CREATES Research Paper 2010-1

Forecasting with nonlinear time series models

Anders Bredahl Kock and Timo Teräsvirta
and Mark Podolskij

School of Economics and Management
Aarhus University
Bartholins Allé 10, Building 1322, DK-8000 Aarhus C
Denmark

Forecasting with nonlinear time series models

Anders Bredahl Kock and Timo Teräsvirta

CREATES, Aarhus University
DK-8000 Aarhus C, Denmark

January 4, 2010

Abstract

In this paper, nonlinear models are restricted to mean nonlinear parametric models. Several such models popular in time series econometrics are presented and some of their properties discussed. This includes two models based on universal approximators: the Kolmogorov-Gabor polynomial model and two versions of a simple artificial neural network model. Techniques for generating multi-period forecasts from nonlinear models recursively are considered, and the direct (non-recursive) method for this purpose is mentioned as well. Forecasting with complex dynamic systems, albeit less frequently applied to economic forecasting problems, is briefly highlighted. A number of large published studies comparing macroeconomic forecasts obtained using different time series models are discussed, and the paper also contains a small simulation study comparing recursive and direct forecasts in a particular case where the data-generating process is a simple artificial neural network model. Suggestions for further reading conclude the paper.

Acknowledgements. This research has been supported by the Danish National Research Foundation. A part of the work for this paper was carried out when the second author was visiting the School of Finance and Economics, University of Technology, Sydney, whose kind hospitality is gratefully acknowledged.

JEL Classification: C22, C45, C52, C53

Keywords: forecast accuracy, Kolmogorov-Gabor, nearest neighbour, neural network, nonlinear regression

Note: This paper has been prepared for *Oxford Handbook on Economic Forecasting*, edited by M.P. Clements and D.F. Hendry, to be published by Oxford University Press.

1 Introduction

Nonlinear time series forecasting is quite common in science. Forecasts of riverflow, meteorological phenomena such as earth or sea temperatures or cloud coverage, sunspots, size of animal populations, future outcomes of biochemical processes, or medical time series, to name a few examples, are very often generated by nonlinear models. The most popular nonlinear forecasting models in these areas are complex dynamic systems based on the concept of chaos, and various neural network models. The former models are deterministic, although they may be assumed to be contaminated by stochastic variation. The latter are stochastic and parametric, and in spirit close to nonparametric models. They belong to the family of universal approximators, that is, they are very flexible and can be used to approximate rather general functional forms arbitrarily accurately.

Economic time series have traditionally been forecast by parametric linear models. More recently, nonlinear univariate and single-equation models have gained in popularity, although linear models still dominate. Perhaps because of the linear tradition, nonlinear forecasting of economic time series is often carried out by parametric nonlinear models such as switching or smooth transition regression models and hidden Markov or Markov-switching models. These models have the property that they nest a linear model and, depending on the application, may have an economic interpretation. They are less frequently used for forecasting outside economics where, as already mentioned, neural networks and nonparametric methods including chaos-based ones are more frequent.

Forecasts from parametric models to be discussed in this chapter are conditional means of the variable to be forecast. If the model and thus its conditional mean are linear, computing the conditional mean is easy. When the conditional mean is nonlinear, multi-step forecasting becomes more complicated, and numerical techniques are called for. This issue deserves a separate discussion.

The plan of this chapter is as follows. In Section 2 we shall consider a number of parametric time series models. Some universal approximators, including neural network models, will be studied in Section 3. Forecasting several periods ahead with nonlinear models is the topic of Section 4, and forecasting with chaotic systems is briefly considered in Section 5. Comparisons of linear and nonlinear forecasts of economic time series are discussed in Section 6 and studies comprising a large number of series in Section 7. Section 8 contains a limited forecast accuracy comparison between recursive and direct forecasts. Final remarks and suggestions for further reading can be found in Section 9.

2 Nonlinear time series models

2.1 Switching regression model

The standard switching regression (SR) model is defined as follows:

$$y_t = \sum_{j=1}^{r} (\boldsymbol{\phi}_j' \mathbf{z}_t + \varepsilon_{jt}) I(c_{j-1} < s_t \leq c_j) \tag{1}$$

where $\mathbf{z}_t = (\mathbf{w}_t', \mathbf{x}_t')'$ is a vector of explanatory variables, $\mathbf{w}_t = (1, y_{t-1}, ..., y_{t-p})'$ and $\mathbf{x}_t = (x_{1t}, ..., x_{kt})'$, s_t is an observable switch-variable, usually assumed to be a continuous stationary random variable, $I(A)$ is an indicator variable: $I(A) = 1$ when A is true, zero otherwise. Furthermore, $c_0, c_1, ..., c_r$ are switch or threshold parameters, $c_0 = -\infty$, $c_r = \infty$. Parameters $\boldsymbol{\phi}_j = (\phi_{0j}, \phi_{1j}, ..., \phi_{mj})'$ are such that $\boldsymbol{\phi}_i \neq \boldsymbol{\phi}_j$ for $i \neq j$, where $m = p + k + 1$, $\varepsilon_{jt} = \sigma_j \varepsilon_t$ with $\{\varepsilon_t\} \sim \text{iid}(0, 1)$, and $\sigma_j > 0$, $j = 1, ..., r$. It is seen that (1) is a piecewise linear model whose switch-points are generally unknown. If they are known, the model is linear. It is also linear if $r = 1$, that is, if there is only one regime. In many economic applications, the number of regimes is two, so eqn. (1) collapses into

$$y_t = (\boldsymbol{\phi}_1' \mathbf{z}_t + \varepsilon_{1t}) I(s_t \leq c_1) + (\boldsymbol{\phi}_2' \mathbf{z}_t + \varepsilon_{2t})\{1 - I(s_t \leq c_1)\}. \tag{2}$$

When $\mathbf{x}_t$ is absent and $s_t = y_{t-d}$, $d > 0$, (1) becomes the self-exciting threshold autoregressive (SETAR, or TAR for short) model. The univariate model has been frequently applied in economics. For a thorough account of the TAR model, see **?**.

A useful special case of the univariate TAR model is the one in which only the intercept is switching, whereas the autoregressive structure remains unchanged. Setting $\mathbf{w}_t = (1, \widetilde{\mathbf{w}}_t')'$ with $\widetilde{\mathbf{w}}_t = (y_{t-1}, ..., y_{t-p})'$, the model can be written as follows:

$$y_t = \sum_{j=1}^{r} \phi_{0j} I(c_{j-1} < y_{t-d} \leq c_j) + \boldsymbol{\phi}' \widetilde{\mathbf{w}}_t + \varepsilon_t \tag{3}$$

where $\boldsymbol{\phi} = (\phi_1, ..., \phi_p)'$ and $\varepsilon_t \sim \text{iid}(0, \sigma^2)$. This specification was suggested by **?** to characterize 'near unit root' behaviour. The switching intercept in (3) causes level shifts in realizations, although the process itself is stationary and ergodic when the roots of the lag polynomial $1 - \sum_{j=1}^{p} \phi_j z^j$ lie outside the unit circle. This model is a useful tool for modelling series that 'look' nonstationary but are stationary and fluctuate within bounds, such as interest rate series. Lanne and Saikkonen fit their model to two interest rate series that by definition cannot have a unit root.

Application of the SR model requires selecting the number of regimes r since it is typically not known in advance. It is a priori possible that $r = 1$, in which case a genuine SR model with $r > 1$ is not identified. Solutions to this specification problem can be found in **?**, **?**, **?** and **?**, see also **?**Teräsvirta, Tjøstheim and Granger (2010, Chapter 16). Estimation of SR models with $r = 2$ is carried out by a set of regressions, see for example **?**, because the threshold parameter has to be estimated using a grid. **?** showed how to do that when there is more than one threshold parameter, in which case $r > 2$. Assuming stationarity and ergodicity of the TAR model, **?** derived the asymptotic properties of maximum likelihood estimators of the parameters of the model. This included showing (in the case $r = 2$) that $\widehat{c}_1$, the maximum likelihood estimator of c_1 is super consistent and $T\widehat{c}_1$ (T is the sample size) is asymptotically independent of $\sqrt{T}\widehat{\boldsymbol{\phi}}_1$ and $\sqrt{T}\widehat{\boldsymbol{\phi}}_2$.

