Industrial Economics.
**Delmar, F.,** Entrepreneurial Behavior & Business Performance.
**Edlund, L.,** The Marriage Market: How Do You Compare?
**Gunnarsson, J.,** Three Studies of Financial Behavior. Research Report.
**Hedborg, A.,** Studies of Framing, Judgment and Choice.
**Holgersson, C., Höök, P.,** Ledarutveckling för kvinnor - Uppföljning av en satsning på Volvo - Research Report.
**Holmberg, C.,** Stores and Consumers - Two Perspectives on Food Purchasing.
**Håkansson, P., Wahlund, R.,** Varumärken. Från teori till praktik.
**Karlsson, A.,** The Family Business as an Heirloom. Research Report.
**Linghag, S.,** Man är handelsstudent. Research Report.
**Molin, J.,** Essays on Corporate Finance and Governance.
**Mägi, A.,** The French Food Retailing Industry - A Descriptive Study.
**Mölleryd, B.,** Så byggdes en världsindustri - Entreprenörskapets betydelse för svensk mobiltelefoni. Research Report.
**Nielsen, S.,** Omkostningskalkulation for avancerede produktions-omgivelser - en sammenligning af stokastiske og deterministiske omkost-ningskalkulationsmodeller.
**Normark, P., Danielsson, L., Larsson, A., Lundblad, P.,** Kooperativa nyckeltal. Research Report.
**Sandin, R.,** Heterogeneity in Oligopoly: Theories and Tests.
**Sandén-Håkansson, U.,** Från kampanjmål till mediemix - en studie av samarbete mellan annonsörer, reklambyråer och mediebyråer. Research Report.
**Stein, J., Söderlund, M.,** Framgång i arbetet, strategier för att utföra arbetsuppgifterna, arbetsuppgifternas karaktär och utbildningskvalitet. En empirisk studie av civil-ekonomer. Research Report.
**Strömberg, P., Thorburn, K.,** An Empirical Investigation of Swedish Corporations in Liquidation Bankruptcy. Research Report.
**Söderlund, M.,** Och ge oss den nöjda kunden. En studie av kundtillfredsställelse och dess orsaker och effekter. Research Report.
**Thodenius, B.,** Användningen av ledningsinformationssystem i Sverige: Lägesbild 1995. Research Report.
**Ulfsdotter, U.,** Internationalisering för expansion eller hemmamarknadsförsvar? De nordiska marknaderna för fruktyoghurt 1982-1994.
**Westelius, A.,** A Study of Patterns of Communication in Management Accounting and Control Projects.
**Wijkström, F.,** Den svenska ideella sektorn och pengarna. Research Report.
**Örtendahl, M.,** Health and Time - A Problem of Discounting.