 Vol. 21: 97–130, 1994

Non-linear Time Series: A Selective Review*

DAG TJØSTHEIM

University of Bergen

ABSTRACT. A selective review of non-linear time series is presented. All of the three phases in the specification–estimation–verification modelling cycle are covered, but much of the emphasis is on non-parametric and restricted non-parametric methods. In particular, recent results on non-parametric tests of linearity and independence are included.

Key words: non-linear time series, non-parametric identification, restricted non-parametric, test of independence, test of linearity

1. Introduction

Until quite recently much of time series modelling has been confined to ARMA models; i.e. processes $\{X_t\}$ generated by a difference equation scheme

$$X_t - \sum_{i=1}^{p} a_i X_{t-i} = e_t + \sum_{i=1}^{q} b_i e_{t-i} \tag{1.1}$$

where $\{e_t\}$ is a sequence of zero-mean independent identically distributed (i.i.d.) random variables. In the book by Box & Jenkins (1970) the analysis was formalized into (i) model specification, (ii) estimation and (iii) model verification. The specification phase consists in determining the degree of differencing needed to make the model stationary and in determining p and q. The estimation stage yields estimates of $\sigma^2 = E(e_i^2)$ and the parameters $\{a_1, \ldots, a_p\}$ and $\{b_1, \ldots, b_q\}$ for given values of p and q. Finally, the model verification amounts to checking whether the estimated residual process $\{\hat{e}_t\}$ can be identified with a white noise process.

The modelling framework can be enlarged to include long range dependence with fractional ARMAs (see e.g. Dahlhaus, 1989; Robinson, 1992 and references therein), multivariate VARMA and VARMAX models (Hannan & Deistler, 1988) and handling of random walk non-stationarities via cointegration (Engle & Granger, 1987). Rather a complete theory has been developed for ARMA models, and much of it can be found in the book by Brockwell & Davis (1987).

If $\{e_t\}$ is Gaussian in (1.1), or in general if $\{X_t\}$ is a Gaussian process, then the conditional mean of X_t given past Xs will be linear, the conditional variance of X_t given previous Xs will be a constant, and essentially the process will be time-reversible (Lawrance, 1991). There are classical data sets such as the sunspot, lynx and blowfly data (Tong, 1990) where these properties are far from being fulfilled, and there is an increasing awareness of deviations from these assumptions in general. In particular, econometricians have assembled increasing evidence for a non-constant (heteroscedastic) conditional variance describing a fluctuating risk structure for financial time series (Bera & Higgins, 1993).

Sometimes it is claimed that one should attempt to describe these non-linear features by non-Gaussian ARMA processes, and indeed such models can be constructed so that they are

*This paper was presented as a Special Invited Lecture at the 14th Nordic Conference on Mathematical Statistics, Røros, June 1992.

non-reversible and have a non-linear conditional mean and variance. While it can be said that the modelling capacity of this class is not exhausted and that it deserves closer investigation (cf. Breidt & Davis, 1991), I think such a framework is basically too restrictive to describe the phenomena of interest. For example, for a given non-linear regression structure, which may have been non-parametrically estimated, it seems to be easier and more natural to fit an explicit non-linear model of the type to be discussed in section 2 rather than to look for a non-Gaussian ARMA model.

There are at least two chief objectives which warrant the introduction of non-linear models for univariate series. The first, and perhaps the most important, is to say something about the structure of the data generating mechanism at hand. Occasionally, but rarely, it is clear from *a priori* considerations of the physics involved that a non-linear model of a certain type will be natural for describing the data (e.g. Bølviken *et al.*, 1992; Karlsen & Tjøstheim, 1990; Tong *et al.*, 1985). A more common situation is where the data indicate that there is a non-linearity present, and one would like to characterize it by fitting an appropriate model. Such a characterization could be applied directly in a classification procedure and in discriminating between non-linearities in pattern recognition. In some cases it could conceivably contain hints of the physical mechanism generating the data.

The second objective is to do optimal forecasts. If the data-generating mechanism is non-linear, one expects forecasts based on non-linear models to be superior, and several such examples have been found (see e.g. Lewis & Stevens, 1991a, b; Chen & Tsay, 1993a). However, caution should be used here. It is tempting to try a large number of non-linear models to minimize forecasts errors for the data given. Often there is only a fairly limited number of observations available, and there is the danger of over-fitting the model. One could argue that only a relatively restricted set of models should be tried (Teräsvirta *et al.*, 1993a), and this means that a good model specification routine should be available. Even when this precaution is taken into account, there is no guarantee it will give better forecasts in general. For example, if the non-linearity is associated with a particular feature of the data, and this feature does not occur in the post-sample evaluation, then the non-linear model may perform on a par with or even worse than a linear model.

The introduction of ARCH (autoregressive conditional heteroscedastic) models has added a new dimension to the prediction problem. By assuming a non-constant conditional variance depending on past observations, it is possible to predict the future variance (risk) as well as the mean. It should be noted, however, that an ARCH model may sometimes be the result of an incorrect specification of the conditional mean, in which case neither of the forecasts is optimal.

Our survey of non-linear models is motivated by the two objectives mentioned above. It is comprehensive in the sense that it will touch on all of the three main phases: specification, estimation and verification. On the other hand it is selective. The emphasis on the various themes and subthemes will be subjective and largely according to the author's current interest. Among important subjects that are virtually not discussed are properties of higher order spectra (Subba Rao & Gabr, 1984) relationships with chaos theory (see the recent issue of *J. Roy. Statist. Soc. Ser. B*, 1992), and with neural network modelling (Stinchcombe & White, 1989), limit cycle analysis (Tong, 1990) and regime models (Bølviken, 1993). Also, I will restrict myself to univariate and stationary series.

An outline of the paper is as follows: the most important parametric and non-parametric model classes are reviewed in section 2. Tests of non-linearity are described in section 3 with emphasis on some work in progress. In the section on specification (section 4) a major portion is on recent non-parametric work. The actual fitting and estimation of a given model with corresponding asymptotic theory are briefly surveyed in sections 5 and 6, whereas model

verification is treated in more depth in section 7; in particular its connection with recently developed independence tests.

2. Classes of non-linear models

It is not entirely clear how a non-linear model should be defined. The difficulty stems from the problem of defining exactly what is meant by a linear model. In this paper I adopt a rather pragmatic approach and essentially consider processes that are not of the ARMA type (1.1) to be non-linear. I first define the main non-linear classes in sections 2.1, 2.2 and 2.3 and then state some properties in section 2.4.

2.1. Parametric models

The parametric models may be somewhat arbitrarily subdivided into four different classes, and I will treat each of them briefly. In each case I look at rather simple, often first order, processes to better illustrate the characteristic properties of each class. I refer to the literature (in particular Tong, 1990) for a much more general and detailed treatment.

Class 1. Parametric models for the conditional mean

The objective is to model the conditional mean of X_t parametrically given previous observations. For Gaussian processes the conditional mean is linear. For a first order model

$$X_t = f(X_{t-1}, \boldsymbol{\theta}) + e_t$$

where $X_0 = x_0$ and where $\{e_t, t \geqslant 1\}$ is a series of i.i.d. random variables, we have

$$E(X_t \mid X_{t-1} = x) = f(x, \boldsymbol{\theta}).$$

Here the function f is known and $\boldsymbol{\theta}$ is an unknown vector parameter to be estimated.