The switching regression model can be generalized to a vector model that may be called the vector switching regression or, in the absence of exogenous variables, the vector threshold autoregressive model. For various forms of this model, including the Threshold Cointegration model, see Teräsvirta et al. (2010, Chapter 3).

The TAR model has been applied to macroeconomic series such as GNP, industrial production, unemployment and interest rate series. **?** contains several examples of fitting a vector TAR model to economic and financial series.

2.2 Smooth transition regression model

The smooth transition regression (STR) model is a nonlinear model that bears some relationship to the switching regression model. The two-regime switching regression model with an observable switching variable is a special case of the standard STR model. The univariate smooth transition autoregressive (STAR) model contains the two-regime TAR model as a special case. The smooth transition regression model originated as a generalization of a switching regression model in the work of **?**. The authors considered two regression lines and devised a model in which the transition from one line to the other as a function of the sole explanatory variable is smooth instead of being abrupt. The STR model is defined as follows:

$$\begin{aligned} y_t &= \boldsymbol{\phi}'\mathbf{z}_t + \boldsymbol{\psi}'\mathbf{z}_t G(\gamma, \mathbf{c}, s_t) + \varepsilon_t \\ &= \{\boldsymbol{\phi} + \boldsymbol{\psi} G(\gamma, \mathbf{c}, s_t)\}'\mathbf{z}_t + \varepsilon_t, t = 1, ..., T \end{aligned} \tag{4}$$

where $\mathbf{z}_t$ is defined as in the preceding section, $\boldsymbol{\phi} = (\phi_0, \phi_1, ..., \phi_m)'$ and $\boldsymbol{\psi} = (\psi_0, \psi_1, ..., \psi_m)'$ are parameter vectors, $\mathbf{c} = (c_1, ..., c_K)'$ is a vector of location parameters, $c_1 \leq ... \leq c_K$, and $\varepsilon_t \sim \text{iid}(0, \sigma^2)$. Furthermore, the so-called transition function $G(\gamma, \mathbf{c}, s_t)$ is a bounded function of s_t, continuous everywhere in the parameter space for any value of the continuous transition variable s_t. The last expression in (4) indicates that the model can be interpreted as a linear model with stochastic time-varying coefficients $\boldsymbol{\phi} + \boldsymbol{\psi} G(\gamma, \mathbf{c}, s_t)$. The logistic transition function has the general form

$$G(\gamma, \mathbf{c}, s_t) = (1 + \exp\{-\gamma \prod_{k=1}^{K} (s_t - c_k)\})^{-1}, \quad \gamma > 0 \tag{5}$$

where $\gamma > 0$ is an identifying restriction. Equations (4) and (5) jointly define the logistic STR (LSTR) model. In applications, typically either $K = 1$ or $K = 2$. For $K = 1$, the parameter vector $\boldsymbol{\phi} + \boldsymbol{\psi} G(\gamma, c_1, s_t)$ changes monotonically from $\boldsymbol{\phi}$ to $\boldsymbol{\phi} + \boldsymbol{\psi}$ as a function of s_t. For $K = 2$, this vector changes symmetrically around the mid-point $(c_1 + c_2)/2$ where this logistic function attains its minimum value. The slope parameter γ controls the steepness and c_1 and c_2 the location of the transition function.

The usefulness of the LSTR model with $K = 1$ (LSTR1 model) is based on the fact that it is capable of characterizing asymmetric cyclical behaviour. It shares this property with the SR model. Suppose for example that s_t measures the phase of the business cycle. Then the LSTR1 model can describe economic growth processes whose dynamic properties are different in expansions from what they are in recessions, and where the effect of the exogenous variables on the growth rate may vary with the business cycle. The LSTR2 model is appropriate in situations where the dynamic behaviour of the process is similar at both large and small values of s_t and different in the middle. A three-regime SR model whose outer regimes are similar to each other while the mid-regime is different also has this property.

When $\gamma = 0$, the transition function $G(\gamma, \mathbf{c}, s_t) \equiv 1/2$, so the STR model (4) nests the linear model. At the other end of the scale, when $\gamma \to \infty$ the LSTR1 model approaches the SR model (2) with two regimes and $\sigma_1^2 = \sigma_2^2$. When $\gamma \to \infty$ in the LSTR2 model, the result is an SR model with three regimes such that the outer regimes are identical and the mid-regime different from the other two.

In practice, the transition variable s_t is a stochastic variable and very often an element of $\mathbf{z}_t$. There is a useful exception: $s_t = t$, which yields a linear model with deterministically changing parameters as seen from the last expression of (4). A univariate model of this type will be called the time-varying autoregressive (TV-AR) model.

When $\mathbf{z}_t = \mathbf{w}_t = (1, y_{t-1}, ..., y_{t-p})'$ in (4) and $s_t = y_{t-d}$ or $s_t = \Delta y_{t-d}$, $d > 0$, the STR model becomes a univariate smooth transition autoregressive

(STAR) model. This model can be generalised in various ways including a generalization to vector models, for discussion see Teräsvirta et al. (2010, Chapter 3).

Application of the STR or STAR model requires a modelling strategy: the model has to be specified, estimated and evaluated. Specification involves testing linearity and, if rejected, selecting the transition variable and K, and the appropriate parameter restrictions in (4). The parameters are estimated using numerical optimization techniques and the estimated model evaluated, among other things, by misspecification tests. This strategy has been discussed in several contributions: see, for example, **???**Teräsvirta (1998, 2004, 2006), Teräsvirta et al. (2010, Chapter 16) and **?**. For vector STAR models, see **?**.

The STAR model has been applied to the same macroeconomic series as the TAR model. Some forecasting applications will be considered in Section 7.

2.3 Markov-switching regression models

The observable regime indicator s_t in the SR model (1) may be replaced by an unobservable discrete stochastic variable θ_t that can take r different values $\{\nu_1, ..., \nu_r\}$, and is independent of ε_t. This gives another switching regression model, called the Markov switching (MS) or hidden Markov regression model. The sequence $\{\theta_t\}$ is assumed to follow a Markov chain, typically of order one, with transition (or staying) probabilities

$$p_{ij} = \Pr\{\theta_t = \nu_j | \theta_{t-1} = \nu_i\}, \; i, j = 1, ..., r. \tag{6}$$

The model is defined as follows:

$$y_t = \sum_{j=1}^{r} (\boldsymbol{\phi}_j' \mathbf{z}_t + \varepsilon_{jt}) I(\theta_t = \nu_j) \tag{7}$$

where $\mathbf{z}_t$ is defined as before, $\varepsilon_{jt} = \sigma_j \varepsilon_t$ with $\{\varepsilon_t\} \sim$ iid $\mathcal{N}(0, 1)$. Often, but not always, it is assumed that $\sigma_j = \sigma > 0$ for $j = 1, ..., r$. **?** considered this model and properties of the maximum likelihood estimators of its parameters. The univariate version of this model was introduced and fitted to a daily IBM stock return series by **?**.

The MS model (7) with (6) is a generalization of a linear dynamic regression model. Analogously, the linear vector autoregressive model may be generalized into a Markov switching (MS-VAR) one. For a comprehensive account of MS-VAR models, see **?**.

It may be noted, however, that the MS model (7) or its univariate version are not the most commonly applied Markov switching models in macroeconomics. Instead, many econometricians have preferred the following specification, due to **?**:

$$\begin{aligned} y_t &= \mu(\theta_t) + \sum_{j=1}^{p} \phi_j \{y_{t-j} - \mu(\theta_{t-j})\} + \varepsilon_t \\ &= \{\mu(\theta_t) - \sum_{j=1}^{p} \phi_j \mu(\theta_{t-j})\} + \sum_{j=1}^{p} \phi_j y_{t-j} + \varepsilon_t. \end{aligned} \tag{8}$$

where $\mu(\nu_i) \neq \mu(\nu_j)$ for $i \neq j$. From (8) it is seen that the flexibility of the parameterization is due to the switching intercept that can obtain r^{p+1} different values, whereas the autoregressive coefficients are constant and the roots of the lag polynomial $1 - \sum_{j=1}^{p} \phi_j \mathsf{z}^j$ lie outside the unit circle. In this respect the model resembles the intercept-switching TAR model of **?**, the difference being that in (8) the switching intercept is controlled by a latent variable. It should be noted that this model is not nested in the univariate autoregressive MS model (7) .

