Nonlinear Econometric Modelling: A Selective Review

Norman R. Swanson and Philip Hans Franses*

Department of Economics, Pennsylvania State University, and
Department of Economics, Erasmus University, Rotterdam, respectively

June 1997

Abstract

In the discipline of economics, the vast array of economic theories which are available to economists often contain nonlinear elements (e.g. first order conditions, functional forms, etc.). In the past, however, many of the testable implications of economic theories, for example were formulated and derived with one of the main objectives being the specification of linear econometric models. This is perhaps not surprising for at least two reasons. First, up until quite recently, the statistical tools available for the examination of economic variables were better able to handle estimation and inference within a linear context. Second, and perhaps just as importantly, the computational ability of early computers (and even earlier, slide rules) was limited enough to make it infeasible to estimate either large or complex econometric models. However, as computers have become more efficient, so have the tools, algorithms, tests, and modelling strategies used by increasing numbers of practicing economists also become more complex. Indeed, one feature of the econometrics profession in recent years seems to be that we are always able to develop estimation and inferential procedures (and associated nonlinear econometric models) which are complex enough to fully tax the capability of even the most powerful computers. Of course, this is not a feature of economics alone, as physics, biology, chemistry, and other "harder" sciences have clearly done likewise. Furthermore, it should be noted that we are not merely developing new and more complex theories in order to take advantage of computational ability. Rather, we are taking advantage of the opportunity of "better" empirically modelling a system which we have always known to be complex and nonlinear. In this sense, the growth in nonlinear modelling in applied economics, for example, seems quite natural. In this chapter, we discuss two varieties of nonlinear models which may be of some interest to economists and econometricians. First we consider what are referred to as "stochastic unit root" (STUR) models. In these models, roots of AR processes are assumed to vary (according to some well defined stochastic process) with average or mean values of unity, rather than being fixed at unity. These models are not "near unit root" models, however, as the unit root in a simple AR(1) process, for example, is not assumed to approach unity as the sample size gets large. Rather, STUR models focus on a reasonable claim that very few economic series are precisely characterized as containing "pure" or constant unit roots, and attempt to examine whether some economic series can be better forecast using more general nonlinear models. Second, we discuss the class of nonlinear models called artificial neural networks (ANNs). These models are nonparametric in the sense that the number of parameters one fits to the data increases with the sample size. Estimation and modelling using ANNs is discussed, some findings are presented. For example, evidence concerning the relative forecasting ability of ANN, STUR, and a variety of simple linear models is presented. Finally, we discuss various empirical nonlinear modelling issues, with particular attention paid to modelling outliers, and the persistence of shocks.

Correspondence to: Norman R. Swanson, 521 Kern Graduate Bldg., Pennsylvania State University, University Park, PA., 16802. email: nswanson@psu.edu.

* Prepared for "Nonlinear Time Series Analysis of Economic and Financial Data". Some parts of the discussion in this chapter are summaries of more detailed discussions which can be found in: "An Introduction to Stochastic Unit Root Processes", "A Model Selection Approach to Assessing the Information in the Term Structure Using Linear Models and Artificial Neural Networks", "A Model Selection Approach to Real Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks", and "Further Developments in the Study of Cointegrated Variables" (see references). Swanson, thus, wishes to thank Clive W.J. Granger and Halbert White, who co-authored all of these papers. Swanson also wishes to thanks the Research and Graduate Studies Office at Pennsylvania State University for financial support during the writing of this article.

KEYWORDS doubly stochastic, stochastic unit root, cointegration, approximate maximum likelihood, artificial neural networks, forecast, mean squared error, nonlinear cointegration, outliers, persistence.

1. Introduction

In recent years nonlinear econometric modelling has received increasing attention in econometrics, and particularly, in applied econometrics. This is not to say, however, that nonlinear phenomena have not be considered an important aspect of economics for many decades. Rather, as statistical and econometric methodology has developed, our ability to formulate, and more importantly, estimate nonlinear models has increased. However, as with all sciences and social sciences, nonlinear theories and models have been around virtually from the beginning. Although many examples can be brought to bear in support of this notion, one need really simply consider business cycles, which have long been known to exhibit various seasonal and cyclical nonlinearites (for example, see Burns and Mitchell (1946)).

In the discipline of economics, the vast array of economic theories which are available to economists often contain nonlinear elements (e.g. first order conditions, functional forms, etc.). In the past, however, many of the testable implications of economic theories, for example were formulated and derived with one of the main objectives being the specification of linear econometric models. This is perhaps not surprising for at least two reasons. First, up until quite recently, the statistical tools available for the examination of economic variables were better able to handle estimation and inference within a linear context. Second, and perhaps just as importantly, the computational ability of early computers (and even earlier, slide rules) was limited enough to make it infeasible to estimate either large or complex econometric models. However, as computers have become more efficient, so have the tools, algorithms, tests, and modelling strategies used by increasing numbers of practicing economists also become more complex. Indeed, one feature of the econometrics profession in recent years seems to be that we are always able to develop estimation and inferential procedures (and associated nonlinear econometric models) which are complex enough to fully tax the capability of even the most powerful computers. Of course, this is not a feature of economics alone, as physics, biology, chemistry, and other "harder" sciences have clearly done likewise. Furthermore, it should be noted that we are not merely developing new and more complex theories in order to take advantage of computational ability. Rather, we are taking advantage of the opportunity of "better" empirically modelling a system which we have always known to be immensely complex. In this sense, the growth in nonlinear modelling in applied economics, for example, seems quite natural. Put another way, the development of nonlinear econometrics and applied economics is a feature of modern econometrics which is not at all undesirable, as it can easily be argued that we are simply attempting to better model an inherantly nonlinear world; and as computational capability increases, so naturally, does the complexity of our models of economic behavior.

One question that is often asked is: "Are we modelling economic behavior any better by using more complex constructs?" This question is difficult to answer decisively at this early juncture in the "nonlinear revolution". However, many of the references given in this chapter attempt to provide at least partial answers to this question by comparing the performance of alternative models and theories. This is often done, for example, by examining the predictive ability of linear and related nonlinear models, or by comparing the ability of linear models to "match" the data as well as more complex nonlinear models.

To fix some ideas concerning what is commonly meant by "nonlinearity", it is perhaps useful to consider the following simple model:

$$y_t \;=\; g(x_{t-i}\,,\, i \ge 1) + h(\varepsilon_{t-i}\,,\, i \ge 0)\ \ , i = 1,...,p,$$

If $g(\cdot)$ and $h(\cdot)$ are affine functions, then y_t follows a standard ARMA or autoregressive integrated moving average (ARIMA) model, depending on whether the process is second order stationary (assuming Gaussianity, for example) or not. If, on the other hand, $x_t = y_t$, and $h(\cdot) = \varepsilon_t$, then y_t follows a simple nonlinear autoregression of order p (NLAR(p)), where p refers to the highest order of lags of y_t appearing in $g(\cdot)$. This process is often referred to as "nonlinear in mean", for obvious reasons. Also, the process y_t is usually called "short-memory" if $\Phi_t(x{:}y_0)$ approaches $\Phi_h(x)$ as t approaches h, for h large, so that the starting value no longer affects the marginal distribution after the process has been running for a long time, where $\Phi_t(x{:}y_0)$ is the conditional distribution function of y_t given y_0 ($Prob(y_t \le x \mid y_0) = \Phi_t(x{:}y_0)$), and $\Phi_t(x)$ is the corresponding marginal distribution of y_t. This process is called "stable" if y_t and y_s have the same marginal distribution for t, s large.