Choosing particular functions f yields some of the "classical" non-linear models, i.e.

$$\text{Threshold AR:} \quad f(x, \boldsymbol{\theta}) = \theta_1 x 1(x \geqslant k) + \theta_2 x 1(x < k)$$

where $1(\cdot)$ is the indicator function and k is a threshold. This model class is very useful, and in our simple example consists of linking two AR models together. It can be generalized in a number of ways. It is also possible to have a transition between two models in a smoother fashion as in

$$\text{Exponential AR:} \quad f(x, \boldsymbol{\theta}) = \{\theta_1 + \theta_2 \exp(-\theta_3 x^2)\}x \tag{2.1}$$

where $\theta_3 > 0$, and where there is a smooth transition between the linear model $X_t = \theta_1 X_{t-1} + e_t$ at infinity and the non-linear behaviour in the finite domain. There are other ways of constructing smooth transition models (Chan & Tong, 1986), and Granger & Teräsvirta (1992b) have looked at these models quite systematically. A flexible one is the smooth logistic transform, where

$$\text{Logistic AR:} \quad f(x, \boldsymbol{\theta}) = \theta_1 x + \theta_2 x\{[1 + \exp(-\theta_3(x - \theta_4))]^{-1} - \tfrac{1}{2}\}, \quad \theta_3 > 0. \tag{2.2}$$

A related general class of higher order models is

$$X_t = \mathbf{a}^{\mathrm{T}} \mathbf{X}_{t-1} + \sum_{i=1}^{k} b_i \phi_i(\boldsymbol{\gamma}_i^{\mathrm{T}} \mathbf{X}_{t-1}) + e_t$$

where $\mathbf{X}_t^{\mathrm{T}} = [X_t, \ldots, X_{t-p+1}]$, with $\mathbf{a}^{\mathrm{T}} = [a_1, \ldots, a_p]$ and $\gamma^{\mathrm{T}} = [\gamma_1, \ldots, \gamma_p]$ being p-dimensional parameter vectors and where the ϕ_is in general are known squashing or transition functions such as probability density functions, logistic functions or sigmoidal functions. This class is essentially the neural network class, which has been used with considerable success in various fields and which can be used to approximate quite arbitrary functions (Stinchcombe & White, 1989). In neural nets $\phi_i = \phi$, $i = 1, \ldots, k$ usually. In that case the model is not identified (estimable) as parameters (b_i, γ_i), (b_j, γ_j) are exchangeable.

Class 2. Parametric models for the conditional variance

The motivation is that it is of interest to predict not only the mean but also the variance of the process given past information on $\{X_t\}$. The variance is an expression of risk or volatility in economic time series, and it is often found that extreme values of X_t lead to larger fluctuations in subsequent observations than more moderate values. Engle (1982) expressed this in the pure ARCH process given by

$$X_t = (\theta_1 + \theta_2 X_{t-1}^2)^{1/2} e_t$$

where $\mathrm{var}\,(X_t \mid X_{t-1} = x) = (\theta_1 + \theta_2 x^2)\sigma^2$, with $\sigma^2 = \mathrm{var}\,(e_t)$. A more general ARCH model could have the form

$$X_t = g(\mathbf{X}_{t-1}, \boldsymbol{\theta}) e_t$$

where g is a known positive function. Usually g is symmetric in some sense but asymmetric functions have also been discussed (Engle & Ng, 1991; Diebolt & Guegan, 1991). Moreover, it is possible to introduce (Bollerslev, 1986) an AR part and recursions in the conditional variance structure of the residuals (GARCH). There is now a considerable literature on these processes, and a survey article is given by Bera & Higgins (1993).

Class 3. Mixed models

A general model containing both a conditional mean and a conditional variance component is the model

$$X_t = f(X_{t-1}, \ldots, X_{t-p}, \boldsymbol{\theta}_1) + g(X_{t-1}, \ldots, X_{t-p}, \boldsymbol{\theta}_2) e_t, \quad g > 0,$$

where f and g are known and where $E(X_t \mid X_{t-1} = x_1, \ldots, X_{t-p} = x_p) = f(x_1, \ldots, x_p, \boldsymbol{\theta}_1)$ and $\mathrm{var}\,(X_t \mid X_{t-1} = x_1, \ldots, X_{t-p} = x_p) = g^2(x_1, \ldots, x_p, \boldsymbol{\theta}_2)\,\mathrm{var}\,(e_t)$. Non-linearity in both the conditional mean and the conditional variance can also be obtained using more "classical" non-linear models such as the bilinear model, which in a very simple situation can be stated as

$$X_t = \theta_1 + \theta_2 X_{t-1} + \theta_3 X_{t-1} e_t + \theta_4 X_{t-1} e_{t-1} + e_t.$$

Even this simple model is quite flexible. It is capable of producing burst-like phenomena, and in general it has both a non-linear conditional mean and a non-linear conditional variance. However, one difficulty is that one cannot in general compute an explicit expression for them, and for a given conditional mean or conditional variance structure, it is not obvious which kind of bilinear model is most appropriate. The fact remains, though, that the bilinear class is a flexible one and therefore of interest in applications.

Class 4. Regime models

Part of the stimulus behind this class consists in a desire to model stochastic changes in a structure, while still retaining stationarity in the compound process. A simple example is

where the autoregressive parameter in an AR(1) process is replaced by a Markov chain $\{\theta_t\}$ having a finite state space; for example, taking two values a_1 and a_2, resulting in the simple doubly stochastic model (Tjøstheim, 1986a)

$$X_t = \theta_t X_{t-1} + e_t.$$

The process $\{X_t\}$ alternates between the AR processes

$$X_t = a_1 X_{t-1} + e_t$$

and

$$X_t = a_2 X_{t-1} + e_t,$$

and the change is regulated by the transition probability of the hidden Markov chain $\{\theta_t\}$; i.e. by an external agent, and not, as in the threshold case, by values of the variable X_{t-1} itself. A special case is where $\{\theta_t\}$ is i.i.d., which results in random coefficient autoregressive (RCA) models (Nicholls & Quinn, 1982). In turn Lawrance & Lewis (1985) define models that are formally subsets of RCA models, but their viewpoint and motivation are rather different. Some early and related references are Cartwright & Newbold (1983) and Tong (1983, p. 62).

More smooth changes in the parameter can be modelled by a Kalman type state space model

$$X_t = \theta_t X_{t-1} + e_t,$$

$$\theta_t = a\theta_{t-1} + \eta_t$$

where $\{\theta_t\}$ is itself modelled by an AR process and where $\{e_t\}$ and $\{\eta_t\}$ are independent i.i.d. processes. It should be noted that the problem of finding parameter values which render the model stationary, is difficult (Karlsen, 1990a, b).

The regime models, or more specifically the hidden Markov chain models, extend far beyond time series analysis and have been used in a variety of situations. It is not obvious how they can be recognized from a given data trace or from plots of the conditional mean and variance. They are probably most appropriate when there is some *a priori* information on the physical structure of the problem, which can give the hidden parameter process a physical interpretation. Two such examples are geological layers and speech recognition (Bølviken *et al.*, 1992; Karlsen & Tjøstheim, 1990; Rabiner, 1989). For more details on regime models refer to the paper by Bølviken (1993).

Why not polynomial models?

Perhaps the most useful and natural non-linear models in ordinary regression analysis are polynomial models. Thus, one should think that models of type

$$X_t = \sum_{i=1}^{k} a_i X_{t-1}^i + e_t$$

and more general ones including cross-terms would be very useful. However, if, as is in general the case, the noise distribution has infinite support, then these models explode due to the persistent feedback of $\{X_t\}$ into itself. For this reason they have not been much used in actual fitting to data, but auxiliary regressions of some linearity tests may have a polynomial form as here, with cross-terms included. Moreover, in fitting to data, polynomials can be used *locally*, constituting the very useful spline models.