The number of regimes as well as the threshold variable in SR or TAR models are typically determined from the data. The number of regimes r in the Markov-switching model (7) or (8) is in principle also unknown *a priori.* Nevertheless, in economic applications it is most often chosen beforehand without any testing even when economic theory behind the model is not specific about the number of regimes. The most common choices are $r = 2$ and $r = 3$.

The model (8) has also been applied to several macroeconomic series such as GNP, industrial production, unemployment rate or interest rate series.

2.4 Other models

The three models already discussed, the SR (or TAR), the STR (or STAR), and the MS model seem to be the most commonly applied nonlinear time series models in economic forecasting. This list, however, must be completed by the artificial neural network model that will be considered separately in Section 3.2. It has been a popular forecasting device in many branches of science and has been applied to economic forecasting problems as well. In what follows, we shall briefly mention two other families of models that have been used in economic forecasting. They are the bilinear model and the family of random coefficient models.

2.4.1 Bilinear model

The bilinear model is a model containing both autoregressive and moving average terms such that the model is nonlinear in variables but linear in parameters. It has the following general form:

$$y_t = \phi_0 + \sum_{j=1}^{p} \phi_j y_{t-j} + \sum_{j=1}^{r} \sum_{k=1}^{s} \gamma_{jk} y_{t-j} \varepsilon_{t-k} + \varepsilon_t \tag{9}$$

where $\{\varepsilon_t\} \sim \mathrm{iid}(0, \sigma^2)$. For a review for bilinear models, see **?**. Due to the moving average terms, invertibility of the model is an issue. It is particularly important when the model is used for forecasting. Due to bilinear terms, analytic invertibility conditions for (9) only exist in some special cases. As mentioned in Teräsvirta et al. (2010, Section 3.5), most often the only way to check invertibility is to do it numerically. Bilinear models with suitable coefficients can generate realizations that display occasional deviating observations or short sequences of them. Such observations are in practice difficult to distinguish from outliers.

The bilinear model can be used as an example of a situation in which a stochastic process is white noise but nevertheless forecastable. Consider the following special case of (9):

$$y_t = \gamma_{21} y_{t-2} \varepsilon_{t-1} + \varepsilon_t.$$

It follows that $\mathsf{E} y_t = 0$ and $\mathsf{E} y_t y_{t-j} = 0$, $j \neq 0$, because $\varepsilon_t \sim \mathrm{iid}(0, \sigma^2)$. However, $\mathsf{E}\{y_{t+1}|\mathcal{F}_t\} = \gamma_{21} y_{t-1} \varepsilon_t$ where $\mathcal{F}_t = \sigma\{(y_{t-j}, \varepsilon_{t-j}) : j \geq 0\}$, so y_t is forecastable. For more examples of forecastable white noise models; see **?**. The bilinear model has not turned out to be very successful in economic forecasting; see however **?** who used the model for short-term forecasting of currency in circulation in Spain.

2.4.2 Random coefficient models

One way of generalizing standard linear models is to assume that their parameters are stochastic. The simplest alternative is that the parameters form a sequence of independent identically distributed random variables. This yields the following model:

$$y_t = \theta_0 + \boldsymbol{\theta}_t' \mathbf{z}_t + \varepsilon_t, \; t = 1, ..., T \tag{10}$$

where $\mathbf{z}_t$ is an $m \times 1$ vector of explanatory variables, $\{\boldsymbol{\theta}_t\} \sim \mathrm{iid}(\boldsymbol{\theta}, \boldsymbol{\Omega})$ with $\boldsymbol{\Omega} = [\omega_{ij}]$ a positive definite matrix and $\varepsilon_t \sim \mathrm{iid}(0, \sigma^2)$. Furthermore, ε_t

and $\boldsymbol{\theta}_t$ are mutually independent. If $\mathbf{z}_t = (y_{t-1}, ..., y_{t-m})'$, the model (10) is called the random coefficient autoregressive model. A notable thing about this model is that by writing $\boldsymbol{\theta}_t = \boldsymbol{\theta} + \boldsymbol{\phi}_t$ where $\{\boldsymbol{\phi}_t\} \sim \text{iid}(\mathbf{0}, \boldsymbol{\Omega})$, equation (10) can be reformulated as

$$y_t = \theta_0 + \boldsymbol{\theta}'\mathbf{z}_t + v_t \tag{11}$$

where $v_t = \varepsilon_t + \boldsymbol{\phi}_t'\mathbf{z}_t$. From this expression it is seen that the model becomes a linear model with constant coefficients but conditional heteroskedasticity. This is because $\mathsf{E}(v_t|\mathbf{z}_t) = 0$ and

$$\text{var}(v_t|\mathbf{z}_t) = \sigma^2 + \mathbf{z}_t'\boldsymbol{\Omega}\mathbf{z}_t. \tag{12}$$

Note that if $\mathbf{z}_t = (\varepsilon_{t-1}, ..., \varepsilon_{t-q})'$ and $\boldsymbol{\Omega} = \text{diag}(\omega_{11}, ..., \omega_{qq})$ in (11), the conditional variance (12) has an ARCH representation of order q.

In economic applications, the coefficient sequence is often not completely random but contains autoregressive structure. A well known special case is (10) where the sequence $\{\boldsymbol{\theta}_t\}$ is a random walk without drift, that is, $\{\Delta\boldsymbol{\theta}_t\}$ is a sequence of independent variables with zero mean and finite variance. For applications of this model to economic forecasting, see Marcellino (2002, 2004). Vector autoregressive models with random walk coefficients have become popular recently; see for example **??**Cogley and Sargent (2001, 2005).

3 Universal approximators

One may assume that y_t, the variable to be forecast, is affected by a vector of variables $\mathbf{z}_t$, but that the functional form of the relationship is unknown. In that case it would be useful to be able to approximate the unknown function by a general parametric function and use that function of $\mathbf{z}_t$ for forecasting y_t. This is where the so-called universal approximators have a role to play. Two such approximators, the Kolmogorov-Gabor polynomial and the single hidden-layer neural network, will be presented in this section.

In order to illustrate the concept of universal approximator, consider a possibly nonlinear function $f(\mathbf{z})$ of the vector of variables $\mathbf{z} = (z_1, ..., z_M)'$ that satisfies some regularity conditions. Suppose there exists another parametric function $g_N(\mathbf{z})$, where N is the number of parameters and $\delta > 0$ an arbitrary constant such that for an appropriate norm $|\cdot|$,

$$|f(\mathbf{z}) - g_N(\mathbf{z})| < \delta \tag{13}$$

for $N \leq N_0 < \infty$. The function $g_N(\mathbf{z})$ is called a universal approximator: it approximates $f(\mathbf{z})$ arbitrarily accurately with a finite number of parameters.

3.1 Kolmogorov-Gabor polynomial

Consider a nonlinear causal relationship between two processes: $\{x_t\}$ (input) and $\{y_t\}$ (output), both observable, and approximate it by the following equation **?**(see Priestley, 1981, p. 869):

$$y_t = \sum_{i=0}^{\infty} \theta_i x_{t-i} + \sum_{i=0}^{\infty}\sum_{j=i}^{\infty} \theta_{ij} x_{t-i} x_{t-j} + \sum_{i=0}^{\infty}\sum_{j=i}^{\infty}\sum_{k=j}^{\infty} \theta_{ijk} x_{t-i} x_{t-j} x_{t-k} + ... \tag{14}$$

The right-hand side of (14) is called the Volterra series expansion. If the lag-length, and thus the number of sums, is finite, it is called the Kolmogorov-Gabor (KG) polynomial. For further discussion, see Teräsvirta et al. (2010, Section 3.5).