On the other hand, we may have $x_t = y_t$, and $g(\cdot)$ an affine function, say of order p, then one might write

$$y_t \;=\; a_0 + a_1 y_{t-1} + \cdots + a_p y_{t-p} + \varepsilon_t .$$

If in addition, ε_t is distributed as $N(0,h_t)$, and

$$h_t = \alpha_0 + \alpha_1 h_{t-1} + \cdots + \alpha_1 h_{t-m} \ ,$$

the resulting model is an autoregressive conditional heteroskedasticity (ARCH) process, as examined by Engle (1982). Indeed, $g(\cdot)$ and $h(\cdot)$ can be further generalized to essentially allow for an infinitely wide class of nonlinear processes. Examples include threshold autoregressive models (Tong (1983, 1990)), bilinear models (Granger and Andersen (1978)), smooth transition autoregression (STAR) and related models (Luukkonen, Saikkonen, and Teräsvirta (1988a, b) and Granger and Ter a:svirta (1993)), Markov switching models (Hamilton (1989) and the references in Hamilton (1994), models related to chaos theory (Brock, Hsieh, and LeBaron (1991) and the references therein), random parameter models, and time varying parameter models, to name but a very few. A number of nonlinear models are also examined in the papers which comprise the special issue of the Journal of Applied Econometrics on Nonlinear Dynamics and Econometrics (1991), including for example Teräsvirta and Anderson (1991), Mizrach (1991), and Rothman (1991), where issues as diverse as constructing nonlinearity tests, modelling nonlinearities in U.S. macroeconomic variables, and forecasting using nearest neighbour techniques are examined. Indeed, it is not unreasonable to say that one can find an example of nonlinear modelling in economics by simply picking up a recent issue of virtually any economics journal. In the above papers and books, various recent nonlinearity tests are also examined and developed, including time reversibility tests, bispectrum tests, neural network nonlinearity tests (see also Lee, White, and Granger (1993)), STAR tests, BDS tests (Brock, Dechert, and Scheinkman (1996), and various others. As misspecification can be in the form of neglected nonlinearity, many other tests can in some cases be interpreted as tests of nonlinearity, including ARCH tests, White heteroskedasticity tests, and Ramsey (1969) RESET tests, to name but a few. Also, nonlinearity can appear in the form of asymmetrically adjusting error-correction models, and various other forms of "nonlinear cointegration" (see Balke and Fomby (1997), Granger and Swanson (1996), and Corradi, Swanson, and White (1997), for example). A very small subset of other recent papers which address nonlinearity in economic variables include: Diebold and Nason (1990), Granger (1993), Neftci (1984), Potter (1995), and Weiss (1996). Also, a particularly good starting point is Granger and Teräsvirta (1993).

In this chapter, we do not attempt to disseminate the vast accumulated knowledge in the area of nonlinear econometrics. For this, the reader is referred to the above works. Rather, we discuss two varieties of nonlinear models which may be of some interest to econometricians. In Section 2, we consider what are referred to as "stochastic unit root" (STUR) models. In these models, coefficients of AR processes are assumed to vary (according to some well defined stochastic process) with average or mean values of unity, and are not fixed to be unity (or to sum to unity in the case of an AR(p) model). These models are not "near unit root" models, however, as the unit root in a simple AR(1) process, for example, is not assumed to approach unity as the sample size gets large. Rather, STUR models focus on a reasonable claim that very few economic series are precisely characterized as containing "pure" or constant unit roots, and attempt to examine whether some economic series can be better forecast using more general nonlinear models. Also, STUR models may provide a good starting point for multivariate nonlinear VAR models. In Section 3, we discuss the class of nonlinear models called artificial neural networks (ANNs). These models are nonparametric in the sense that the number of parameters one fits to the data increases with the sample size. ANN models already have a rich history in the biological sciences, and were fist constructed in an effort to model the inner workings of the human brain. Estimation and modelling using ANNs is discussed, and some earlier findings are presented. In Section 4 we review some practical issues in empirical modelling, such as the effects of outliers, and model selection.

2. Stochastic Unit Root Processes

2.1 Theoretical Considerations

Consider a series x_t generated by

$$x_t = a_t x_{t-1} + \varepsilon_t \ , \tag{2.1}$$

where ε_t is zero mean, *i.i.d.* with variance σ_ε^2, and where

$$a_t = \exp(\alpha_t) \tag{2.2}$$

with α_t a Gaussian stationary series having mean m, variance σ_α^2 and power spectrum $g_\alpha(\omega)$. This is an example of a doubly stochastic process as considered by Tjøstheim (1986), and has been specifically analyzed by Brandt (1986), and Pourahmadi (1986, 1988). Model (2.1) can also be viewed as an extension of the periodic integration model (Franses (1996)) where α_t takes different values in different seasons. However, the exponential form (2.2) is a convenient case for consideration of the long-run properties of the series. A particular example that will be used to illustrate the results is the case where α_t is given by an AR(1) process

$$\alpha_t = \mu + \rho\, \alpha_{t-1} + \eta_t \tag{2.3}$$

where $|\rho| < 1$, and η_t is *i.i.d.* normally distributed $N(0,\sigma_\eta^2\,)$ and is independent of the series ε_t. It will be assumed throughout this section that α_t is generated exogenously from x_t, so that

$$g\ (\,\alpha_{t+1} \mid \alpha_{t-j}\,,\, x_{t-j}\,,\, j \geq 0\,) = g\ (\,\alpha_{t+1} \mid \alpha_{t-j}\,,\, j \geq 0\,)$$

where $g(\alpha \mid I)$ is a conditional distribution. For the example (2.3), $m = \mu/(1-\rho)$ and $\sigma_\alpha^2 = \sigma_\eta^2/(1-\rho^2)$. (2.1) can be solved as

$$x_t = \varepsilon_t + a_t\, \varepsilon_{t-1} + a_t a_{t-1}\, \varepsilon_{t-2} + \ldots + a_t a_{t-1} \cdot\cdot\cdot\cdot a_{t-k+2}\, \varepsilon_{t-k+1} + a_t a_{t-1} \cdot\cdot\cdot\cdot a_{t-k+1} x_{t-k}$$

or,

$$x_t = \varepsilon_t + \pi_{t,1}\, \varepsilon_{t-1} + \pi_{t,2}\, \varepsilon_{t-2} + \ldots + \pi_{t,k-1}\, \varepsilon_{t-k+1} + \pi_{t,k} x_{t-k} \tag{2.4}$$

for any integer k, $0 \leq k \leq t$, where,

$$\pi_{t,j} = \exp(S_{\alpha t}(j))\ ,$$

and,

$$S_{\alpha t}(j) = \sum_{i=0}^{j-1} \alpha_{t-i}$$

with the notation $S_{\alpha t}(0) = 1$. It should be noted that for j>0

$$E_j \equiv E\,[S_{\alpha t}(j)] = jm\ \ , \tag{2.5}$$

and,

$$V_j = var[S_{\alpha t}(j)] = \sigma_\alpha^2[j + \sum_{r=1}^{j-1}(j-r)\rho_\alpha(r)] \tag{2.6}$$

where $\rho_\alpha(r) = corr(\alpha_t\,,\, \alpha_{t-r})$. For j large (2.6) can be approximated as

$$V_j \approx j f_\alpha(0) \tag{2.7}$$

as shown in Koopmans (1974, page 171) where $f_\alpha(w) = \dfrac{1}{2\pi} g_\alpha(w)$.

A standard result that is useful in this analysis is that if a random variable X is distributed $N(m\ ,\ \sigma^2)$ then

$$E\,[\exp(kX)] = \exp(km + \tfrac{1}{2}k^2\sigma^2). \tag{2.8}$$

It follows that for k large,

$$E[\pi_{t,k}] = \exp(E_k + \tfrac{1}{2}V_k) \approx \exp(k\theta)\ \ , \tag{2.9}$$

where,

$$\theta = m + \tfrac{1}{2}f_\alpha(0). \tag{2.10}$$

Assuming for convenience that x_0 is deterministic, and taking expectations of (2.4) with $k = t$ gives

$$E[x_t] = x_0 \exp(t\theta)$$

and, as the ε_t are *i.i.d.*,

$$var(x_t) \;=\; \sigma_\varepsilon^2 \sum_{j=0}^{t-1} E[\pi_{t,j}^2] \;=\; \sigma_\varepsilon^2 \sum_{j=0}^{t-1} \exp(2jm + 2V_j) \;\approx\; \sigma_\varepsilon^2 \sum_{j=0}^{t-1} \exp(2j\phi) \;, \tag{2.11}$$

where, $\phi = m + f_\alpha(0)$.[1] Thus,

$$\left.\begin{array}{lll} var(x_t) & = \sigma_\varepsilon^2 \left[\dfrac{1 - \exp(t2\phi)}{1 - \exp(2\phi)} \right] & \text{if } \phi \neq 0 \\ & = t\,\sigma_\varepsilon^2 & \text{if } \phi = 0 \end{array}\right\} \tag{2.12}$$

Turning to consideration of (linear) regression coefficients $\beta_{t,k}$ as in:

$$x_t \;=\; \beta_{t,k} x_{t-k} + \text{error}$$

it can be noted from (2.4) that

$$E_L(x_t \mid x_{t-j} \,, j \geq k) \;=\; x_{t-k} E_L[\pi_{t,k} \mid x_{t-j} \,, j > k] \;=\; x_{t-k} E[\pi_{t,k}] \tag{2.13}$$

where $E_L(\cdot)$ is an expectation using only the components of the conditioning sets in a linear form and $E(\cdot)$ is an unconditional expectation. Thus, from (2.9),

$$\beta_{t,k} \;=\; E[\pi_{t,k}] \;=\; \exp(k\theta) \,, \;\; \text{for large } k \tag{2.14}$$

Using the identity $corr(x_t \,, x_{t-k}) \;=\; \beta_{t,k} \sqrt{\dfrac{var(x_{t-k})}{var(x_t)}}$ thus gives the following approximate expressions for this autocorrelation when t and k are both large:

	$\theta < 0$	$\theta = 0$	$\theta > 0$
$\phi > 0$	$\exp(-kf)$	$\exp(-kf)$	$\exp(-kf)$
$\phi = 0$	$\exp(k\theta)(1 - k/t)^{1/2}$	1	-
$\phi < 0$	$\exp(k\theta)$	-	-

where $\theta = m + \frac{1}{2} f$, $\phi = \theta + \frac{1}{2} f$,and $f \equiv f_\alpha(0) = \dfrac{1}{2\pi} g_\alpha(0)$ where $g_\alpha(0)$ is the spectrum of α_t at zero frequency. Empty cells are cases that cannot occur, since by definition, $\phi \geq \theta$.