2.2. Non-parametric modelling

Much of the emphasis for the classes of section 2.1 was on parametric modelling of the conditional mean and the conditional variance, separately and jointly. Alternatively, these quantities may be estimated non-parametrically, and in a way this is more natural since it allows for great flexibility without confining oneself to a special parametric model.

The most common way of estimating the conditional mean and variance non-parametrically is probably via the kernel estimator. For given observations $\{X_1, \ldots, X_n\}$ the conditional mean $M(x_1, \ldots, x_p) = E(X_t \mid X_{t-1} = x_1, \ldots, X_{t-p} = x_p)$ is estimated by

$$\hat{M}(x_1, \ldots, x_p) = \frac{(n-p)^{-1} \sum_{t=p+1}^{n} X_t \prod_{i=1}^{p} K_h(X_{t-i} - x_i)}{(n-p+1)^{-1} \sum_{t=p+1}^{n+1} \prod_{i=1}^{p} K_h(X_{t-i} - x_i)} \tag{2.3}$$

where $K_h(x) = h^{-1}K(h^{-1}x)$, with K being a kernel function, e.g. Gaussian or Epanechnikov, and h is the bandwidth of the kernel, and where $h \to 0$ as $n \to \infty$. We have used a product kernel although such a kernel may not be optimal. The conditional variance $V(x_1, \ldots, x_p) = \text{var}(X_t \mid X_{t-1} = x_1, \ldots, X_{t-p} = x_p)$ is obtained likewise as

$$\hat{V}(x_1, \ldots, x_p) = \frac{(n-p)^{-1} \sum_{t=p+1}^{n} X_t^2 \prod_{i=1}^{p} K_h(X_{t-i} - x_i)}{(n-p+1)^{-1} \sum_{t=p+1}^{n+1} \prod_{i=1}^{p} K_h(X_{t-i} - x_i)} - \hat{M}^2(x_1, \ldots, x_p). \tag{2.4}$$

Despite the flexibility of this modelling approach, there are a couple of drawbacks which hinder its general applicability. First, as can be expected and as is shown explicitly in section 6, the asymptotic mean square error of non-parametric estimates is larger than that of parametric ones, and it is increasing, relatively speaking, with the dimension p of the model. Second, and more serious, is the curse of dimensionality: As p increases, an extremely large sample size is needed to get a sufficient number of points in each unit cube of p-space. This means that there cannot be too many lags included in the model. In turn this implies that it will be important to single out significant and, in general, non-consecutive lags of the model. Even for this case, a non-parametric model obtained from (2.3) and (2.4) is not likely to be an end product of the analysis, but rather a starting point for fitting a suitable parametric model. If this point of view is adopted, the general estimates of conditional mean and variance, and of functionals thereof, will be of help mainly in the specification of a parametric or restricted non-parametric model to be used at a later stage. This point is elaborated further in section 4.

2.3. Restricted non-parametric and semi-parametric models

By restricting the scope somewhat there are a number of ways in which the curse of dimensionality can be evaded. Here the main restricted models are listed and then returned to in section 5. These models have been developed in a regression context, but they have recently been modified to include time series.

Additive models

In the simplest case of these models the conditional mean function is written as

$$E(X_t \mid X_{t-1} = x_1, \ldots, X_{t-p} = x_p) = M(x_1, \ldots, x_p) = \sum_j f_j(x_j). \tag{2.5}$$

For all the models the f_js are assumed to be unknown and are estimated non-parametrically. Since they are one-dimensional, the curse of dimensionality is avoided. A similar model can be formulated for the conditional variance, and in fact the ARCH model is a special case of an additive conditional variance function where the f_js are known. Additive models have been used extensively in regression analysis, and generalizations with interaction terms exist (Hastie & Tibshirani, 1990). Chen & Tsay (1993a) have analysed them in a time series situation. A related class is the ACE (alternating conditional expectation) models, where

$$E(h(X_t) \mid X_{t-1} = x_1, \ldots, X_{t-p} = x_p) = \sum_j f_j(x_j) \tag{2.6}$$

for some unknown function h. However, the algorithm is probably best suited to a regression context, where there is a clear distinction between the input and output variables. The method was developed by Breiman & Friedman (1985). Some curious aspects of the ACE algorithm are highlighted in Hastie & Tibshirani (1990, pp. 184–186).

Projection pursuit

These are additive models of linear combinations of past values, i.e.

$$X_t = \sum_i f_i(\gamma_i^{\mathrm{T}} \mathbf{X}_{t-1}) + e_t$$

where $\mathbf{X}_{t-1}^{\mathrm{T}} = [X_{t-1}, \ldots, X_{t-p}]$. The γ_is are unknown parameter vectors and the f_is are unknown functions which are estimated non-parametrically (Friedman & Stuetzle, 1981). Since they are functions of a scalar argument, the curse of dimensionality is avoided. Projection pursuit models are related to the neural network models of section 2.1, but for the latter the functions f_i are known. An application to time series involving simulations is contained in the paper by Granger & Teräsvirta (1992a).

Regression trees, splines and MARS

Assume a model

$$X_t = f(\mathbf{X}_{t-1}) + e_t$$

and approximate $f(\mathbf{x})$ in terms of simple basis functions $B_j(\mathbf{x})$ so that $f_{appr}(\mathbf{x}) = \Sigma_j c_j B_j(\mathbf{x})$. In the regression tree approach (Breiman *et al.*, 1984) f_{appr} is built up recursively from indicator functions $B_j(\mathbf{x}) = 1\ (\mathbf{x} \in R_j)$, where the regions R_j are partitioned in the next step of the algorithm according to a certain pattern. As can be expected, there are problems in fitting simple smooth functions like the linear model. Friedman (1991) in his MARS (multivariate adaptive regression splines) methodology has made at least two important new contributions. First, to overcome the difficulty in fitting simple smooth functions Friedman proposed not to eliminate the parent region, R_j, automatically in the above recursion scheme for creating subregions. In subsequent iterations both the parent region and its corresponding subregions are eligible for further partitioning. This allows for much greater flexibility. The second contribution is to replace step functions by products of linear left- and right-truncated regression splines. The products make it possible to include interaction terms. For a detailed discussion the reader is referred to Friedman (1991). Lewis & Stevens (1991a, b) have applied MARS to time series data, both simulated and real.

Stepwise series expansion of conditional densities

In a sense the conditional density $p(X_t \mid \mathbf{X}_{t-1})$ is the most natural quantity to look at in a non-linear modelling of $\{X_t\}$, since predictive distributions as well as the conditional mean and variance can all be derived from this quantity. Gallant & Tauchen (1990) used this as their starting point. The conditional density is estimated by expanding it in Hermite polynomials to avoid the curse of dimensionality. As a first approximation they are assumed to be of linear Gaussian and ARCH type, respectively. Applications are given in Gallant *et al.* (1990).

Semi-parametric models

Another way of trying to eliminate the difficulties in evaluating high-dimensional conditional quantities is to assume non-linear and non-parametric dependence in some of the predictors, and parametric and usually linear in others. An illustrative example is given by Engle *et al.* (1986), who modelled electricity sales using a number of predictor variables. It is natural to assume the impact of temperature on electricity consumption to be non-linear, as both high and low temperatures lead to increased consumption, whereas a linear relationship may be assumed for the other regressors. A similar situation arose in Shumway *et al.* (1988) in a study of mortality as a function of weather and pollution variables in the Los Angeles region.

New developments

Breiman (1992) and Breiman & Friedman are currently investigating the possibility of fitting "hinge" and "ramp" type functions to regression data. The method is designed to tackle data in very high dimensions, and they do have natural generalizations to univariate and multivariate time series.