The KG polynomial has been used to describe the (unknown) functional relationship between y and the vector $\mathbf{z}$. The polynomial model of y_t of order k then becomes

$$\begin{aligned} y_t &= \sum_{i_1=1}^{M} \alpha_{i_1} z_{i_1 t} + \sum_{i_1=1}^{M}\sum_{i_2=i_1}^{M} \alpha_{i_1 i_2} z_{i_1 t} z_{i_2 t} + ... \\ &+ \sum_{i_1=1}^{M}\sum_{i_2=i_1}^{M} ... \sum_{i_k=i_{k-1}}^{M} \alpha_{i_1 i_2 ... i_k} z_{i_1 t} z_{i_2 t} ... z_{i_k t} + \varepsilon_t \end{aligned} \tag{15}$$

where ε_t is the error term that is white noise. The KG polynomial is a universal approximator of the function $f(\mathbf{z})$, where $\mathbf{z} = (z_1, ..., z_M)'$, in the sense that under mild conditions, it satisfies the condition (13) when k, the order of the polynomial, is sufficiently high. It may be mentioned that the well known translog production function is based on a second-order KG polynomial. Note that if $z_{it} = y_{t-i}$, $i = 1, ..., M$, the estimated version of (15) for $k > 1$ is generally explosive and not useful in forecasting, except possibly in the very short run.

Although popular for instance in engineering, KG polynomial approximations to unknown nonlinear functional forms have not been in common use in economic forecasting. New developments in automated model selection, however, see for instance **?**, have generated interest in them. **?** used KG polynomials as a starting point for nonlinear model selection. The idea was to approximate well known nonlinear models such as the LSTR1 model introduced in Section 2.2 by appropriate special cases of these polynomials. **?** discussed linearity tests based on KG polynomials as they nest the linear model. In particular, the authors focussed on ways of parsimonious approximations to KG polynomials as testing tools. Interest in such approximations arises from the fact that the number of parameters in the KG polynomial increases quite rapidly with the number of variables. This in turn implies that the dimension of the null hypothesis of the linearity tests grows accordingly. When the number of variables is not small, tests relying on such approximations while having reasonable power are very useful. It may be mentioned that parameter-saving approximations to KG polynomials have interested researchers for a long time; see for example Ivakhnenko (1970, 1971). **??**

3.2 Artificial neural networks

As mentioned in the Introduction, artificial neural networks are a popular forecasting method in many branches of science. The family of neural network models is large, and many books and reviews have been devoted to them. For a review written for econometricians, see **?**.

Artificial neural networks are universal approximators. In this section we focus on two versions of a simple artificial neural network (ANN) model, the so-called 'single hidden-layer feedforward' model. It has the following form

$$y_t = \boldsymbol{\beta}_0'\mathbf{z}_t + \sum_{j=1}^{q} \beta_j G(\boldsymbol{\gamma}_j'\mathbf{z}_t) + \varepsilon_t \tag{16}$$

where y_t is the output series, $\mathbf{z}_t = (1, y_{t-1}, ..., y_{t-p}, x_{1t}, ..., x_{kt})'$ is the vector of inputs, including the intercept and lagged values of the output, $\boldsymbol{\beta}_0'\mathbf{z}_t$ is a linear unit with $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, ..., \beta_{0,p+k})'$. Furthermore, β_j, $j = 1, ..., q$, are parameters, called 'connection strengths' in the neural network literature. In (16), the component

$$g_N(\mathbf{z}_t, q) = \boldsymbol{\beta}_0'\mathbf{z}_t + \sum_{j=1}^{q} \beta_j G(\boldsymbol{\gamma}_j'\mathbf{z}_t)$$

satisfies the condition (13) for some $q \leq q_0 < \infty$. Only mild regularity conditions are then required for the unknown function $f(\mathbf{z})$, see for example **?**, **?**, **?** for discussion.

The function $G(\cdot)$ in (16) is a bounded, asymptotically constant function and $\boldsymbol{\gamma}_j$, $j = 1, ..., q$, are parameter vectors. It is often chosen to be a 'symmetric sigmoid', such as the logistic function. The errors ε_t are often assumed iid(0,σ^2). Often $\boldsymbol{\beta}_0 = (\beta_{00}, 0, ..., 0)'$ but in time series applications omitting the linear unit $\boldsymbol{\beta}_0'\mathbf{z}_t$ may not always be sensible. The term 'hidden layer' refers to the structure of (16). While the output y_t and the $m \times 1$ input vector $\mathbf{z}_t$ are observable, the linear combination $\sum_{j=1}^{q} \beta_j G(\boldsymbol{\gamma}_j'\mathbf{z}_t)$ is not. It thus forms a hidden layer between the 'output layer' y_t and 'input layer' $\mathbf{z}_t$.

Another rather common variant for $G(\cdot)$ in (16) is the radial-basis function. This function is radially symmetric around a centre $\mathbf{c}$, for example

$$G(\mathbf{z}_t, \mathbf{c}, \varphi) = \exp\{-\varphi^{-1}||\mathbf{z}_t - \mathbf{c}||\} \tag{17}$$

where $|| \cdot ||$ is the quadratic norm, $\mathbf{c}$ is a vector of parameters defining the centre and $\varphi > 0$ is the radius. When $\mathbf{z}_t = \mathbf{c}$, $G(\mathbf{z}_t, \mathbf{c}, \beta)$ obtains its maximum value unity and approaches zero when the distance of $\mathbf{z}_t$ from the centre increases. It has been generalized to the elliptic-basis function

$$G(\mathbf{z}_t, \mathbf{c}, \varphi) = \exp\{-||\mathbf{\Phi}^{-1/2}(\mathbf{z}_t - \mathbf{c})||\} \tag{18}$$

where $\mathbf{\Phi} = \mathrm{diag}(\varphi_1, ..., \varphi_m)'$, $\varphi_j > 0$, $j = 1, ..., m$. In (18), the deviations from the centre can have different weights for different elements of $\mathbf{z}_t$. Note, however, that the weights can also be changed by applying variance-changing transformations to these elements in (17). For more information, see for example Park and Sandberg (1991, 1994). **??**The radial- and elliptic-basis functions are universal approximators as well.

A statistical property separating the artificial neural network model (16) from the nonlinear time series models discussed in Section 2 is that it is only locally identified. This is because the hidden units are exchangeable. For example, letting any $(\beta_i, \boldsymbol{\gamma}_i')'$ and $(\beta_j, \boldsymbol{\gamma}_j')'$, $i \neq j$, change places in the equation does not affect the value of the likelihood function. Thus for $q > 1$ there always exists more than one observationally equivalent parameterization and, consequently, the likelihood function has $q!$ identical global maxima. Which one of these is reached when the likelihood function is maximized does not matter, but the existence of multiple maxima may sometimes cause problems in numerical maximization of the log-likelihood. For further discussion of identification of ANN models, see **?**.

Assume now that $\mathbf{z}_t = \mathbf{w}_t = (1, y_{t-1}, ..., y_{t-p})'$ in (16) , and let $\{\varepsilon_t\} \sim \mathrm{iid}(0, \sigma^2)$. Equation (16) can be written as follows:

$$y_t = \beta_0(\mathbf{w}_t) + \sum_{j=1}^{p} \beta_{0j} y_{t-j} + \varepsilon_t \tag{19}$$

where $\beta_0(\mathbf{w}_t) = \beta_{00} + \sum_{j=1}^{q} \beta_j G(\boldsymbol{\gamma}_j' \mathbf{w}_t)$. This shows that the switching-intercept TAR model (3) mentioned in Section 2.1 may be viewed as a special case of the more general autoregressive model (19). Another special case of (19) is obtained by assuming that the intercept in (19) equals $\beta_0(t/T) = \beta_{00} + \sum_{j=1}^{q} \beta_j G_j(t/T)$, where $G_j(t/T) = (1 + \exp\{-\gamma_j(t/T - c_j\})^{-1}$, $\gamma_j > 0$, and the roots of $1 - \sum_{j=1}^{p} \beta_{0j} z^j$ lie outside the unit circle. This yields a nonstationary autoregressive model with a deterministically fluctuating intercept, the so-called Switching-Mean Autoregressive (SM-AR) model, see ?. ? recently used the SM-AR model for medium-term forecasting of the euro area and UK inflation using monthly year-on-year time series.