The autocorrelations decline exponentially (approximately) in all cases except when $\theta = \phi = 0$, so that $f_\alpha(0) = 0$ and $m = 0$. This special case occurs, for example, if $\alpha_t = 0$, all t, since it then follows that $E[e^{\alpha_t}] = 1$, and x_t is a perfect unit root process.

Two alternative characterizations of the STUR process given by (2.1) and (2.2) are

(i) STURA: α_t is such that $E[e^{\alpha_t}] = 1$ so that

[1] This is an approximation of the true variance of x_t, since $E[\pi_{t,j}^2] \neq \exp(2j\phi)$ for small j.

$$m + \tfrac{1}{2}\sigma_\alpha^2 = 0 \qquad (2.15)$$

(ii) STURB: α_t is such that $E[\pi_{t,k}] = 1$, k large, where $\pi_{t,k}$ is given in (2.4), so that

$$\theta = 0 \quad \text{or} \quad m + \tfrac{1}{2} f_\alpha(0) = 0 \qquad (2.16)$$

The persistence which arises when the restrictions given by (2.15) or (2.16) are met is similar to the persistence exhibited by perfect unit root processes. For instance, as in the case of a random walk, shocks from the distant past continue to have an impact on current values of the process.

More precisely, the properties of STURA depend on the relative size of σ_α^2 and $f_\alpha(0)$. Suppose that $f_\alpha(0) > \sigma_\alpha^2$, so that, essentially, α_t is smoother than white noise. It follows that θ and ϕ are both positive, so that from (2.12), the variance of x_t is increasing explosively, although probably mildly so, and $E[\pi_{t,k}]$ is increasing with k. Thus, distant shocks have greater impact on the current value of the series than do recent shocks. This property also holds for linear I(d), d>1, processes, for example, and might be called "super-persistence".

A STURB process will have $E[e^{\alpha_t}] < 1$ if $f_\alpha(0) > \sigma_\alpha^2$, so that a regression of

$$x_t = \beta_{t,k} x_{t-k} + error$$

will yield $\beta_{t,k}$ values which are nearer to unity for large values of k than for small values of k. The persistence property is now particularly interesting. Recall that

$$x_t = \sum_{j=0}^{k-1} \pi_{t,j} \varepsilon_{t-j} + \pi_{t,k} x_0$$

and

$$\pi_{t,j} = \exp[S_{\alpha t}(j)], \quad j \geq 1 \quad \text{and} \quad \pi_{t,0} = 1$$

where, $S_{\alpha t}(j)$ has mean jm, variance $jf_\alpha(0)$, for j large. If α_t is normally distributed, then $\pi_{t,j}$ will have a lognormal distribution. With the constraint that $E[\pi_{t,j}] = 1$, which implies that m is negative, standard results then give

$variance(\pi_{t,j}) = \exp(jf_\alpha(0)) - 1$

$mode(\pi_{t,j}) = \exp(-3/2\; jf_\alpha(0))$

$median(\pi_{t,j}) = \exp(-1/2\; jf_\alpha(0))$

Also, a 95% confidence interval (about the median) is

$$\exp[-\tfrac{1}{2} jf_\alpha(0) \pm 2(jf_\alpha(0))^{1/2}]$$

Thus, as j increases, $\pi_{t,j}$ has a constant mean, an increasing variance, a mode and median which tend towards zero, and a $100(1-\delta)\%$ confidence interval that tends towards very small values. This implies that the form of persistence which is displayed is very fragile.

From a theoretical perspective, stochastic unit roots may be seen to arise quite naturally in economics. As an example, it is well known that many basic formulations of the permanent income hypothesis involve the term $\frac{1+\delta}{1+R_t}$ (see Hall (1978), for example). Using the standard approach, δ is the marginal rate of time preference, and R_t is the real rate of interest. A frequently made assumption is that this term is identically unity. However, when that assumption is relaxed, and R_t is allowed to vary stochastically, then the term may well vary around unity, and probably has a mean very close to unity, as expected in the STUR case.

To summarize, the above results indicate that the properties of STUR processes are often markedly different from comparable properties of perfect unit root processes. For example, some properties mimic those of the ARCH model in (1.3). Another characteristic of stochastic unit roots is that they are quite difficult to distinguish from perfect unit roots. This is not surprising given that evidence presented below indicates that variances of stochastic unit roots are often quite small. In this sense the usual power failures associated with unit root tests should apply. Nevertheless, before discussing an appropriate testing strategy for the presence of STUR, we provide an example of a STUR process which is readily mistaken for a pure unit root process based on a standard Dickey-Fuller test.

2.2 Empirical Considerations

In this section, we discuss estimation of and testing for the presence of stochastic unit root processes. First, one method for estimating the parameters of STUR processes is summarized. For further details, and discussion of alternative estimation strategies, the reader is referred to Granger and Swanson (1997). The method which we discuss below is closely related to the approximate maximum likelihood method discussed in Guyton, Zhang, and Foutz (1986). In contrast to Guyton et al., though, conditions are not found for the existence of a *stationary* generalized autoregressive process. Rather, the parameters of a nonstationary STUR process are estimated. Assume that

$$x_t \;=\; a_t x_{t-1} + \varepsilon_t \;, \quad a_t \;=\; \exp(\alpha_t) \quad \text{and} \quad E(a_t) \;=\; 1 \quad \text{or} \quad \theta \;=\; 0 \;, \tag{2.17}$$

where

$$\alpha_t \;=\; \mu + \sum_{i=1}^{p} \rho_i \alpha_{t-i} + \eta_t \;,$$

and α_t is a stationary stochastic process, thus generalizing (2.3) to be an AR(p) process. (The STUR model may also be extended in the current context by adding lags of x_t, for example.) Of course, for the sake of estimation, we will not assume that $E(a_t) = 1$ or that $\theta = 0$, as we do not wish to impose that the process must be STURA or STURB according to (2.15) or (2.16). Another possible variation of the above STUR process is the addition of a time varying intercept, say c_t, to the measurement equation. In this way, the upward trending behavior of many economic series could, potentially, be better modeled. As we will see below, such generalizations pose no problem for the approximate maximum likelihood (AML) estimation technique.

Assume that the ε_t and the η_t are normally distributed with zero means, and variances σ_ε^2 and σ_η^2, respectively, and are independent. Thus, a_t is distributed as a lognormal random variable. The unknown parameters are gathered into a vector, $\Omega = (\mu, \rho, \sigma_\varepsilon^2, \sigma_\eta^2)$, where $\rho = (\rho_1, \rho_2, ..., \rho_p)$. In order to construct a likelihood function, use the well known property that $f(x, a \mid \Omega) = f(x \mid a, \Omega) f(a \mid \Omega)$, where $f(\cdot)$ is a joint density. Then, noting that likelihood is proportional to probability, assuming a value of unity for the constant of proportionality, carrying the conditioning argument one step further, and introducing a randomization parameter, K, yields

$$L(x \mid \Omega) \;=\; \int L(x,\, a \mid \Omega) da \;=\; \int L(x \mid a,\, \Omega) f(a \mid \Omega) da \;=\; E_{a|\Omega} L(x \mid a,\, \Omega)$$

and,

$$\hat{E}_{a|\Omega} L(x \mid a,\, \Omega) \;=\; \hat{L}(x \mid \Omega) \;=\; \frac{1}{K} \sum_{k=1}^{K} \left[\frac{1}{\sqrt{2\pi}} \right]^T \sigma_\varepsilon^{-T} e^{-\frac{1}{2\sigma_\varepsilon^2} \sum_{t=1}^{T} \hat{\varepsilon}_{t,k}^2} \tag{2.18}$$

$$\hat{\varepsilon}_{t,k} \;=\; x_t - \hat{a}_{k,t} x_{t-1} \tag{2.19}$$

where by $E_{a|\Omega}$ we mean the expectation over a conditional on the parameter Ω, and assuming that the density of x_1 is the same as the conditional density of x_t, given x_{t-1}, and that $x_0 = 0$. This is a standard result, except that the a_t are unobserved so that we construct an estimate, say $\hat{L}(\cdot)$, by randomly sampling across different distributions of a_t.