2.4. *Some probabilistic properties*

In this section we mention briefly some questions relating to stationarity and mixing for the models introduced. These two properties are important in any asymptotic analysis.

Most of the models discussed in this paper can be restated as vector Markov chains, and hence the Markov theory developed for continuous state space Markov chains (see e.g. Nummelin, 1984) is at our disposal. Some features of this theory are highlighted in Tjøstheim (1990).

For example, for the model given by

$$X_t = f(X_{t-1}, \ldots, X_{t-p}) + e_t \tag{2.7}$$

for $t \geqslant 1$ and with $X_i = x_i$ for $i = -p+1, \ldots, 0$, the conditions

(i) f is bounded on compact sets,
(ii) $f(\mathbf{x}) = \mathbf{a}^{\mathrm{T}}\mathbf{x} + o(\|\mathbf{x}\|)$ as $\|\mathbf{x}\| \to \infty$, where the linear model $\mathbf{a}^{\mathrm{T}}\mathbf{x}$ is stable in the sense that $z^p - \Sigma_{i=1}^{p} a_i z^{p-i}$ has its zeros within the unit circle ($\mathbf{a}$ may of course be the zero vector),
(iii) the density of e_t is positive everywhere and $E(|e_t|) < \infty$,

imply the existence of a stationary distribution for $\mathbf{X}_0 = [X_0, X_{-1}, \ldots, X_{-p+1}]$, so that when $\{X_t\}$ is started in this distribution, it is strictly stationary.

These are sufficient conditions, nowhere close to being necessary. They are based on Tweedie-type criteria (Tweedie, 1975, 1983; Tjøstheim, 1990), and they are obtained trivially from those criteria using the test function $g(\mathbf{x}) = \|\mathbf{x}\|$ and a vectorization of the process.

The above conditions actually imply that the process is geometrically ergodic, so that the n-step transition probability converges to the invariant distribution geometrically fast as $n \to \infty$. This in turn (Pham, 1985) means that $\{X_t\}$ is absolutely regular, which implies that it is strongly (α) mixing. This last property is important since it allows the central limit theorem for strongly mixing processes to be applied, and one obtains a powerful instrument for analysing the asymptotic distribution of non-parametric as well as parametric estimates (see e.g. Masry & Tjøstheim, 1992).

For a parametric model where $f(x_1, \ldots, x_p) = f(x_1, \ldots, x_p; \boldsymbol{\theta})$ with f being known, the conditions (i)–(iii) translate into restrictions on the parameter $\boldsymbol{\theta}$ to obtain stationarity and geometric ergodicity. We refer to Tjøstheim (1990) for treatment of some non-linear parametric models. There are still many open problems in finding good conditions for $\boldsymbol{\theta}$ in general non-linear vector processes. Similar techniques can be used for regime models, but the problems appear to be even more difficult here (Karlsen, 1990a, b; Holst *et al.*, 1994).

If a non-constant conditional variance function is introduced in (2.7), the Tweedie criterion can still be used and sufficient conditions similar to (i)–(iii) can be stated (cf. Masry & Tjøstheim, 1992). Work in this direction on ARCH-related models has been done by Diebolt & Guegan (1991) and Bougerol & Picard (1992).

3. Tests of linearity

A first natural step in analysing a time series $\{X_t\}$ is to decide whether to use a non-linear model or not. This decision could be based on graphical plots of say the conditional mean and the conditional variance and/or on formal tests. Plots will be discussed mainly in the next section, and formal tests here. The existing tests can be divided roughly into parametric tests and non-parametric tests.

3.1. Parametric tests of linearity

We concentrate our exposition on the Lagrange multiplier tests (Saikkonen & Luukkonen, 1988; Luukkonen *et al.*, 1988a, b) since several, but not all, of the other tests come out as special cases of this procedure.

The Lagrange multiplier tests will be discussed in the somewhat simplified setting of the model

$$X_t = \mathbf{a}^{\mathrm{T}}\mathbf{X}_{t-1} + f(\boldsymbol{\theta}, \mathbf{X}_{t-1}) + e_t \tag{3.1}$$

where $\mathbf{X}_{t-1}^{\mathrm{T}} = [X_{t-1}, \ldots, X_{t-p}]$ and $\mathbf{a}^{\mathrm{T}} = [a_1, \ldots, a_p]$. The parameter $\boldsymbol{\theta}$ is generally a vector parameter such that $f(\boldsymbol{\theta}, \cdot) \equiv 0$ when $\boldsymbol{\theta}' = \mathbf{0}$, where $\boldsymbol{\theta}'$ is a suitable sub-vector of $\boldsymbol{\theta}$. In the logistic AR model of (2.2), for example, $\boldsymbol{\theta}$ could be taken as $\boldsymbol{\theta} = (\theta_2, \theta_3, \theta_4)$ and $\boldsymbol{\theta}'$ as θ_2 or θ_3.

In (3.1) $\mathbf{a}$ and $\boldsymbol{\theta}$ are unknown parameters, whereas f is assumed to be known. In other words, we are testing a specific non-linear parametric alternative by testing the hypotheses H_0: $\boldsymbol{\theta}' = \mathbf{0}$. A Lagrange multiplier principle is used to test this hypothesis, and an asymptotic theory can be developed under certain regularity conditions. We refer to the above papers, where there are also a number of simulation experiments for sample sizes ranging from 50 to 200, and where it is shown how some of the other parametric tests (e.g. Keenan, 1985: Tsay, 1986) come out as special cases by choosing f appropriately. A problem for this class of tests is its sensitivity to a wrongly specified f; both its functional form and its dimension. In addition there are certain identification problems. Other parametric non-linearity tests not quite covered by the Lagrange class include those of Teräsvirta *et al.* (1993b).

3.2. Non-parametric tests, the spectral domain

These tests originated with Subba Rao & Gabr (1980) and were improved by Hinich (1982). In particular Hinich has continued to work with them, and a number of results have been obtained by him and his co-workers (Ashley *et al.*, 1986; Brockett *et al.*, 1988).

The spectral density of a stationary process with an absolutely summable covariance function K is given by

$$f(\omega) = (2\pi)^{-1} \sum_{t=-\infty}^{\infty} K(t) \exp(-i\omega t)$$

where

$$K(t) = \operatorname{cov}(X_t, X_0) = \int_{-\pi}^{\pi} \exp(i\omega t) f(\omega)\, d\omega.$$

Similarly, for a zero-mean process with an absolutely summable third moment function $K(s, t) = E(X_s X_t X_0)$ the bispectrum is defined by

$$f_B(\omega_1, \omega_2) = (2\pi)^{-2} \sum_{s=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} K(s, t) \exp(-i\omega_1 s - i\omega_2 t)$$

so that

$$K(s, t) = \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \exp(i\omega_1 s + i\omega_2 t) f_B(\omega_1, \omega_2)\, d\omega_1\, d\omega_2.$$

It is well known that for a Gaussian process

$$f_B(\omega_1, \omega_2) \equiv 0. \tag{3.2}$$

As noted by Subba Rao & Gabr (1980) and Hinich (1982) for a linear process

$$X_t = \sum_{i=0}^{\infty} \alpha_i e_{t-i}$$

we have that the quantity

$$B(\omega_1, \omega_2) = \frac{|f_B(\omega_1, \omega_2)|^2}{f(\omega_1) f(\omega_2) f(\omega_1 + \omega_2)} \equiv c \tag{3.3}$$

where c is a constant, and the expressions (3.2) and (3.3) can now be taken as starting points for tests of Gaussianity and linearity.