Specification and estimation of ANN models may be a complicated exercise. Several algorithms, often computationally intensive, have been proposed in the literature as well as a specific-to-general technique based on statistical inference ?(Medeiros et al., 2006). Recent work by ?, however, considerably simplifies ANN modelling. White's idea was to convert the complicated specification and nonlinear estimation problem into a linear model selection problem. This was achieved by treating hidden units as variables by fixing their parameters, creating a very large set of them (the parameters in each one were fixed), and developing a specific-to-general algorithm called QuickNet for selecting the hidden units with the highest explanatory power from this set. It may be mentioned that instead of QuickNet, the recent automatic model selection algorithm Autometrics (Doornik 2008, 2009)?? may also be used for this purpose; see Section 8 for a small example.

Claims of success of ANN models in economic forecasting include ? who forecast daily exchange rates using ANN models and reported that their out-of-sample forecasts had a smaller root mean square forecast error (RMSFE) than a simple random walk model. On the other hand, ?, who considered both economic and noneconomic series, found that the ANN models often did not offer any improvement in forecast accuracy over the corresponding linear model. ?, who surveyed the area, also found the evidence mixed. Large-scale applications of ANN models in macroeconomic forecasting will be discussed separately in Section 7.

4 Forecasting with nonlinear time series models

4.1 Analytical point forecasts

Forecasts from a nonlinear model for more than one period ahead can most often only be obtained recursively using numerical techniques. Consider the following nonlinear model

$$y_t = g(x_{t-1}) + \varepsilon_t \tag{20}$$

where $\{\varepsilon_t\}$ is white noise, $\mathsf{E}\varepsilon_t = 0$ and $\text{var}(\varepsilon_t) = \sigma^2$. Assume further that the set of conditioning information $\mathcal{F}_t = \sigma\{x_{t-j} : j \geq 0\}$ and that ε_t is independent of $\mathcal{F}_{t-1}$. Given that the loss function is quadratic, the minimum average loss forecast for y_{T+h} from (20) at time T equals the conditional mean

$$y_{T+h|T} = \mathsf{E}\{y_{T+h}|\mathcal{F}_T\} = \mathsf{E}\{g(x_{T+h-1})|\mathcal{F}_T\}. \tag{21}$$

When $h = 1$ in (21), $y_{T+1|T} = g(x_T)$. When $h \geq 2$, however, the conditional expectation (21) can in general only be calculated numerically. Nevertheless, there are cases, in which the forecast can still be obtained analytically. As an example assume, for simplicity, that x_t follows a stationary first-order autoregressive model:

$$x_t = \phi x_{t-1} + \eta_t \tag{22}$$

where $|\phi| < 1$, and $\{\eta_t\} \sim \text{iid}(0, \sigma_\eta^2)$. Furthermore, assume that

$$g(x_t) = \alpha_1 x_t + \alpha_{11} x_t^2 \tag{23}$$

that is, a second-order KG polynomial. Then the forecast for y_{T+2} equals

$$\begin{aligned} y_{T+2|T} &= \mathsf{E}\{y_{T+2}|\mathcal{F}_T\} = \mathsf{E}\{g(x_{T+1}) + \varepsilon_{t+2}|\mathcal{F}_T\} \\ &= \alpha_1 \mathsf{E}\{\phi x_T + \eta_{T+1}|\mathcal{F}_T\} + \alpha_{11}\mathsf{E}\{(\phi x_T + \eta_{T+1})^2|\mathcal{F}_T\} \\ &= \alpha_1 \phi x_T + \alpha_{11}(\phi^2 x_T^2 + \sigma_\eta^2). \end{aligned}$$

Generally, for $h \geq 3$,

$$y_{T+h|T} = \alpha_1 \phi^{h-1} x_T + \alpha_{11}\{\phi^{2(h-1)} x_T^2 + (1 + ... + \phi^{2(h-2)})\sigma_\eta^2\}.$$

As another example, consider the following bilinear model:

$$y_t = \varepsilon_t + \gamma_{11} y_{t-1}\varepsilon_{t-1} + \gamma_{22} y_{t-2}\varepsilon_{t-2}.$$

Then,

$$y_{T+2|T} = \gamma_{11}\mathsf{E}\{y_{T+1}\varepsilon_{T+1} + \gamma_{22} y_T \varepsilon_T|\mathcal{F}_T\} = \gamma_{11}\sigma^2 + \gamma_{22} y_T \varepsilon_T$$

and, for $h \geq 3$,

$$y_{T+h|T} = \sigma^2(\gamma_{11} + \gamma_{22}) = \mathsf{E} y_T$$

Other examples of analytical forecasts for nonlinear models can be found in **?**.

4.2 Recursive point forecasts

Consider again the nonlinear model (20) and the forecast (conditional mean) (21) for $h = 2$. Assume that η_t has a continuous distribution with the density $f(\eta_t)$. Then the forecast for two periods ahead becomes

$$\begin{aligned} y_{T+2|T} &= \mathsf{E}\{g(x_{T+1})|\mathcal{F}_T\} = \mathsf{E}g(\phi x_T + \eta_{T+1}|\mathcal{F}_T) \\ &= \int_{-\infty}^{\infty} g(\phi x_T + \eta_{T+1}) f(\eta_{T+1}) \mathrm{d}\eta_{T+1}. \end{aligned} \tag{24}$$

The forecast can be obtained by numerical integration when $f(\eta_{T+1})$ is known. This so-called 'exact' method becomes computationally more complex, however, when $h > 2$, as the integral becomes a multiple integral. Three other methods for obtaining $y_{T+2|T}$ have been suggested in the literature; see for example **?**Granger and Teräsvirta (1993, Chapter 9) or Teräsvirta et al. (2010, Chapter 14). The first method is called 'naïve', which means that the presence of η_{T+1} in (24) is ignored by putting its value to zero. This implies assuming $\mathsf{E}g(x) = g(\mathsf{E}x)$, which is true when $g(x)$ is affine but not generally. The naïve forecast $y^n_{T+2|T} = g(\phi x_T)$.

As already mentioned, numerical integration becomes tedious when the forecast horizon increases. It will then be easier to obtain the forecast by simulation. This is called the Monte Carlo method. The 'Monte Carlo forecast' is

$$y^{MC}_{T+2|T} = \frac{1}{N}\sum_{j=1}^{N} g(\phi x_T + z_j)$$

where $z_j, j = 1, \ldots, N$, are random numbers drawn independently from the distribution of η_{T+1}. This forecast is an approximation to $y_{T+2|T}$, and for N large enough, it should be practically identical to the one obtained by the exact method.

It is also possible to compute the forecast by the bootstrap. In that case,

$$y^{B}_{T+2|T} = \frac{1}{N_B}\sum_{j=1}^{N_B} g(\phi x_T + \eta_t^{(j)})$$

where $\widehat{\eta}_t^{(j)}$, $j = 1, \ldots, N_B$, are the N_B independent draws with replacement from the set of residuals $\{\widehat{\eta}_t\}_{t=2}^T$ estimated from (22) over the sample period $[1, T]$. Compared to the Monte Carlo method, this variant has the advantage that the errors η_t can be allowed to be unconditionally heteroskedastic.

These techniques can clearly be used for multi-step forecasts ($h > 2$). For example the exact three-step forecast is

$$\begin{aligned} y_{T+3|T} &= \mathsf{E}\{g(x_{T+2}) + \varepsilon_{T+3})|\mathcal{F}_T\} = \mathsf{E}\{g(\phi x_{T+1} + \eta_{T+2})|\mathcal{F}_T\} \\ &= \mathsf{E}\{g(\phi^2 x_T + \phi\eta_{T+1} + \eta_{T+2})|\mathcal{F}_T\}. \end{aligned} \tag{25}$$

The naïve forecast is easy to compute because it ignores η_{T+1} and η_{T+2}, but the exact forecast (25) now involves a double integral. The Monte Carlo method requires draws from a bivariate distribution, but with independent components. The recommended bootstrap forecast based on pairs of subsequent residuals $(\widehat{\eta}_t^{(j)}, \widehat{\eta}_{t+1}^{(j)})$:

$$y_{T+3|T}^B = \frac{1}{N_B}\sum_{j=1}^{N_B} g(\phi^2 x_T + \phi\widehat{\eta}_{t+1}^{(j)} + \widehat{\eta}_t^{(j)}).$$

This implies that the order of the observed residuals is retained, in case the assumption of the independence of errors is incorrect. The Monte Carlo and the bootstrap method are computationally easier than the exact method when h increases. If the distribution of η_t is known, the Monte Carlo method will be the better of the two, but otherwise the bootstrap is the preferred alternative.