In order to calculate values of the likelihood function for various points in the parameter space, various sequences, $a_k = \{a_{k,t}\}$, are generated for each $\{\mu, \rho_{i,}\ i=1,...,p, \sigma_\eta^2\}$ in the parameter space. Thus, the approximate maximum likelihood function given as (2.18) is a random variable, and is an estimator of $E_{a\,|\Omega}L(x \mid a, \Omega)$. Maximization of (2.18) proceeds by first generating K sequences $\{\eta_{t,k}\}$, t=1,...,T. Then, $\hat{L}(\cdot \mid \cdot)$ and its derivatives are calculated for each point in the parameter space $\hat{\Omega} = (\hat{\mu}, \hat{\rho}_i\ , i=1,...,p\ , \hat{\sigma}_\varepsilon^2\ , \hat{\sigma}_\eta^2)$. This iterative procedure is carried out until a $\hat{\Omega}$ is found which optimizes $\hat{L}(\cdot \mid \cdot)$, according to some convergence criterion. Following accepted techniques from the simulation-based estimation literature, the *same* K sequences, $\{\eta_{t,k}\}$, t=1,...,T, are used throughout the optimization. Approximate maximum likelihood methods are also considered in the disequilibrium literature. For example, Laroque and Salanie (1994) show that simulated pseudo ML estimators have good finite sample properties, even for values of K as low as 15. It should be noted that σ_ε^2 cannot be concentrated out of the approximate likelihood function (since the sum of logs is not equal to the log of sums).

The finite sample properties of the AML method were examined using a series of Monte Carlo experiments. STUR processes were generated according to (2.17) with p=1, $\rho_1 = 0.6$, $E(a_t) = 1$, $\eta_t \sim$ *i.i.d.* N(0, 0.0001) and N(0, 0.000001), and $\varepsilon_t \sim$ *i.i.d.* N(0, 1.0). One thousand trials were carried out, for sample sizes of 100, 250, and 500 observations, and for randomization parameter, K, values of 10, 25, 100, and 250. Parameters were estimated using the GAUSS computer program and the Broyden, Fletcher, Goldfarb, Shanno (BFGS) algorithm. The results of this simulation exercise are found in Granger and Swanson (1997), and their findings can be summarized as follows. First, the AML method appears to produce fairly reasonable average point estimates of the STUR parameters, in particular for the larger sample sizes of 250 and 500, and for the larger K values of 100 and 250. For example, when K=10 and 25 the average point estimates of σ_η^2 are not as close to their true values as when higher randomization parameters are used. Interestingly, selecting K=250 does not generally result in substantial improvement over the K=100 case. Second, σ_ε^2 is estimated very well, even for small K and T. Finally, it should be noted that the parameters are often estimated fairly imprecisely, particularly when smaller values of σ_η^2 are used. This suggests that AML estimates should be used with caution, in practical applications.

Turning now to the issue of testing, it is first worth pointing out that ADF tests have low power against some STUR alternatives, since the test is well known to exhibit limited power against a variety of "near" I(1) processes. In the STUR case it would be ideal if one were to construct a test which has STUR as the null hypothesis. This is difficult, though, given the rather complex nature of STUR processes, and is not considered here. The approach adopted here has difference stationarity as the null hypothesis, and a random coefficient AR(p+1) model as the alternative. Leybourne, McCabe, and Tremayne (1994) develop the LM type test which we will discuss. Interestingly, the test is also applicable to STUR processes, and is robust to various departures under the null such as ARCH and GARCH. Of note is that the test starts by assuming that a series is non-stationary. In this way, the test has the null of a fixed unit root (or a linear I(1) process) and the alternative of some more general non-stationary data generating mechanism.

The model treated by Leybourne et al. has

$$x_t = \sum_{i=1}^{q+1} \delta_{i,t} x_{t-i} + \varepsilon_t \tag{2.20}$$

$$\delta_{1,t} = (a_t + \pi_1)\ ,\ \delta_{i,t} = (\pi_i - a_t\pi_{i-1})\ , i=2,\ldots,q\ ,\ \delta_{q+1,t} = -a_t\pi_q\ ,$$

where $a_t \sim$ i.i.d. $(1, \omega^2)$ and $\varepsilon_t \sim$ i.i.d. $(1, \sigma_\varepsilon^2)$, and q is a non-negative integer. Trends may also be incorporated into (2.20), and some dependence is allowed between the a_t and ε_t. Also, note that under the null hypothesis that $\omega^2 = 0$, x_t is seen to be an AR(q+1) process with a unit root, since all of the coefficients sum to unity. The version of (2.20) which we will consider has q=1 and π_1=1:

$$x_t = a_t x_{t-1} + \varepsilon_t\ , \tag{2.21}$$

and is examined by Leybourne, McCabe and Tremayne (1994) and McCabe and Tremayne (1995). In this framework, consider: H_0: $\omega^2 = 0$ against H_1: $\omega^2 > 0$. Before writing the statistic, three features of the test should be mentioned. First, Leybourne et al. show that the test statistic below remains unchanged when the natural log of a_t in (2.21) is allowed to follow a stationary AR(1) process, as in (2.1), (2.2) and (2.3). Second, randomized coefficient models such as (2.20) and (2.21) can be written as models of conditional variation. However, the lagged level of the process itself drives the conditional heteroskedasticity, while in many G(ARCH) formulations it is the innovations that drive the conditional variation.

Leybourne et al. have shown that assuming the ε_t are stationary is sufficient to ensure that the test has no power against G(ARCH). They have also shown that the empirical size of the test in the G(ARCH) case is less than the nominal value so that the test is conservative. Unfortunately, this result does not hold when the form of heteroskedasticity is IGARCH. These and other related issues will be left to future work. Third, it should also be noted that the test is augmented in the same way as ADF tests to allow for various forms of serial correlation in the differences of x_t.

Assuming normality, using the standard conditioning argument, and substituting $\omega^2 = 0$ in the first order condition, yields

$$\partial L(\cdot)/\partial\omega^2 \,|_{\omega^2=0} \;=\; \sigma_\varepsilon^{-4} \sum_{t=q+2}^{T} \Big(\sum_{j=1}^{t-1} \varepsilon_j\Big)^2 (\varepsilon_t^2 - \sigma_\varepsilon^2) + o_p(1)$$

The statistic is

$$\hat{Z}_T \;=\; T^{\frac{-3}{2}}\, \hat{\sigma}_\varepsilon^{-2} \hat{\kappa}^{-1} \sum_{t=q+3}^{T} \Big(\sum_{j=q+2}^{t-1} \hat{\varepsilon}_j\Big)^2 (\hat{\varepsilon}_t^2 - \hat{\sigma}_\varepsilon^2) \;,$$

where, $\hat{\varepsilon}_t$ is the residual from a regression of Δx_t on a constant, a trend (and lags of Δx_t), with the number of lags, q chosen accordingly (i.e. so that the ε_t are not autocorrelated), $\hat{\sigma}_\varepsilon^2 = \frac{1}{T}\sum \hat{\varepsilon}_t^2$ and $\hat{\kappa}^2 = \frac{1}{T}\sum(\hat{\varepsilon}_t^2 - \hat{\sigma}_\varepsilon^2)^2$. Leybourne et al. show that the size and power of the statistic is not significantly affected by overfitting. Thus, they suggest that to err on the side of too many lags is better than under-specifying the number of lags. $\hat{Z}_T$ converges weakly to a sum of functionals of generalized Brownian bridges and Brownian motions, given by

$$\int_0^1 G_1(r)^2 dG_2(r) - \int_0^1 G_1(s)^2 ds G_2(1) \;,$$

where $G_1(r) = W_1(r) - rW_1(1) + 6r(1-r)[W_1(1)/2 - \int_0^1 W_1(s)ds]$, $G_2(r) = \psi W_1 + (1-\psi^2)^{1/2} W_2(r)$, the limit processes W_1 and W_2 are independent Brownian motions, G_1 is a generalized Brownian bridge, and the nuisance parameter ψ is the covariance between ε_t and ε_t^2. The critical values for this test statistic have been simulated under the assumption that ε_t is symmetrically distributed so that ψ=0 (see Leybourne et al. (1994)). However, the distribution could be tabulated for different values of ψ, and an estimated value used to calibrate the critical value. Interestingly, if $\psi = 0$, $\hat{Z}_T$ can be modified so that it converges to a random variable with a standard normal distribution. In particular, note that the asymptotic distribution given above can be written as