It is known that (3.2) and (3.3) may hold also in non-Gaussian and non-linear cases, but the main criticism levelled against the bispectrum test has been that it needs considerably more observations to match the power of the best parametric tests for a given alternative. Thus, typically, in the simulation experiments of Ashley *et al.* (1986) they operate with sample sizes in the range 250–1000, and there seems to be widespread belief that they should not be applied for much smaller sample sizes. The so-called BDS test could also be thought of as a non-parametric linearity test (Brock & Potter, 1992), but I have stressed the independence aspect of this test, and the reader is referred to section 7.2 for more details.

3.3. A non-parametric test based on the conditional mean and the conditional variance

In a more informal approach to the testing problem one can plot estimates of the conditional mean $M_k(x) = E(X_t \mid X_{t-k} = x)$ and $V_k(x) = \operatorname{var}(X_t \mid X_{t-k} = x)$ for various lags k (see e.g.

Tong, 1990) to get an idea about the non-linearity. The quantities $M_k(x)$ and $V_k(x)$ are not always helpful in identifying the detailed structure of the model (Tjøstheim & Auestad, 1994a and section 4), but it does give a hint in many cases whether or not the process is non-linear and an idea about the strength of the non-linearity. Being one-dimensional, these quantities are easy to display graphically, and they are not influenced by the curse of dimensionality, as are the perhaps more natural quantities

$$M(x_1, \ldots, x_p) = E(X_t \mid X_{t-1} = x_1, \ldots, X_{t-p} = x_p)$$

and

$$V(x_1, \ldots, x_p) = \text{var}\,(X_t \mid X_{t-1} = x_1, \ldots, X_{t-p} = x_p).$$

An example with plots of $\hat{M}_1(x)$ for the exponential AR process

$$X_t = [0.5 + b \exp(-0.5X_{t-1}^2)]X_{t-1} + e_t$$

for $b = 0.0, 0.3, 0.6$ and 1.0 is given in Fig. 1. Here $\{e_t\}$ is a sequence of i.i.d. Gaussian variables with mean 0 and variance 1. The sample size $n = 1000$. To get an idea of the stochastic variation and the discriminatory power of such plots we have also plotted $\hat{M}_1(x)$ for five independent realizations of the process

$$X_t = 0.5X_{t-1} + e_t.$$

The curved behaviour close to the end points of Fig. 1 is due to boundary effects.

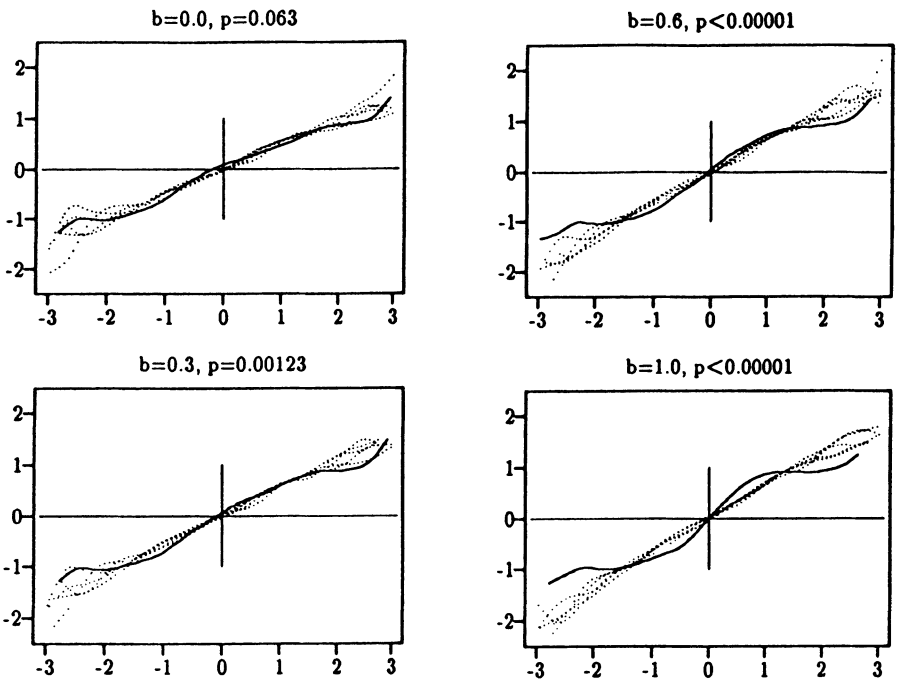


Fig. 1. Solid line: $\hat{M}_k(x)$ plotted against x for the exponential AR process $X_t = [0.5 + b \exp(0.5X_{t-1}^2)]X_{t-1} + e_t$ using $n = 1000$ observations. Dotted lines: $\hat{M}_k(x)$ plotted against x for the AR process $X_t = 0.5X_{t-1} + e_t$.

A more formal way of assessing stochastic fluctuations would be to construct confidence intervals, based on asymptotic theory, but they can be grossly misleading (Franke & Wendel, 1990). Still another possibility is to use the blockwise bootstrap (Kũnsch, 1989). The bootstrap captures much, but not all, of the individual variations of truly independent realizations.

How more formal test procedures can be constructed based on $\hat{M}_k(x)$ and $\hat{V}_k(x)$ is examined in this subsection. This is work still in progress, and more details can be found in Hjellvik & Tjøstheim (1994). For the four particular realizations of the exponential AR process depicted in Fig. 1 the formal tests resulted in p-values of 0.063, 0.0012 and less than 0.00001 for $b = 0.0$, 0.3, 0.6 and 1.0, respectively.

The test is based on measuring differences betweem $\hat{M}_k(x)$ and $\hat{\rho}_k x$, where $\rho_k = \text{corr}\,(X_t, X_{t-k})$ and

$$\hat{\rho}_k = \frac{\sum_t (X_t - \bar{X})(X_{t-k} - \bar{X})}{\sum_t (X_t - \bar{X})^2}$$

with $\bar{X}$ being the sample mean. Such a procedure amounts to measuring differences between the estimated linear and non-linear predictor of X_t based on X_{t-k}. We have constructed the functional

$$\hat{I}_k = n^{-1} \sum_{t=1}^{n} (\hat{M}_k(X_t) - \hat{\rho}_k X_t)^2 w(X_t)$$

where, corresponding to (2.3),

$$\hat{M}_k(x) = \frac{(n-k)^{-1} \sum_{t=k+1}^{n} X_t K_h(X_{t-k} - x)}{(n-k+1)^{-1} \sum_{t=k+1}^{n+1} K_h(X_{t-k} - x)},$$

and the idea is to compare the value of this functional with the null distribution of it under the hypothesis of an ARMA model. Here w is a weight function usually assumed to be an indicator function on a fixed compact support. The weight function serves a double purpose. It screens off some of the extreme observations, and it makes asymptotic arguments easier. Moreover, it should be noted that we always scale $\{X_t\}$ so that it has zero mean and standard deviation equal to one, and the bandwidth h has been chosen as $h = n^{-\frac{1}{5}}$.

Hjellvik & Tjøstheim (1994) showed that under regularity conditions

$$\hat{I}_k \xrightarrow{\text{a.s.}} I_k \doteq \int (M_k(x) - \rho_k x)^2 w(x)\, dF(x)$$

where F is the marginal distribution function of X_t. It is considerably more difficult to obtain results on the asymptotic distribution function of $\hat{I}_k$, and this problem has only been partially solved.