4.3 Direct point forecasts

There exists yet another forecasting method, the so-called 'direct' method. It appears computationally attractive, because the recursions are avoided altogether. As an example, assume that $h = 2$ and, furthermore, that the nonlinear model (20) is approximated by

$$y_{t+2} = g_2(x_t, y_t) + \varepsilon_t^*$$

where $\{\varepsilon_t^*\}$ may be autocorrelated but has mean zero and does not depend on x_t or y_t. Since $(x_t, y_t) \in \mathcal{F}_t$, the minimum average loss forecast of y_{T+2} given $\mathcal{F}_t$ becomes

$$y_{T+2|T}^D = g_2(x_t, y_t).$$

The function $g_2(\cdot)$ has to be specified and estimated separately, rather than derived from the one-step representation (20). A different function is required for each forecast horizon. A linear function is a popular choice in practice. The study by **?** constitutes an exception: the authors used STAR and neural network models for the purpose. Results of this study will be discussed in Section 7.1.

5 Forecasting with complex dynamic systems

The forecasting method of this section has its origins in chaos theory. The method was introduced by **?** and has also been applied to economic time series. It builds on the assumption that $\{y_t\}$ is a chaotic process. Consider a block of length m from the available past of the observed series y_t, $t = 1, ..., T$, and denote it by $\mathbf{y}_t^m = (y_t, y_{t-1}, \ldots, y_{t-(m-1)})$. The choice of m is left to the user. There are $T - m + 1$ such blocks. The most recent block is thus $\mathbf{y}_T^m$, and it is the one used for (short-term) forecasting. The one-step-ahead forecast is obtained as follows. First, carry out a search to find the k earlier blocks that are closest to $\mathbf{y}_T^m$, according to some distance measure $d(\mathbf{y}_T^m, \mathbf{y}_t^m)$, such as

$$d(\mathbf{y}_T^m, \mathbf{y}_t^m) = \sum_{i=0}^{m-1} |y_{T-i} - y_{t-i}| \, .$$

These are the k nearest neighbours $\mathbf{y}_{t_1}^m, ..., \mathbf{y}_{t_k}^m$ of $\mathbf{y}_T^m$. Second, run the regression

$$y_{t_i+1} = \beta_0 + \sum_{j=0}^{m-1} \beta_j y_{t_i-j} + \varepsilon_{t_i+1}, \; i = 1, ..., k \tag{26}$$

where $k > m$. The nearest-neihgbour forecast of y_{T+1} has the form

$$y_{T+1|T}^{NN} = \widehat{\beta}_0 + \sum_{j=1}^{m} \widehat{\beta}_j y_{T+1-j}$$

where $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_m$ are OLS estimates of the parameters of (26). The method can be generalized to the case where $h > 1$, but in practice its accuracy deteriorates quite rapidly with an increasing forecast horizon. **?** applied the method for one-month-ahead forecasting of the growth of the US industrial production. They found that their forecasts were generally more accurate than the ones obtained from linear vector autoregressive models, but the improvement was not statistically significant. **?** considered European daily exchange rates and reported nearest neighbour forecasts that were superior to ones obtained from simple linear models.

The nearest neighbour method is a nonparametric forecasting method. It may also be viewed as a quantitative version of analogy forecasting, in which a system to be forecast is compared to another system that has been in the same state as the current system is at the moment of forecasting. Analogy forecasting has been used for instance in weather or technology forecasting and in economics forecasting economic development of a country or countries.

6 Comparing linear and nonlinear point forecasts

A frequently asked question in forecasting economic time series with nonlinear time series models is whether nonlinear models yield forecasts that are more accurate than the ones obtained from linear models. A number of small-scale studies comparing a limited number of nonlinear models with corresponding linear ones exist, and some of them are listed in **?**. Many of these studies find that nonlinear models do not perform better than linear ones. This suggests that nonlinearity in economic series is often not 'strong' enough to make a difference in forecasting. For example, suppose that the 'nonlinear episodes', that is, events that require a nonlinear time series in order to be properly explained by a time series model, are relatively rare. Then gains from forecasting with an appropriate nonlinear model may not show in standard measures of accuracy of point forecasts such as the RMSFE. This is illustrated by the findings of **?**. They considered forecasting the US unemployment rate and found that the nonlinear threshold autoregressive and the Markov-switching autoregressive model of **?** outperformed the linear autoregressive model during periods of rapidly increasing unemployment but not elsewhere.

? conducted a simulation experiment to investigate this idea further. The authors generated one million observations from the following stationary LSTAR model:

$$y_t = -0.19 + 0.38(1 + \exp\{-10 y_{t-1}\})^{-1} + 0.9 y_{t-1} + 0.4\varepsilon_t \tag{27}$$

where $\{\varepsilon_t\} \sim \text{iid}\mathcal{N}(0,1)$. This model is a smooth transition version of the switching intercept TAR model (3) of **?** and has the property that the realizations intermittently fluctuate around two local means, -1.9 and 1.9, respectively, moving from one regime to the other only rarely. The 'nonlinear observations' were defined to be the ones corresponding to large (> 0.2 or < -0.2) changes in the value of the transition function. The number of such observations was 6282. Every nonlinear observation was chosen to be the last observation in a 1000 observations long subseries, which gave 6282 subseries. An LSTAR1 model was fitted to each subseries and forecasts up until 20 periods ahead generated from the model.

Due to infrequent and relatively rare switches, standard unit root tests when applied to these series typically did not reject the unit root hypothesis. For this reason the authors fitted a linear first-order autoregressive model to the corresponding differenced series Δy_t and model to the subseries and generated forecasts from that ARI(1,1) model as well. Figure 1 contains the

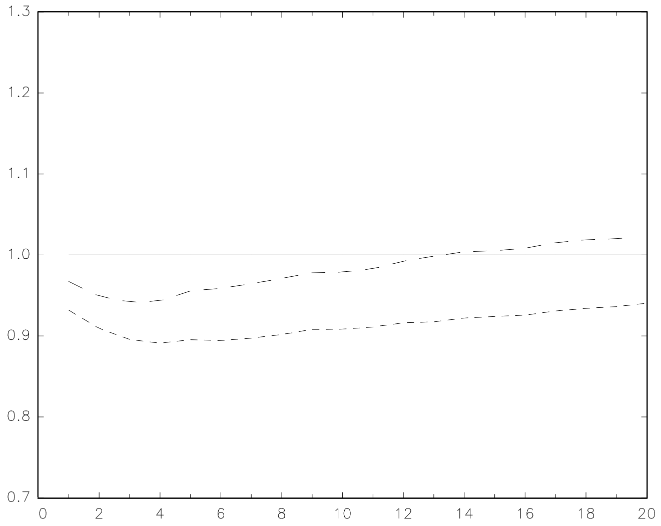

Figure 1: Ratio of the RMSFEs of the forecasts from the LSTAR model (27) and the ARI(1,1) model estimated from 6282 time series whose last observation represents a large change. Dashed curve, long dashes: the parameters of the LSTAR model are estimated, Dashed curve, short dashes: the parameters are known. Source: Lundbergh and Teräsvirta (2002).

ratio of the RMSFEs of the LSTAR and ARI forecasts in two cases: (i) the parameters of the LSTAR model are known, and (ii) they are unknown and estimated. As may be expected, the forecasts from the LSTAR model are more accurate than the ones from the ARI model when the parameters of the former model are known. When they are estimated (the realistic case), the model loses its predictive edge when the forecast horizon exceeds thirteen. This shows that a part of the advantage gained by knowing the nonlinear model disappears when the parameters of the model are estimated. That the ARI model is so competitive against the LSTAR model may appear surprising at first. A plausible explanation is given in **?**Clements and Hendry (1999, Chapter 5). They argued that first differences are useful in forecasting in case of structural breaks in levels, because the model based on them is flexible and adapts quickly to the new level even if it is misspecified. Rare but reasonably rapid shifts in levels are characteristic for realizations from the simulated STAR model (27), which offers an obvious explanation to the success of the linear ARI model. This simulation experiment thus illustrates two things. First, even a misspecified linear model may sometimes yield more accurate point forecasts than a correctly specified nonlinear model. Second, estimation of the parameters of a nonlinear model may have a large effect on forecast accuracy when the time series are short. In the present example it still has a non-negligible effect, although the series are rather long.