$$\int_0^1 G_1(r)^2 dG_2^*(r) = \int_0^1 J(r) dG_2(r),$$

where

$$G_2^*(r) = G_2(r) - rG(1)$$

and

$$J(r) = G_1(r)^2 - \int_0^1 G_1(s)^2 ds \ .$$

Thus, $\hat{Z}_T$ is characterized by a mixture of normals distribution, $N(0, \int_0^1 J(r)^2 dr)$. This suggests that an alternative test statistic is:

$$\hat{Z}_T^* = [\sum_{t=p+3}^{T} \hat{S}_{t-1}^2 (\hat{nu}_t^2 - \hat{\sigma}_\nu^2)]/[\hat{\kappa}(\sum_{t=p+3}^{T} \hat{S}_{t-1}^4 - (\sum_{t=p+3}^{T} \hat{S}_{t-1}^2)^2)]^{1/2}$$

$$= [\sum_{t=p+3}^{T} \hat{J}_{t-1}^* \hat{\nu}_t^2]/[\hat{\kappa}(\sum_{t=p+3}^{T} \hat{J}_{t-1}^{*2})^{1/2}] \ ,$$

where $\hat{S}_t = \sum_{j=p+2}^{t} \hat{\nu}_j$, $\hat{J}_t = \hat{S}_t^2$, and $\hat{J}_t^* = \hat{J}_t - T^{-1} \sum_{j=1}^{T} \hat{J}_j$. Then it follows that $\hat{Z}_T^*$ converges weakly to

$$[\int_0^1 J(r) dG_2(r)]/[(\int_0^1 J(r)^2 dr)^{1/2}]$$

which is the N(0,1) distribution in the special case where $\psi = 0$. This result suggests that the above modified STUR test is very easy to apply, and does not require the simulation of critical values for a non-standard distribution, when ν_t is known to have a symmetric distribution.

Based on Monte Carlo simulations of the STUR model specified by (2.1), (2.2) and (2.3), where the alternative hypothesis considered is (2.20), but with $a_t = exp(\alpha_t)$ and

$$\alpha_t = \mu + \rho \, \alpha_{t-1} + \eta_t$$

where $|\rho| < 1$, $\eta_t \sim i.i.d. \ N(0, \sigma_\eta^2)$ and is independent of the series ε_t, and $E(a_t) = 1$, the following findings were reported in Granger and Swanson (1997). Since the test has difference stationarity as the null, it may have power against STUR, as STUR processes are clearly not difference stationary, in theory. As might be expected $\hat{Z}_T$ has very good power against the STUR alternative for larger values of σ_η^2 and for larger sample sizes (i.e. σ_η^2= 0.1 - 0.001, T=500, and 1000). As σ_η^2 falls, the power of the test tails off significantly. Also, overall, the evidence from a number of other ADF test simulations suggests that the difference-stationarity test has more power against STUR alternatives than standard ADF tests. For example, for a sample size of 250 observations, the power of the $\hat{Z}_T$ test is fairly low (0.152 for nominal size of 5%), but is still around three times as high as the power of the ADF test. In summary, while the above test is only a starting point, it nevertheless serves as a good indicator of whether (macro)economic time series are well modelled as containing stochastic unit roots.

Using the above test as a guide, Granger and Swanson (1997) find that approximately one half of a randomly selected group of macroeconomic variables reject the null of difference stationarity in favor of the alternative (STUR). Also, a forecasting experiment based on 1- and 4-step ahead forecasts suggests that the mean-square forecast error (MSFE) "best" model is STUR just as often as it is a Kalman filter based model, a simple linear autoregressive model (with lag structure chosen so as to *whiten* the residuals of the estimated model), or a random walk model. This finding applies particularly to various interest rate and price series. On the other hand, linear models are preferred when forecasting unemployment.

3. Artificial Neural Network Models

3.1 Methodology

Cognitive scientists have proposed a class of flexible nonlinear models inspired by certain features of the way that the brain processes information. (A good introduction to the cognitive science literature is Rumelhart and McClelland (1986).) Because of their biological inspiration, these models are referred to as "artificial neural network models" or simply "artificial neural networks" (ANNs). Because of their flexibility and simplicity, and because of demonstrated successes in a variety of empirical applications where linear models fail to perform well (see White (1989) and Kuan and White (1994) for some specifics), ANNs have become the focus of considerable attention as a possible vehicle for forecasting financial variables as well as for pattern recognition. Among recent applications are those of White (1988), Dutta and Shekhar (1988), Moody and Utans (1991), Dorsey, Johnson and van Boening (1991), Dropsy (1992), Franses and Draisma (1997), Franses and Van Homelen (1997), Kuan and Liu (1992), and Ripley (1994). See also the recent book by Bishop (1995) and Trippi and Turbau (1993).

For those interested in a detailed discussion of ANNs and their econometric applicability, we refer to Kuan and White (1994). For this chapter, it suffices to treat these models as a potentially interesting black box, delivering a specific class of nonlinear regression models. In particular, the ANN nonlinear regression models considered here have the form:

$$f(x,\,\theta) = \tilde{x}'\alpha + \sum_{j=1}^{q} G(\tilde{x}'\,\gamma_j)\,\beta_j \tag{3.1}$$

where $\tilde{x}$ is a (column) vector of explanatory variables, $\tilde{x} = (1,\, x')'$ augments x by the inclusion of a constant term, $\theta = (\alpha', \beta', \gamma')'$, $\beta = (\beta_1, \ldots, \beta_q)'$, $\gamma = (\gamma', \ldots, \gamma_q')'$, q is a given integer and G is a given nonlinear function, in our case, the logistic cumulative distribution function (c.d.f.) $G(z) = 1/(1 + \exp(-z))$.

A network interpretation of (3.1) is as follows. "Input units" send signals $x_0(=1), x_1, \ldots, x_r$ over "connections" that amplify or attenuate the signals by a factor ("weight") γ_{ji}, $i=0, \ldots, r$, $j=1, \ldots, q$. The signals arriving at "intermediate" or "hidden" units are first summed (resulting in $\tilde{x}'\gamma_j$) and then converted to a "hidden unit activation" $G(\tilde{x}'\gamma_j)$ by the operation of the "hidden unit activation function" G. The next layer operates similarly, with hidden activations sent over connections to the "output unit." As before, signals are attenuated or amplified by weights β_j and summed. In addition, signals are sent directly from input to output over connections with weights α. A nonlinear activation transformation at the output is also possible, but we avoid it here for simplicity. Of note is that two or more hidden layers, as well as alternative activation functions are also frequently used (for example, see Chen and Swanson (1997)).

In network terminology, $f(x,\,\theta)$ is the "network output activation" of a "hidden layer feedforward network" with "inputs" x and "network weights" θ. The parameters γ_j are called "input to hidden unit weights," while the parameters β_j are called "hidden to output unit weights." The parameters α are called "input to output unit weights."

Hornik, Stinchcombe and White (1989, 1990) (among others, see also Cybenko 1989, Carroll and Dickinson 1989 and Funahashi 1989) have shown that functions of the form (3.1) are capable of approximating arbitrary functions of x arbitrarily well given q sufficiently large and a suitable choice of θ. This "universal approximation" property is one reason for the successful application of ANNs. In fact, White

(1990) establishes that ANN models can be used to perform nonparametric regression, consistently estimating any unknown square integrable conditional expectation function.

In Swanson and White (1995), model (3.1) is applied to the problem of forecasting $R_{t+\tau} - R_{t+1}$ or $R_{t+\tau} - R_{t+\tau-1}$, lags of these variables, as well as forward rates observed at time *t*, where $R_{t+\tau}$, the 1-month spot interest rate observed at time $t + \tau-1$; and $F_{\tau,t}$, the forward rate for month $t+\tau$ observed at time *t*, and all data are end of month U.S. Treasury bill rates (see e.g. Mishkin (1988)). The basic idea behind this approach is that the inclusion of the nonlinear terms $G(x' \gamma_j)$ should enhance forecasting ability relative to standard linear models, if overfitting is properly avoided and the data truly have nonlinear features. In their analysis, a model selection approach is taken to fitting the artificial neural network models, thus shedding light not only on the usefulness of the current forward rate for predicting the future spot rate, but also on the usefulness of the Schwarz Information Criterion (SIC) for selecting "best" forecasting models ("best" in their paper refers to the best model in the context of confusion matrices, MSFEs, and various profit measures - see Swanson and White (1995, 1997) for further details of these types of "model selection" criteria). Their findings can be summarized as follows: First, for a relatively long out-of-sample period, the "best" model contains forward rates in four out of five forecast horizons, based on a forecast mean squared error performance measure, and five out of five horizons, when chosen using a confusion performance measure. The best models from a trading profitability standpoint contain the forward rate for three of the five horizons. Forward rates are thus useful in predicting future changes in spot rates, to this extent. Second, windows of observations (used in the estimation of the models) of less than maximal size occasionally appear as forecast mean squared error-optimal, generally appear as confusion-optimal and often appear as profit-best, suggesting instability in the relationships of interest. Third, the in-sample SIC does not appear to be a reliable guide to out-of-sample performance, so it fails to offer a convenient shortcut to true out-of-sample performance measures for selecting models, and for configuring nonlinear artificial neural network models. Finally, artificial neural network (ANN) models appear to be promising for use in this forecasting context, particularly as they often lead to the "least-confused" models, and further refinement and application of ANN methods is warranted.