Actually, exact theoretical results on the asymptotic distribution seem to be of limited practical value in so far as the dominant term of the Taylor expansion giving the mean and variance of $\hat{I}_k$ can be off by a factor of 1.5–2 from the corresponding simulated quantities for sample sizes in the range 50–500 and for quite a wide number of simulation experiments. This means that it will be problematic to use asymptotic theory to set critical values for the test corresponding to a certain nominal size. Use of asymptotic normality in this manner actually leads to gross underestimation of the level in most cases, and similar problems occur

if the asymptotic expressions are fitted to a two-parameter gamma distribution. It should be observed that these problems are similar but worse than those encountered in Skaug & Tjøstheim (1993a) for an independence test (see section 7). For that test the asymptotics are easier to obtain and have been worked out in some detail.

To avoid these problems we have resorted to a bootstrap-based test. For initial data $\{X_t, t = 1, \ldots, n\}$ we first fit a linear AR (or ARMA) model to the data, so that

$$X_t - \sum_{i=1}^{p} \hat{a}_i X_{t-i} = \hat{e}_t.$$

Under the null hypothesis of a linear model, linear meaning in effect here autoregressive, the estimated residuals $\hat{e}_t$ correspond to an i.i.d. process. We therefore centre and bootstrap them to obtain $\{e_t^*\}$ and then form the linear bootstrap replicas $\{X_t^*\}$ by

$$X_t^* = \sum_{i=1}^{p} \hat{a}_i X_{t-i}^* + e_t^*.$$

Subsequently we compute the corresponding test variables $I_k^* = n^{-1} \Sigma_t (\hat{M}_k^*(X_t^*) - \hat{\rho}_k^* X_t^*)^2 w(X_t^*)$. There are now two possibilities. Either critical values can be found from the bootstrap empirical distribution of I_k^*, or this distribution can be parametrized in terms of the bootstrapped mean and standard deviation of I_k^*. In Hjellvik & Tjøstheim (1994) we have mainly chosen the last possibility to avoid creating too many bootstrap replicas. We have fitted both a normal distribution and a gamma distribution to the bootstrapped mean and standard deviation. The gamma distribution gives the best approximation to the nominal size of the test, which is not surprising since the empirical distribution I_k^* is skewed. Using this approach for $n = 100$, the simulated level is generally within 1–2% of the nominal level at the 5 and 10% level with somewhat larger relative discrepancies at the 1% level. The test is now completed by computing

$$\hat{I}_k = n^{-1} \sum_t (\hat{M}_k(X_t) - \hat{\rho}_k X_t)^2 w(X_t)$$

and comparing it with the critical value in the right hand tail of the bootstrap distribution of $\{I_k^*\}$.

A similar test for the conditional variance is constructed by considering the functional

$$\hat{J}_k = n^{-1} \sum_{t=1}^{n} (\hat{V}_{k,\hat{e}}(\hat{e}_t) - \hat{\sigma}_{\hat{e}}^2)^2 w(\hat{e}_t)$$

where, corresponding to (2.4),

$$\hat{V}_{k,\hat{e}}(x) = \frac{(n-k)^{-1} \sum_{t=k+1}^{n} \hat{e}_t^2 K_h(\hat{e}_{t-k} - x)}{(n-k+1)^{-1} \sum_{t=k+1}^{n+1} K_h(\hat{e}_{t-k} - x)} - \hat{M}_{k,\hat{e}}^2(x),$$

and where $\hat{\sigma}_{\hat{e}}^2 = n^{-1} \Sigma_t (\hat{e}_t - \bar{\hat{e}})^2 = 1$ due to the normalization of $\{\hat{e}_t\}$. This means that we have a conditional variance test that acts directly on the residuals. This is akin to the procedure used in ARCH modelling. A null distribution is formed by bootstrap replicating J_k as

$$J_k^* = n^{-1} \sum_t (\hat{V}_{k,e^*}(e_t^*) - \hat{\sigma}_{e^*}^2)^2 w(e_t^*).$$

In this case very accurate estimates of the level of the test were obtained for the examples studied. The test is carried out by comparing $\hat{J}_k$ to the critical value of the bootstrap distribution.

Several comments are in order.

1. To get a test independent of a specific lag we have also looked at test statistics going over several lags, like $\tilde{I}_k = \Sigma_{i=1}^{k} \hat{I}_i$ and $I_k^{\dagger} = \sup_{i \leqslant k} \hat{I}_i$ and similarly for $\tilde{J}_k$ and $J_k^{\dagger}$.
2. We have not sought to justify the bootstrap on asymptotic grounds, and there may in fact be problems with that. Instead a number of simulation experiments have been undertaken to compare our test with others. Thus it was compared to the eight parametric tests (four of which are Lagrange multiplier tests) of Luukkonen *et al.* (1988a) for 10 different first order parametric models chosen by them and for sample sizes of 50, 100 and 200, and to two versions of the bispectrum test for seven examples considered by Ashley *et al.* (1986) for sample sizes of 250, 500 and 1000.
3. For one particular bilinear process of Luukkonen *et al.* (1988a), where the estimated conditional mean is essentially linear and the conditional variance virtually flat, so that the non-linearity does not manifest itself in terms of the non-linearity indexes we are considering, our test was without power, whereas the Lagrange multiplier test with a correct bilinear alternative performed very well. For all of the other nine examples the power of our test based on $\hat{I}_1$ or $\hat{J}_1$ was comparable and in some cases better than the most powerful of the eight parametric tests. This was true even for 50 observations, which is somewhat surprising in view of its non-parametric character.
4. For the bispectrum examples, the non-linearity typically comes in at a higher lag. In all cases the new test performed better, and in some cases considerably better, than the bispectrum tests.

For more details and some real data examples we refer to Hjellvik & Tjøstheim (1994), where we also point out some possible pitfalls of investigations of this sort.

4. Model specification, empirical identification

If a linearity test indicates non-linearity, there still remains the question of whether a parametric, restricted non-parametric or general non-parametric modelling should be used, and in the first two cases which of the several possible model classes should be employed. In addition we have the problem, relevant also for linear analysis, of which lags should be included. The difficulties of these specification problems are an order of magnitude larger than for linear models, and few formal results exist. The results of sections 4.1 and 4.2 are of a preliminary nature and much work remains to be done. In a way it is sensible to look at the lag problem first, since if we know which lags are most relevant, then the rest of the specification procedure can be concentrated to those specific lags.

4.1. Selecting significant lags

For linear models this is typically done using an FPE (final prediction error) (Akaike, 1969) or AIC type criterion, and in this case a fairly complete asymptotic theory is available. Such a criterion has also been applied to order determination and in comparisons (Priestley, 1988) involving non-linear parametric models, but in that situation its theoretical properties are less clear, as the AIC criterion requires the availability of maximum likelihood estimates and the satisfaction of standard conditions, which are problematic for some of the non-linear cases.

It is possible to approach the problem of selecting order and significant lags in a model-free or close to model-free context. The problem is then to select an "order" p and lags $i_1, \ldots, i_p$ such that $\hat{M}(X_{t-i_1}, \ldots, X_{t-i_p}) = \hat{E}(X_t \mid X_{t-i_1}, \ldots, X_{t-i_p})$ approximates X_t as well as possible, and where this selection is done for $i_k \leqslant L_1$ and $p \leqslant L_2$. Here the upper limit

L_2 on the number of allowable lags has to be small to avoid the curse of dimensionality, which is entering the problem implicitly, although we look at a functional evaluated only at the data points $\{X_t, 1 \leqslant t \leqslant n\}$. On the other hand, L_1 may be large so that it includes possible long period features. In our examples (Tjøstheim & Auestad, 1994b) with n ranging from 120 to 500 we have used $L_2 = 6$ and $L_1 = 15$ or 30.