7 Large comparisons of economic forecasts

7.1 Forecasting with a separate model for each forecast horizon

As already mentioned, linear and nonlinear forecasts of economic time series have been compared in a number of studies. These studies have typically involved a small number of models and time series. In this section we shall consider a number of large-scale forecast comparisons that involve a large amount of series, several forecast horizons, and a large number of forecasts. One of the most extensive investigations of that kind appears to be the study by **?** who forecast 215 monthly US macroeconomic variables using the direct method described in Section 4.3. The observation period was mostly 1959(1)–1996(12), although some time series were shorter than that. The series covered most types of macroeconomic variables from production, consumption, money and credit series to stock returns. All series were seasonally adjusted. In two large studies, Marcellino (2002, 2004)**??** focussed on forecasting macroeconomic variables of the countries of the European Union. The forecasts were also generated using the direct method.

The study of **?** involved two types of nonlinear models: a 'tightly parameterized' model which was the LSTAR model of Section 2.2 and a 'loosely parameterized' one, which was the autoregressive neural network model. The authors experimented with two families of AR-NN models: one with a single hidden layer as the model (16) in Section 3.2, and a more general family with two hidden layers. Various linear autoregressive models were included as well as models of exponential smoothing. Several methods of combining forecasts were included in the comparisons. The total number of models or methods to forecast each series was 63.

The models were either completely specified in advance or the number of lags was specified using AIC or BIC. The models were built on either levels of the series or first differences. The forecast horizons were 1, 6 and 12 months. The series were forecast every month, beginning after an initial period of 120 observations. The authors trimmed their forecasts, which means that if the forecast was outside the range of observations in the sample, the forecast was set equal to the sample mean. The same idea, although in a slightly different form, was already applied in forecasting exercises of **???**Swanson and White (1995, 1997a, b) who compared forecasts from linear and ANN models. They called their trimming device the insanity filter: it 'replaced insanity with ignorance'.

The forecasting methods were ranked according to several loss functions. The individual nonlinear models did not seem to perform better than the linear ones. In one comparison, the 63 different models and methods were ranked on forecast performance using three different loss functions, the absolute forecast errors raised to the power one, two, or three, and the three forecast horizons. The best ANN forecast had rank around ten, whereas the best STAR model typically did worse than that. The best linear models were better than the STAR models and, at longer horizons than one month, better than the ANN models.

A noteworthy result was that combining the forecasts from all nonlinear models generated forecasts that were among the most accurate in rankings. They were among the top five in 53% (models in levels) and 51% (models in differences) of all cases when forecasting one month ahead. This was by far the highest fraction of all methods compared. In forecasting six and twelve months ahead, these percentages were lower but still between 30% and 34%. At these horizons, the combinations involving all linear models had a comparable performance. No single model performed equally well.

A general conclusion that can be drawn from the study of Stock and Watson is that there was some exploitable nonlinearity in the series under consideration but that it was too diffuse to be captured by a single nonlinear model. Interestingly, a similar conclusion emerges from the study of **?**. In their case the nonparametric forecasts generated as described in Section 5 were not the most accurate ones individually, but combining them yielded superior forecasts.

? reported results on forecasting 480 variables representing the economies of the twelve countries of the European Monetary Union. There were 58 models but, unlike Stock and Watson and **?**, combining forecasts from them was not considered. In addition to purely linear models, linear models with stochastic coefficients, each following a random walk, ANN models and logistic STAR models were included in the study.

The results of the study were based on rankings of model performance measured using five different symmetric loss functions. Neither neural network nor LSTAR models appeared in the overall top-10. But then, both the fraction of neural network models and LSTAR models that appeared in top-10 rankings for individual series was greater than the same fraction for linear or stochastic-coefficient AR models. One was able to conclude that nonlinear models in many cases work very well but that they can also relatively often perform rather poorly.

These studies suggest some answers to the question of whether nonlinear models perform better than linear ones in forecasting macroeconomic series. The results in **?** indicated that using a large number of nonlinear models and combining forecasts from them increases forecast accuracy compared to relying on single nonlinear models. It also seemed that this may lead to better forecasting performance than what is achieved by linear models. But then, the results in **?** did not unanimously support this conclusion. From **?** one was able to conclude that nonlinear models are uneven performers but that they can do well in forecasting some types of macroeconomic series such as unemployment rates.

7.2 Forecasting with the same model for each forecast horizon

Contrary to the articles considered in the previous section, the studies reported in **?** and **?** were based on recursive multi-period forecasts. The authors of the former study were interested in the effects of careful specification of the model on the forecast accuracy. This implied, among other things, testing linearity and choosing a nonlinear model only if linearity was rejected. Thus it occurred that a linear model was employed for some periods, and due to respecification of the model when new observations became available, even the structure of the nonlinear model was varying over time.

? considered seven monthly macroeconomic variables of the G7 countries. They were industrial production, unemployment, volume of exports, volume of imports, inflation, narrow money, and short-term interest rate. The number of time series was 47 as two series were too short to be considered. The series were seasonally adjusted with the exception of the inflation and short-term interest rate series. As in **?**, the series were forecast every month and the models respecified every 12 months.

The models applied were the linear autoregressive model, the LSTAR model and the ANN model (16). For some series linearity was never rejected. It was rejected somewhat more frequently against the LSTAR than the ANN model. In order to find out whether modelling was a useful idea, the investigation also included a set of models with a predetermined form and lag structure that did not change over time.

Results were reported for four forecast horizons: one, three, six and 12 months. They indicated that careful modelling does improve the accuracy of forecasts compared to selecting fixed nonlinear models, when the loss function is the RMSFE. The LSTAR model turned out to be the best model overall, better than the linear or neural network model, which was not the case in

? or **?**. There were series/country pairs, however, for which other models performed clearly better than the LSTAR model. Nevertheless, as in **?**, the LSTAR model did well in forecasting the unemployment rate.

? also considered combinations of forecasts and found that in many cases, but not systematically, they did improve forecast accuracy compared to individual models. But then, the combined forecasts only consisted of pairs of forecasts, so the results cannot be regarded as very informative in assessing the usefulness of combined forecasts.

The dataset of **?** consisted of 47 macroeconomic time series from the G7 and Scandinavian countries. His forecast comparisons also included the KG polynomial model whose performance turned out to be roughly comparable to that of ANN models with logistic hidden units. The results supported the notion of combining forecasts from nonlinear models. The forecasts thus obtained were on the average more accurate than ones from linear AR models.

8 Comparing recursive and direct forecasts

There is not much literature on comparing the direct forecasting method with the recursive ones. Using 170 US macroeconomic series **?** found that on the average the recursive method generated more accurate point forecasts than the direct method. Their results are valid for linear models, but comparable results for nonlinear data-generating processes do not seem to exist in the literature.