In a closely related analysis, Swanson and White (1997) use a model selection approach to compare real-time forecasts (see Swanson (1996)) from 9 macroeconomic variables using various adaptive and nonadaptive models, linear and potentially nonlinear (artificial neural network) models, and professional constructed forecasts (from the Survey of Professional Forecasters (SPF), see Croushore (1993)). As above, they rely on various model selection criteria, for which summaries and discussions can be found in Diebold and Mariano (1995), Engle and Brown (1986), Fair and Shiller (1990), Henriksson and Merton (1981), Keane and Runkle (1990), Leitch and Tanner (1991), Pesaran and Timmerman (1994a,b), Stekler (1991, 1994), and Zarnowitz and Braun (1992). Their conclusions can be summarized as follows: First, even when their econometric models are constrained to include information available only on a real-time basis, econometric predictions still outperform SPF predictions for many of the variables, based on mean squared forecast error and mean absolute deviation measures. In particular, adaptive multivariate linear (and to a lesser extent nonlinear artificial neural network) models tend to outperform SPF, no change, and nonadaptive univariate and multivariate linear models. It should be noted, though, that the SPF forecasts appear to perform more or less as well as a number of econometric models when comparing predictions of the direction of change in a variable, and when minimizing the mean absolute percentage error model selection criterion. Nevertheless, based on predictions of the direction of change, the multivariate adaptive linear and nonlinear artificial neural network models (when grouped together) dominate all other (linear) models combined, providing the least confused forecasts for the majority of variables examined, for forecast horizons of both 1 quarter and 1 year. Overall, our results, which include Diebold-Mariano loss differential tests, χ^2 tests of independence, and sign tests, indicate that model selection should proceed on a case by case basis, with adaptive, nonadaptive, and SPF forecasts alternately dominating depending on which variable is being examined. Second, windows of observations less than the maximal size rarely appear in prediction-"best" models, suggesting relative stability in the relationships of interest. Third, they again find that the in-sample SIC does not appear to offer a convenient shortcut to true out-of-sample performance measures for selecting models, or for configuring adaptive neural network models, when forecasting macroeconomic variables. Fourth, the use of unrevised data in real-time forecasting appears to offer a valid guide for comparing real-time professionally available forecasts with econometric predictions. This is contrary to the common practice of building econometric models using the latest fully revised data, which is a mixture of unrevised, partially revised, and fully revised data, and has the feature that future data may have been used (perhaps inadvertantly) to revise earlier data (such as when revised seasonal factors and revised benchmark figures are used). Finally, multivariate adaptive models appear to be promising for use in this context, although it should perhaps be noted that little evidence that explicitly supports the exclusive use of artificial neural network models is found. Nevertheless, it is suggested that further refinement and application of adaptive linear and nonlinear methods for modeling macroeconomic variables appears warranted, particularly in the context of truly *ex ante* or real-time forecasts.

In addition to being useful for modelling conditional mean, ANN models can also be used to help in the examination of conditional variance. For example, Franses and Van Homelen (1997) apply ANNs to forecasting exchange rates. Such financial data often have GARCH (generalized autoregressive heteroskedasticity) features, and it is well known that GARCH (if not corrected for) can lead to the suggestion (based on standard nonlinearity tests) that the data should be modelled nonlinearly in mean. Franses and Van Homelen (1997) show using Monte Carlo analysis, that when data exhibit nonlinearity only in the form of undetected GARCH, then ANNs do not lead to improved forecasts, relative to simpler linear models. However, when data are nonlinear in mean, ANNs will exploit this, yielding improved forecasts. Franses and Van Homelen (1997) then consider using ANNs to model daily exchange rate data, and find that ANNs do not lead to improved forecasts relative to linear models, when the models are compared using tests of independence in the context of 2x2 contingency tables which measure the "directional forecast accuracy" (see Pesaran and Timmerman (1992), and above). Finally, as ANNs appear not to be fooled by GARCH, Franses and Van Homelen (1997) recommend the evaluation of out-of-sample forecasting using ANNs as a diagnostic for checking for nonlinearity.

3.2 Estimation Considerations

In estimating ANN models of the form discussed above, it is inappropriate to simply fit the network parameters with $q = 4$ hidden units by least squares, as the resulting network typically will have more parameters than observations, achieving a perfect or nearly perfect fit in sample, with disastrous performance out-of-sample. To enforce a parsimonious fit, the ANN models were estimated by a process of forward stepwise (nonlinear) least squares regression, using the SIC to determine included regressors and the appropriate value for q. Specifically, a forward stepwise linear regression is performed first, with regressors added one at a time until no additional regressor can be added to improve the SIC. The linear regression coefficients are thereafter fixed. Next a single hidden unit is added (i.e. q is set to 1), and regressors are selected one by one for *connection* to the first hidden unit, until the SIC can no longer be improved. Then a second hidden unit is added and the process repeated, until four hidden units have been tried, or the SIC for q hidden units exceeds that for $q-1$ hidden units. This ANN model selection procedure is begun anew each time the data window moves forward one period. A different set of regressors and a different number of hidden units connected to different inputs may therefore be chosen at each point in time. We thus simulate a fairly sophisticated real time ANN forecasting implementation. We should expect the ANN models to have SIC values superior to (i.e smaller than) those of the linear models, as the ANN model can choose any of the linear models as a special case.

It should be stressed, though, that the estimation strategy outlined here is one of a plethora of available techniques for estimating neural network models. Indeed, a whole host of more sophisticated techniques which rely on various types of cross validation, etc. are now also available, often in "canned" computer packages. For references to other estimation strategies, the reader need look no further than any recent book or periodical on neural networks (of which there are many).

4. Further Empirical Issues

4.1 Data Transformations

Many macroeconomic data are seldom available to the applied econometrician in raw format. For example, quarterly and monthly data often consist of aggregated quantities, averaged prices, or even interpolated data. Also, seasonally observed data are frequently only available after some form of seasonal adjustment has been applied. In brief, many seasonal correction methods apply sequences of outlier removal and moving average filtering techniques to discretely measured data. Such correction methods may consist of the following steps. First, a trend is estimated and removed. Second, outliers are removed and two-sided moving average filters are applied. Third, Steps 1 and 2 are repeated in a number of iterations. In any step, irregular data are treated according to the functional form of the estimated trend and seasonal factors. Perhaps not surprisingly, in many cases these types of procedures may result in seasonally adjusted time series which have properties quite different for the original unadjusted series. For example, Ghysels, Granger, and Siklos (1996) show that seasonal adjustment may introduce nonlinearity into an otherwise linear process.

On the other hand, seasonal adjustment may actually reduce the relevance of switching regimes, say, leading to a finding of "less" nonlinearity. This arises as sequences of moving average filters clearly smooth away the effects of structural shifts. Since many nonlinear models are designed to model regime switches, one may expect that such models are less useful for seasonally adjusted data. Intuitively, the application of a sequence of moving average filters implies that a current, say recessionary, observation is replaced by some weighted average of past and future observations, some of which may lie within expansionary phases. This in turn suggests that the seasonally adjusted observations dampen the oscillatory affect that economic expansions and contractions have on economic variables, for example. In terms of switching regime models, one may expect that seasonal adjustment decreases the probability of switching regimes (transition probability), and thus the probability that one stays within the same regime increases. Using simulations and empirical examples, Franses and Paap (1996) confirm this conjecture for the Markov switching model of Hamilton (1989).

In summary, we recommend that extreme caution be taken when using seasonally adjusted data, as the incorrect use of such data may lead to quite misleading findings concerning to presence of nonlinearity, for example.