There are several ways of approaching this problem. One is to use a generalization of the FPE or AIC criterion (Auestad & Tjøstheim, 1990, 1991; Tjøstheim & Auestad, 1994a, b). A related approach is to use cross-validation (Cheng & Tong, 1992; Yao & Tong, 1992). The point of departure taken in Auestad & Tjøstheim (1990) and Tjøstheim & Auestad (1994b) is to look at

$$\hat{R}(i_1, \ldots, i_p) = n^{-1} \sum_t (X_t - \hat{M}(X_{t-i_1}, \ldots, X_{t-i_p}))^2 w(X_{t-i_1}, \ldots, X_{t-i_p}) \tag{4.1}$$

where again w is a weight function designed to screen off extreme observations. This is the estimated prediction error based on the non-linear least squares predictor, and intuitively one would just pick the lags $i_1, \ldots, i_p$, $i_p \leqslant L_1, p \leqslant L_2$, minimizing $\hat{R}$. However, this amounts to data mining, since by choosing the bandwidth h small enough in (2.3) for the estimate of M, one can make the prediction error in (4.1) as small as one likes. This is possible since X_t is allowed to enter in the computation of its own predictor $\hat{M}(\cdot)$ in (4.1) (this is the case in the corresponding parametric situation as well, but with less disastrous consequences, since there is then no bandwidth parameter). This effect can be avoided by omitting X_t in the computation of $\hat{M}(\cdot)$ in the estimation of $E(X_t \mid X_{t-i_1}, \ldots, X_{t-i_p})$, which leads to cross-validation, or by adding a penalizing term by an FPE-type argument, where $\hat{M}$ computed from all X's is allowed to act on an independent process $\{Y_t\}$ having the same structure as $\{X_t\}$. As in the parametric case, these two approaches are related. I refer to Cheng & Tong (1992) and Tjøstheim & Auestad (1994b).

The form of the penalty factor depends on the structure of the process. If the conditional variance is constant, it is shown in Tjøstheim & Auestad (1994b) that a suitable criterion can be obtained by modifying $\hat{R}(i_1, \ldots, i_p)$ in (4.1) to

$$\widehat{FPE}(i_1, \ldots, i_p)$$

$$= n^{-1} \sum_t (X_t - \hat{M}(X_{t-i_1}, \ldots, X_{t-i_p}))^2 w(X_{t-i_1}, \ldots, X_{t-i_p}) \frac{1 + (nh^p)^{-1} J^p B_p}{1 - (nh^p)^{-1}[2K^p(0) - J^p]B_p}$$

where

$$J = \int K^2(x)\, dx, \quad B_p = n^{-1} \sum_t \frac{w^2(X_{t-i_1}, \ldots, X_{t-i_p})}{\hat{p}(X_{t-i_1}, \ldots, X_{t-i_p})}.$$

Here B_p essentially represents the dynamic range of the data in p-dimensional space and $\hat{p}(\cdot)$ is the estimated density function. For each set of lags $i_1, \ldots, i_p$ one can now choose a smoothing bandwidth h minimizing $\widehat{FPE}(i_1, \ldots, i_p)$ and then select the set of lags which, for the appropriately inserted bandwidth, minimizes $\widehat{FPE}$ as $(i_1, \ldots, i_p)$ ranges over the possible values $i_k \leqslant L_1, p \leqslant L_2$. This procedure is very computer-intensive, and we have therefore selected lags stepwise, only entering one lag at a time, and once a lag is selected, it is not allowed to be excluded again. Computational speed can be greatly improved if h is chosen automatically according to asymptotic optimality theory valid in a simpler situation. We have no guarantee that such a theory holds in our more general situation, but quite similar results to those obtained by a data driven choice of h have been obtained in Tjøstheim & Auestad (1994b).

A more general formula holds for the heterogeneous case and is given in Tjøstheim & Auestad (1994b) to which the reader is also referred for details of derivation and examples with simulated and real data. From the limited set of simulation experiments (see also Cheng & Tong, 1992 for the cross-validated case) the procedure seems to work fairly well, even in some cases where the curse of dimensionality should make it hard to obtain meaningful results. We attribute this tentatively to two factors: (1) in the functional $\widehat{FPE}$ we do not need $\hat{M}(\cdot)$ for all arguments $x_1, \ldots, x_p$, only for the observed data points; (2) there seems to be some residual information on lag structure even when $M(x_1, \ldots, x_p)$ is not very accurately estimated.

In Tjøstheim & Auestad (1994b) are also shown results for selecting significant lags with AIC-like criteria and for the conditional variance function. Although Cheng & Tong (1992) have shown the consistency of the order estimate, there are still many open theoretical as well as data-analytic problems, and some of them are pointed out in Tjøstheim & Auestad (1994b). There are obvious generalizations to the multivariate case as mentioned in Teräsvirta *et al.* (1993a).

4.2. Specification of functional form

The non-parametric and restricted non-parametric approach represent more of an omnibus approach to modelling. The specification problem is therefore primarily relevant in choosing between parametric classes, since each class is often defined to handle particular features of the data. It should be noted, however, that the boundary between parametric and restricted non-parametric models is not sharp, as witnessed for example, by the neural networks models, the MARS procedure and the STAR,LSTAR,ESTAR models of Chan & Tong (1986) and Teräsvirta (1990).

In certain situations *a priori* physical knowledge may be available which makes one model class, usually a parametric one, more natural than others. For example in Karlsen & Tjøstheim (1990), a hidden Markov chain model is a natural choice to describe the data originating from a drill hole penetrating various geological layers. The structure of the layers changes abruptly, and the sequencing and thickness of the layers are assumed to be described by a Markov chain, whereas the structure of the data within each layer can be described by an ARMA model. For certain other data there may be a natural physical threshold, e.g. a temperature threshold in river flow modelling (Tong *et al.*, 1985). In still other situations, seismic series being one of them, sudden bursts of high intensity of random length may occur.

If there is not an obvious parametric model suggested by physical factors, the technique already mentioned for selecting lags in combination with non-parametric analysis can be used as a guide to choose. Again, it must be stressed that so far we have worked with special cases, and I do not want to leave the impression that we possess a technique that solves the identification problem in general, nor does it bridge the gap in full generality between selecting significant lags and choosing an appropriate parametric model. Our method is mainly useful for additive models, though it does work for the particular threshold model (which is non-additive) given in (4.4b). Modifications will be needed for general non-additive models. A different alternative is of course to stick to one particular model class from the outset and use specification techniques designed for that specific class, see e.g. Tsay (1989) for the threshold class.

A good method of specification can hardly be based on the one-dimensional quantities $M_k(x)$ and $V_k(x)$ used in testing theory. For example, for an additive model

$$X_t = \sum_{k=1}^{p} f_k(X_{t-i_k}) + e_t \tag{4.2}$$

we do not in general have $M_{i_k}(x) = f_k(x)$. Neither do we have $M_j(x) \neq 0$ for $j \neq i_k$, $k = 1, \ldots, p$. A better quantity to use would be

$$M_{i_1, \ldots, i_p}(x_1, \ldots, x_p) = E(X_t \mid X_{t-i_1} = x_1, \ldots, X_{t-i_p} = x_p).$$

In the additive model case

$$M_{i_1, \ldots, i_p}(x_1, \ldots, x_p) = \sum_{k=1}^{p} f_k(x_k).$$

However, $M_{i_1, \ldots, i_p}(\cdot)$ cannot be displayed graphically for $p > 2$, and it only gives sums of components for an additive model, not the components itself. In Auestad & Tjøstheim (1991) and Tjøstheim & Auestad (1994a) we introduced so-called projections to circumvent this problem.