We conducted a small simulation study that sheds some light on this issue. It will be a part of a considerably larger study that at this time is still in progress. We chose a strongly nonlinear model from **?**. These authors took the well known annual Wolf's sunspot number series and, after transforming the observations as in **?**, fitted an ANN model (16) with two hidden units to the transformed series. This means that our data-generating process (DGP) is

$$\begin{aligned} y_t &= -0.17 + 0.85y_{t-1} + 0.14y_{t-2} - 0.31y_{t-3} + 0.08y_{t-7} \\ &\quad +12.80G_1(\mathbf{y}_{t-1}) + 2.44G_2(\mathbf{y}_{t-1}) + \varepsilon_t \end{aligned} \tag{28}$$

where the two hidden units were

$$\begin{aligned} G_1(\mathbf{y}_{t-1}) &= (1+\exp\{-0.46(0.29y_{t-1} - 0.87y_{t-2} + 0.40y_{t-7} - 6.68)\})^{-1} \\ G_2(\mathbf{y}_{t-1}) &= (1+\exp\{-1.17\times 10^3(0.83y_{t-1} - 0.53y_{t-2} - 0.18y_{t-7} + 0.38)\})^{-1} \end{aligned}$$

and $\varepsilon_t \sim \mathrm{iid}\mathcal{N}(0, 1.89^2)$. We generated 100 realisations with 600 observations from this model, specified and estimated an ANN model for each realisation and forecast with it.

The ANN models were built using two different methods. One was QuickNet of **?** as described in Section 3.2. We thus formed a pool of 1002 hidden units that included the two in the DGP (28) and let the algorithm select the relevant hidden units from that pool. The other was Autometrics, see Doornik (2008, 2009). This algorithm relied on the same pool of hidden units as QuickNet, but the rules of selecting the units were different. One important difference was that while QuickNet sequentially adds hidden units to the model, Autometrics can also remove them if necessary.

It is known, see for example **?**, that an estimated ANN model with a linear unit can sometimes be explosive even when the time series looks stationary. For this reason we considered not only the original forecasts but also the ones obtained by applying the insanity filter resembling that of Swanson and White (1995a, b, 1997). Our filter works as follows. If the difference $y_{T+h|T} - y_T$, where y_T is the last observation, lies outside the observed bounds defined by the minimum and maximum of $y_t - y_{t-h}$ in the set of the last 120 observations, the forecast $y_{T+h|T}$ is set equal to the last obseration y_T ('no change' forecast). The filter was strictly speaking only necessary in generating recursive forecasts, although it was also applied to direct forecasts.

Table 1: Ratios of the root mean square forecast error of the linear AR(10) model and the benchmark and two ANN models and the benchmark. No insanity filter applied

Model	Horizon		
	1	2	5
True model (16)	(1.7915)	(2.7345)	(4.0849)
AR(10), recursive	1.3326	1.3353	1.3543
AR(10), direct	1.3326	1.3356	1.3064
QuickNet, recursive	1.1651	1.2372	180.8
QuickNet, direct	1.1651	1.2466	1.1169
Autometrics, recursive	1.0247	0.9732	3.5097
Autometrics, direct	1.0247	1.1928	1.1308

Figures in parentheses are the RMSFEs from the benchmark model

The results can be found in Tables 1 and 2. Table 1 contains the relative RMSFEs of the original forecasts. A number of observations can be made from the table. First, a small number of both QuickNet and Autometrics recursive forecasts were explosive. This invalidates a straightforward RMSFE comparison between recursive and direct forecasts when the forecast horizon is five years but underlines the importance of checking the properties of the estimated ANN model before forecasting. It may be noted, however, that in forecasting two periods ahead, the recursive forecasts from the ANN-Autometrics model are clearly superior to their direct counterparts. Second, the recursive and the direct method applied to the AR(10) model yield very similar RMSFEs. This may appear strange at first, but there is an explanation. The model (28) generates very regular cyclical variation with the cycle length of about 11 years. When the AR model contains ten lags, the loss of accuracy due to the 'lag gap' in the direct model is nicely filled by observations from 'the previous cycle'. If the AR model had been an AR(1), say, the outcome would have been quite different. Finally, both ANN models yield more accurate direct forecasts than the linear model. At the one- and two-year horizons, Autometrics appears superior to QuickNet as a method of selecting the 'network architecture'.

Table 2 contains the results obtained by applying the insanity filter. In the ANN case, the recursive method is clearly superior to the direct method when the forecast horizon is sufficiently long. But then, the insanity filter substantially lowers the accuracy of the direct forecasts from all three models when the forecast horizon is five years. Inferior results obtained using our insanity filter are due to the fact that the 'no change' forecast is generally not a reasonable one in forecasting several periods ahead when the data-generating process generates cycles. Constructing a useful filter for such a situation remains an open problem.

Table 2: Ratios of the root mean square forecast error of the linear AR(10) model and the benchmark and two ANN models and the benchmark. Insanity filter was applied

Model	Horizon		
	1	2	5
True model (16)	(1.7915)	(2.7345)	(4.0849)
AR(10), recursive	1.3326	1.3353	1.9846
AR(10), direct	1.3326	1.3356	2.1596
QuickNet, recursive	1.1651	1.2383	1.5176
QuickNet, direct	1.1651	1.2466	1.9272
Autometrics, recursive	1.0247	0.9691	1.4510
Autometrics, direct	1.0247	1.1928	1.8994

Figures in parentheses are the RMSFEs from the benchmark model

9 Final remarks and further reading

In this article we consider forecasting with nonlinear parametric models. We present different models and various methods for obtaining multi-period forecasts recursively. We also briefly discuss the differences in accuracy between recursive forecasts and direct ones obtained by building a separate model for each forecast horizon. Our considerations are restricted to conditional mean forecasts.

It is often argued that the value of nonlinear models in forecasting lies in their ability to generate asymmetric forecast densities. Such densities may sometimes be more informative to decision makers than symmetric densities generated by a linear model. For space reasons, forecast densities have not been considered here, but some discussion about their usefulness in nonlinear models can be found in **?**.

For a general view of nonlinear time series models and their use in economics, the reader may consult **?** or **?**. They contain a large number of additional references. The latter volume should replace **?**. **?** offers a good exposition of nonlinear nonparametric models. **?** contains a comprehensive discussion on threshold autoregressive models. These two books are written for statisticians rather than econometricians. **?** has, among other things, a useful chapter on chaos. Nonlinear forecasting is also discussed in **?** and, more specifically, forecasting with STAR models in **?**. The literature on forecasting with neural network models includes **?**, **?** and a survey by **?**. **?** is a relevant survey of forecast combination, another topic not treated in this article.

Research Papers
2010

CREATES
Center for Research in Econometric
Analysis of Time Series

2009-47: Mark Podolskij and Mathias Vetter: Understanding limit theorems for semimartingales: a short survey

2009-48: Isabel Casas and Irene Gijbels: Unstable volatility functions: the break preserving local linear estimator

2009-49: Torben G. Andersen and Viktor Todorov: Realized Volatility and Multipower Variation

2009-50: Robinson Kruse, Michael Frömmel, Lukas Menkhoff and Philipp Sibbertsen: What do we know about real exchange rate non-linearities?

2009-51: Tue Gørgens, Christopher L. Skeels and Allan H. Würtz: Efficient Estimation of Non-Linear Dynamic Panel Data Models with Application to Smooth Transition Models

2009-52: Torben G. Andersen, Dobrislav Dobrev and Ernst Schaumburg: Jump-Robust Volatility Estimation using Nearest Neighbor Truncation

2009-53: Florian Heinen, Philipp Sibbertsen and Robinson Kruse: Forecasting long memory time series under a break in persistence

2009-54: Tue Gørgens and Allan Würtz: Testing a parametric function against a nonparametric alternative in IV and GMM settings

2009-55: Michael Jansson and Morten Ørregaard Nielsen: Nearly Efficient Likelihood Ratio Tests for Seasonal Unit Roots

2009-56: Valeri Voev: On the Economic Evaluation of Volatility Forecasts

2009-57: Jesper Rangvid, Maik Schmeling and Andreas Schrimpf: Global Asset Pricing: Is There a Role for Long-run Consumption Risk?

2009-58: Olaf Posch: Risk premia in general equilibrium

2009-59: Christian M. Dahl and Emma M. Iglesias: Modelling the Volatility-Return Trade-off when Volatility may be Nonstationary

2009-60: Ole E. Barndorff-Nielsen, José Manuel Corcuera and Mark Podolskij: Limit theorems for functionals of higher order differences of Brownian semi-stationary processes

2010-01 Anders Bredahl Kock and Timo Teräsvirta: Forecasting with nonlinear time series models