4.2 Outliers

A typical feature of economic time series is that some observations are irregular. For example, the May 1968 strike in France resulted in a very obvious downward spike. Taken one step further, one may, in some cases, wish to classify two successive negative valued observations of gdp growth as outliers, if in the middle of an extremely long expansionary phase, for example. One nice feature of nonlinear models is that they may be explicitly designed to jointly describe the different dynamics associated with expansions and recessions. However, since nonlinear models already tend to be constructed using a substantial number of parameters one may wish to stipulate that parameters are not used to fit only one or two (outlying) observations. If such a strategy is not followed, difficulties at the estimation stage may arise. Further, in such cases, out-of-sample forecasts may be poor relative to forecasts constructed using simpler more parsimonious models. Given these considerations, it becomes relevant to test for nonlinearity in the presence of outliers. Van Dijk, Franses, and Lucas (1996a, b) propose tests for nonlinearity and ARCH in the presence of outliers. Their tests are based on the generalized M-estimator, where possibly irregular data points are given less weight. Extensive Monte Carlo evidence in these two papers shows that these types of robust tests statistics have good size properties, and suffer little from diminished power. It is further shown that standard tests for nonlinearity and ARCH can have empirical sizes ranging from 20% to 100%, if not corrected for outliers!

The application of these robust tests for nonlinearity to the industrial production data used by Anderson and Teräsvirta (1992) reveals that for Austria, Belgium, and the U.S.A. there actually appears to be no nonlinearity after all. Upon further inspection of the data, it appears that the nonlinearity found using standard nonrobust tests is attributable to a small number of outlying observations. For financial data, Franses and Van Dijk (1997) document that one may also often find ARCH because of neglected outliers. In particular, they examine 22 weekly and monthly exchange rate series, as well as 13 stock market indices for samples of length 5 years. Their main result is that spurious GARCH is found over 50% of the time. Monte Carlo experiments are also run, providing further evidence that Franses and Van Dijk's (1997) results are indeed driven by outliers.

In another related study, Franses and Draisma (1997) find that hidden layers in ANN activation functions are frequently activated by only a single observation. This suggests that the hidden unit is used to fit only one observation, which may lead to the problems discussed at the beginning of the section. From a practical perspective, we recommend careful investigation into the possibility that findings of nonlinearity are driven by one or few irregular observations. This can be done quite easily by using the robust nonlinearity tests discussed above, for example. However, there is a clear tradeoff between how many outliers are admitted into the DGP and the usefulness of nonlinear models for capturing asymmetries and nonlinearities which may characterize a data series. In particular, allowing too few outliers also has drawbacks, as one may argue that certain so called outliers are precisely those data which are not well treated using linear models, and which we wish to model using nonlinear models.

4.3 Model Selection

Probably one of the most difficult (and currently unresolved) issues in empirical modelling is that of model selection. Once we leave the linear world, there are a virtually endless collection of nonlinear models available, making comparison of all nonlinear alternatives essentially impossible. Indeed, comparing only a few nonlinear alternatives is already, in many cases, quite time consuming. Broadly speaking, there are two strategies which one may wish to consider in such cases. The first involves fitting a linear model and applying nonlinearity tests which have power in the direction in which the investigator is interested. One drawback of this approach is that many of the currently available nonlinearity tests are portmanteau tests, having some power in many directions, but little power in any particular direction. Other tests are powerful in the direction of interest, but also have power against selected other alternatives. An example here is the STAR (smooth transition autoregressive) test (see Granger and Teräsvirta (1993) and the references contained therein), which has power against ARCH, while the ARCH-LM test has power against STAR. A second possible strategy is just to fit one's "favorite" nonlinear model, and investigate coefficient restrictions (such as zero restrictions). However, it is often difficult to exactly identify a nonlinear model, leading to some difficulty when implementing this strategy, in many cases. Also, when estimated parameters have population coefficients equal to zero, some of the estimation techniques used in practice break down. An applied strategy which is often used in place of the above strategies (see also Sections 2 and 3 above) is to use forecast based comparison, or to use the SIC and AIC, for example. These criteria are quite popular, even though much remains unknown with respect to their small sample and consistency properties, in a nonlinear context. However, the SIC and AIC are based on mean-square-error loss, and when somewhat complex nonlinear models (such as ANN and STAR) models are examined, it is not clear if such loss criteria are relevant. For example, in many cases asymmetric loss may be of interest. Further, as competing nonlinear models often have very different characteristics (and thus strengths and weaknesses), one must be careful to compare either (i) all features of the competing models, or (ii) only those features which are of particular interest. Boswijk and Franses (1997) suggest that one time series characteristic which is often of interest is the impulse response function (IRF), and they propose using IRFs as an aid to model selection. For example, seemingly very different models may have similar IRFs, while models which seem alike in many respects may have very different IRFs. While IRFs cannot be used to form a final decision when comparing models, they can serve as an aid, and may used, in addition to the various other model selection criteria discussed above, when selecting among competing models.

In summary, when selecting among nonlinear models, it is sensible to first select model selection criteria based on carefully constructed loss functions. Individual end users generally require models which are useful for many different purposes, from forecasting mean to maximizing profits. Second, it is reasonable to assume than any single model selection criteria only sheds light on a small "part" of the overall picture. Using a number of model selection criteria may thus be useful when comparing models. Third, there are an infinite variety of models which may be compared. When using nonlinear models, it is perhaps sensible to begin with some basic benchmark linear models, and then consider a small set of alternative nonlinear models which are of particular interest, given economic theory, and other considerations.

References

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

Boswijk, H.P. and P.H. Franses, 1997, Common Persistence in Nonlinear Autoregressive Models, Econometric Institute Report 9702.

Brandt, A., 1986, The Stochastic Equation $Y_{n+1} = A_n Y_n + B_n$ with Stationary Coefficients, Advances In Applied Probability 18, 211-220.

Brock, W.A., W.D. Dechert, and Scheinkman, J.-A., 1996, A Test for Independence Based on the Correlation Dimension, Econometric Reviews 15, pp. 197-235.

Brock, W.A., D.A. Hsieh, and B. LeBaron, 1991, *Nonlinear Dynamics, Chaos, and Instability*, Cambridge: MIT Press.

Burns, A.F. and W.C. Mitchell, 1946, *Measuring Business Cycles*, Columbia University Press.

Carroll, S.M. and B.W. Dickinson,, 1989, Construction of Neural Nets Using the Radon Transform, in *Proceedings of the International Joint Conference on Neural Networks*, Washington DC. New York: IEEE Press, pp. 607-611.

Chen, X. and N.R. Swanson, 1997, Semiparametric ARX Neural Network Models With an Application to Forecasting Inflation, Pennsylvania State University, working paper.

Corradi, V., N.R. Swanson, and H. White, 1997, Testing for Stationarity Ergodicity and for Comovements Between Nonlinear Discrete Time Markov Processes, Working Paper, Department of Economics, Pennsylvania State University.

Cybenko, G., 1989, Approximation by Superpositions of a Sigmoid Function, Mathematics of Control Signals and Systems 2, 303-14.

Croushore, D., 1993, Introducing: The Survey of Professional Forecasters, The Federal Reserve Bank of Philadelphia Business Review, November-December, 3-15.

Diebold, F.X. and R.S. Mariano, 1995, Comparing Predictive Accuracy, Journal of Business and Economic Statistics 13, 253-263.

Diebold, F.X., and J.A. Nason, 1990, Nonparametric Exchange Rate Prediction, Journal of International Economics 28, 315-332.

Dorsey, R.E., J.D. Johnson, and M.V. van Boening, 1994, The Use of Artificial Neural Networks for Estimation of Decision Surfaces in First Price Sealed Bid Auctions, in W.W. Cooper and A. Whinston eds., *New Directions in Computational Economics.* Boston: Kluwer, pp. 19-40.

Dropsy, V., 1992, Exchange Rates and Neural Networks, California State University Fullerton Department of Economics Working Paper 1-92.

Dutta, S. and S. Shekhar, 1989, Bond Rating: A Non-Conservative Application of Neural Networks, in *Proceedings of the IEEE International Conference on Neural Networks*, San Diego. New York: IEEE Press, pp. 443-450.

Engle, R.F., 1982, Autoregressive Conditional Heteroskedasticity With Estimates of the Variance of U.K. Inflation, Econometrica 50, 987-1008.

Engle, R.F. and S.J. Brown, 1986, Model Selection for Forecasting, Applied Mathematics and Computation 20, 313-327.

Fair, R.C. and R.J. Shiller, 1990, Comparing Information in Forecasts from Econometric Models, American Economic Review 80, 375-389.

Franses, P.H., 1996, *Periodicity and Stochastic Trends in Economic Time Series*,Oxford, Oxford University Press.

Franses, P.H. and D. Van Dijk, 1997, Do We Often Find ARCH Because of Neglected Outliers?, Econometric Institute Report 9615, Erasmus University.

Franses, P.H. and G. Draisma, 1997, Recognizing Changing Seasonal Patterns Using Artificial Neural Networks, Journal of Econometrics, forthcoming.

Franses, P.H. and R. Paap, 1996, Does Seasonal Adjustment Change Inference from Markov Switching Models?, Econometric Institute Report 9615, Erasmus University.

Franses, P.H. and P. Van Homelen, 1997, On Forecasting Exchange Rates Using Neural Networks, Applied Financial Economics, forthcoming.

Funahashi, K., 1989, On the Approximate Realization of Continuous Mappings by Neural Networks, INeural Networks 2, 183-92.

Ghysels, E., C.W.J. Granger, and P.L. Siklos, 1996, Is Seasonal Adjustment a Linear or Nonlinear Data Filtering Process?, Journal of Business and Economic Statistics 14, 374-386.

Granger, C.W.J., 1993, Strategies for Modeling Nonlinear Time Series Relationships, The Economic Record 69, 233-238.

Granger, C.W.J. and A.P. Andersen, 1978, *An Introduction to Bilinear Time Series Models*, Göttingen, Vandenhoeck and Ruprecht.

Granger, C.W.J. and N.R. Swanson, 1996, Further Developments in the Study of Cointegrated Variables, Oxford Bulletin of Economics and Statistics 58, 537-553.

Granger, C.W.J. and N.R. Swanson, 1997, An Introduction to Stochastic Unit Root Processes, Journal of Econometrics, forthcoming.

Granger, C.W.J. and T. Teräsvirta, 1993, *Modelling Nonlinear Economic Relationships*, New York: Oxford.

Guyton, D.A., N.-F. Zhang and R.V. Foutz, 1986, A Random Parameter Process for Modeling and Forecasting Time Series, Journal of Time Series Analysis 7, 105-15.

Hall, R.E., 1978, Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence, Journal of Political Economy 96, 971-87.

Hamilton, J.D., 1989, A New Approach to the Economic Analysis of nonstationary Time Series and the Business Cycle, Econometrica 57, 357-384.

Hamilton, J.D., 1994, *Time Series Analysis*, New Jersey, Princeton University Press.

Henriksson, R.D. and R.C. Merton, 1981, On Market Timing and Investment Performance. II. Statistical Procedures for Evaluating Forecasting Skills, Journal of Business 54, 513-533.

Hornik, K., M. Stinchcombe, and H. White, 1989, Multilayer Feedforward Networks are Universal Approximators, Neural Networks 2, 359-66.

Hornik, K., M. Stinchcombe, and H. White, 1990, Universal Approximation of an Unknown Mapping and its Derivatives Using Multilayer Feedforward Networks, Neural Networks 3, 551-60.

Journal of Applied Econometrics, 1992, *Special Issue on Nonlinear Dynamics and Econometrics*, December, New York, Wiley.

Keane, M.P. and D.E. Runkle, 1990, Testing the Rationality of Price Forecasts, American Economic Review 80, 714-735.

Koopmans, L.H., 1974, The Spectral Analysis of Time Series (Academic Press, New York).

Kuan, C.-M. and T. Liu, 1995, Forecasting Exchange Rates Using Feedforward and Recurrent Neural Networks, Journal of Applied Econometrics 10, 347-364.

Kuan, C.-M. and H. White, 1994, Artificial Neural Networks: An Econometric Perspective, Econometric Reviews 13, 1-91.

Laroque, G. and B. Salanie, 1994, Estimating the Canonical Disequilibrium Model: Asymptotic Theory and Finite Sample Properties, Journal of Econometrics 62, 165-210.

Lee, T.-H., H. White, and C.W.J. Granger, 1993, Testing for Neglected Nonlinearity in Time Series Models: A Comparison of Neural Network Methods and Alternative Tests, Journal of Econometrics 56, 269-290.

Leitch, G. and J.E. Tanner, 1991, Economic Forecast Evaluation: Profits Versus the Conventional Error Measures, American Economic Review 81, 580-590.

Leybourne, S.J., B.P.M. McCabe and A.R. Tremayne, 1994, Can Economic Time Series be Differenced to Stationarity?, mimeo, University of Nottingham.

Luukkonen, R., P. Saikkonen and T. Teräsvirta, 1988a, Testing Linearity Against Smooth Transition Autoregressive Models, Biometrika 75, 491-9.

Luukkonen, R., P. Saikkonen and T. Teräsvirta, 1988b, Testing Linearity in Univariate Time Series Models, Scandinavian Journal of Statistics 15, 161-175.

Mishkin, F.S., 1988, The Information in the Term Structure: Some Further Results, Journal of Applied Econometrics 3, 307-14.

Mizrach, B., 1992, Multivariate Nearest-Neighbour Forecasts of EMS Exchange Rates, Journal of Applied Econometrics 7, 151-164.

Moody, J. and J. Utans, 1991, Principled Architecture Selection for Neural Networks: Applications to Corporate Bond Rating Predictions, in J.E. Moody, S.J. Hanson and R.P. Lippmann, eds., *Advances in Neural Information Processing Systems* 4. San Mateo: Morgan Kaufman, pp. 683-690.

Neftci, S.N., 1984, Are Economic Time Series Asymmetric Over the Business Cycles?, Journal of Political Economy 92, 307-328.

Pesaran, M.H. and A.G. Timmerman, 1992, A Simple Nonparametric Test of Predictive Performance, Journal of Business and Economic Statistics 10, 461-465.

Pesaran, M.H. and A.G. Timmerman, 1994a, The Use of Recursive Model Selection Strategies in Forecasting Stock Returns, mimeo (1994a).

Pesaran, M.H. and A.G. Timmerman, 1994b, A Generalization of the Non-Parametric Henriksson-Merton Test of Market Timing, Economic Letters 44, 1-7.

Potter, S.M., 1995, A Nonlinear Approach to US GNP, Journal of Applied Econometrics 10, 109-125.

Pourahmadi, M., 1986, On Stationarity of the Solution of a Doubly Stochastic Model, Journal Of Time Series Analysis 7, 123-131.

Pourahmadi, M., 1988, Stationarity of the Solution of $X_t = A_t X_{t-1} + \varepsilon_t$ and Analysis of Non-Gaussian Dependent Random Variables, Journal of Time Series Analysis 9, 225-230.

Ripley, B.D., 1994, Neural Networks and Related Methods for Classification, Journal of the Royal Statistical Society Series B 56, 409-456.

Rothman, P., 1992, The Comparative Power of the TR Test Against Simple Threshold Models, Journal of Applied Econometrics 7, 187-195.

Rumelhart, D. E. and J.L. McClelland, 1986, *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, Cambridge: MIT Press.

Stekler, H.O., 1991, Macroeconomic Forecast Evaluation Techniques, International Journal of Forecasting 7, 375-384.

Stekler, H.O., 1994, Are Economic Forecasts Valuable?, Journal of Forecasting 13, 495-505.

Swanson, N.R., 1996, Forecasting Using First Available Versus Fully Revised Economic Time Series Data, Studies in Nonlinear Dynamics and Econometrics 1, 47-64.

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 and Economic Statistics 13, 265-275.

Swanson, N.R. and H. White, 1997, A Model Selection Approach to Real Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks, Review of Economics and Statistics, forthcoming.

Teräsvirta, T. and H.M. Anderson, 1992, Characterizing Nonlinearities in Business Cycles Using Smooth Transition Autoregressive Models, Journal of Applied Econometrics 7, 119-136.

Tjøstheim, D.,1986, Some Doubly Stochastic Time Series Models, Journal of Time Series Analysis 7, 51-72.

Tong, H., 1983, *Threshold Models in Non-Linear Time Series Analysis*, New York, Springer-Verlag.

Tong, H., 1990, *Non-Linear Time Series: A Dynamical System Approach*, Oxford, Oxford University

Press.

Trippi, R. and E. Turbau, 1993, *Neural Networks in Finance and Investing* Chicago: Probus Publishing Company.

Van Dijk, D., P.H. Franses, and A. Lucas, 1996a, Testing for Smooth Transition Autoregression in the Presence of Outliers, Econometric Institute Report 9622, Erasmus University.

Van Dijk, D., P.H. Franses, and A. Lucas, 1996b, Testing for ARCH in the Presence of Additive Outliers, Econometric Institute Report 9659, Erasmus University.

Weiss, A.A., 1996, Estimating Time Series Models Using the Relevant Cost Function, Journal of Applied Econometrics 11, 539-560.

White, H., 1988, Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns, in *Proceedings of the IEEE International Conference on Neural Networks*, San Diego. New York: IEEE Press, pp. 451-458.

White, H., 1989, Learning in Artificial Neural Networks: A Statistical Perspective, Neural Computation 1, 425-64.

White, H., 1990, Connectionist Nonparametric Regression: Multilayer Feedforward Networks CVan Learn Arbitrary Mappings, Neural Networks 3, 535-549.

Zarnowitz, V. and P. Braun, 1992, Twenty-Two Years of the NBER-ASA Quarterly Economic Outlook Surveys: Aspects and Comparisons of Forecasting Performance, NBER Working Paper Number 3965.