Let $i_1, \ldots, i_p$ be the lags of interest, singled out by, for example, a procedure as outlined in section 4.1, and let $\hat{M}(x_1, \ldots, x_p) = \hat{M}_{i_1, \ldots, i_p}(x_1, \ldots, x_p)$ be the kernel estimate of M computed analogously to (2.3). We then define the projector at lag i_k by

$$\begin{aligned} P_{k,w}(x_k) &= \int M(x_1, \ldots, x_p) w(x_1, \ldots, x_p)\, dF_{(k)}(x_1, \ldots, x_p) \\ &= E_{(k)}[M(X_{t-i_1}, \ldots, x_k, \ldots, X_{t-i_p}) w(X_{t-i_1}, \ldots, x_k, \ldots, X_{t-i_p})] \end{aligned} \tag{4.3}$$

where $F_{(k)}$ is the joint cumulative distribution function for $(X_{t-i_1}, \ldots, X_{t-i_{k-1}}, X_{t-i_{k+1}}, \ldots, X_{t-i_p})$ and w is a weight function. The projector $P_{k,w}(x)$ can be estimated by taking empirical averages

$$\hat{P}_{k,w}(x) = n^{-1} \sum_t \hat{M}(X_{t-i_1}, \ldots, x, \ldots, X_{t-i_p}) w(X_{t-i_1}, \ldots, x, \ldots, X_{t-i_p}),$$

and it is shown in Masry & Tjøstheim (1992) that under the regularity assumptions stated there, $\hat{P}_{k,w}(x) \xrightarrow{\text{a.s.}} P_{k,w}(x)$ for all x when w has compact support. We conjecture that asymptotic normality holds as well.

From (4.2) and (4.3) it is seen that for an additive model

$$P_{k,w}(x) = f_k(x) \int w(\mathbf{x})\, dF(\mathbf{x}_{(k)}) + \sum_{j \neq k} \int f_j(x_j) w(x_1, \ldots, x_p)\, dF_{(k)}(x_1, \ldots, x_p)$$

where $\mathbf{x}_{(k)} = [x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p]$. By letting $w(\mathbf{x}) = 1\ (\mathbf{x} \in S)$ and by letting the support S of w be large, $\int w(\mathbf{x})\, dF(\mathbf{x}_{(k)}) \to 1$ and

$$\sum_{j \neq k} \int f_j(x_j) w(x_1, \ldots, x_p)\, dF_{(k)}(x_1, \ldots, x_p) \to \sum_{j \neq k} E\{f_j(X_t)\} = E(X_t) - E\{f_k(X_t)\}$$

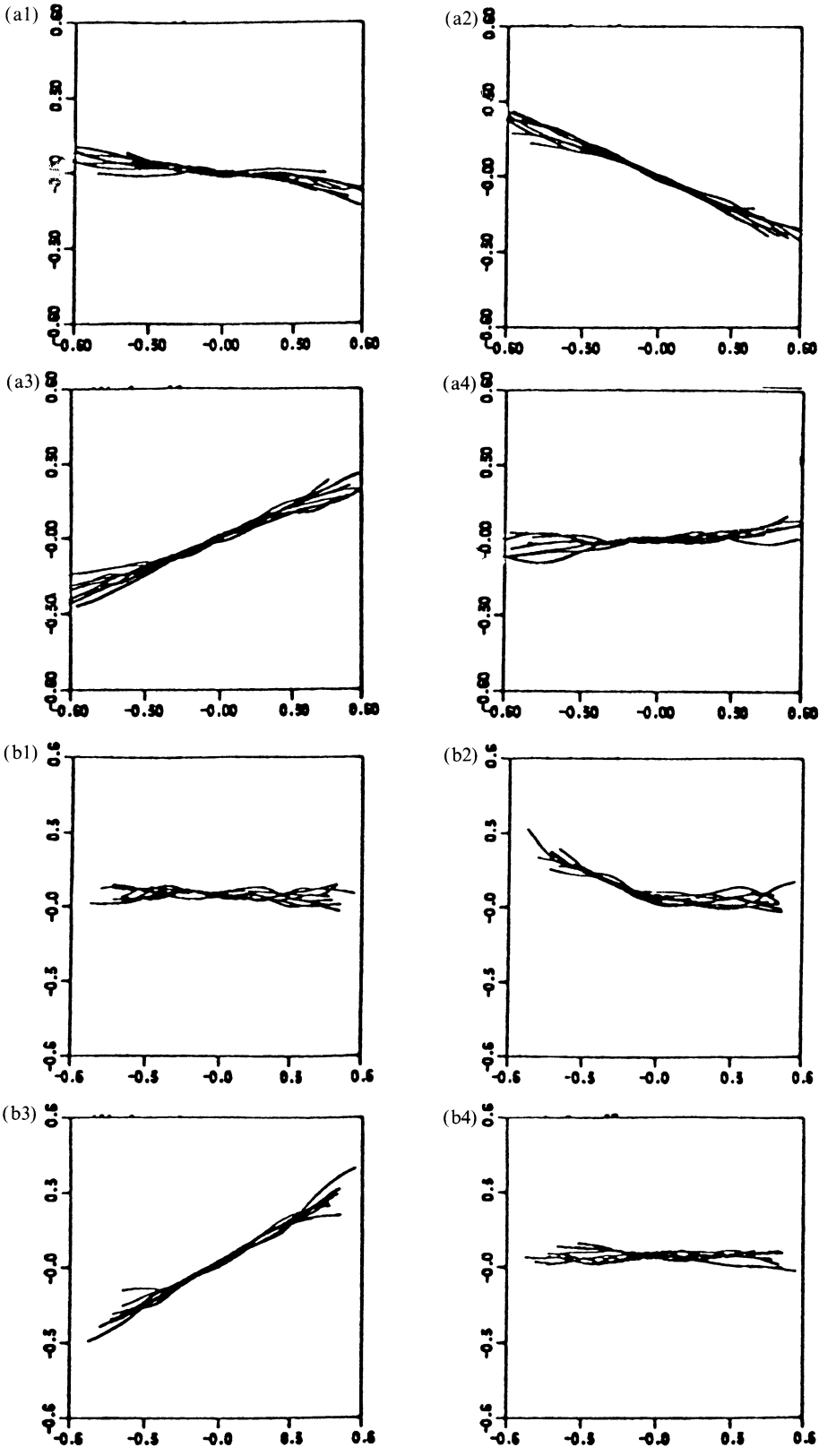
so that $P_{k,w}(x) \to f_k(x) + c$, where c is a constant independent of x.

In this way we can obtain an indication of the functional form of each $f_k(x)$ for an additive model at the significant lags. Several simulation experiments which illustrate this are given in Tjøstheim & Auestad (1994a), and for the examples tested the procedure works fairly well for a moderate number of lags. As p increases, the curse of dimensionality starts to make itself felt, but even for $p = 8$ and $p = 16$ there seems to be some residual information left for $n = 500$, although the plot is heavily biased. An example from Tjøstheim & Auestad (1994a) is displayed in Fig. 2 for the linear, threshold and exponential AR processes:

$$X_t = 0.5X_{t-6} + 0.5X_{t-10} + e_t \tag{4.4a}$$

$$X_t = (-0.5X_{t-6} + 0.5X_{t-10}) 1(X_{t-6} \leq 0) + (0.8X_{t-10}) 1(X_{t-6} > 0) + e_t \tag{4.4b}$$

$$X_t = \{0.4 - 2.0 \exp(-50X_{t-6}^2)\} X_{t-6} + \{0.5 - 0.5 \exp(-50X_{t-10}^2)\} X_{t-10} + e_t \tag{4.4c}$$



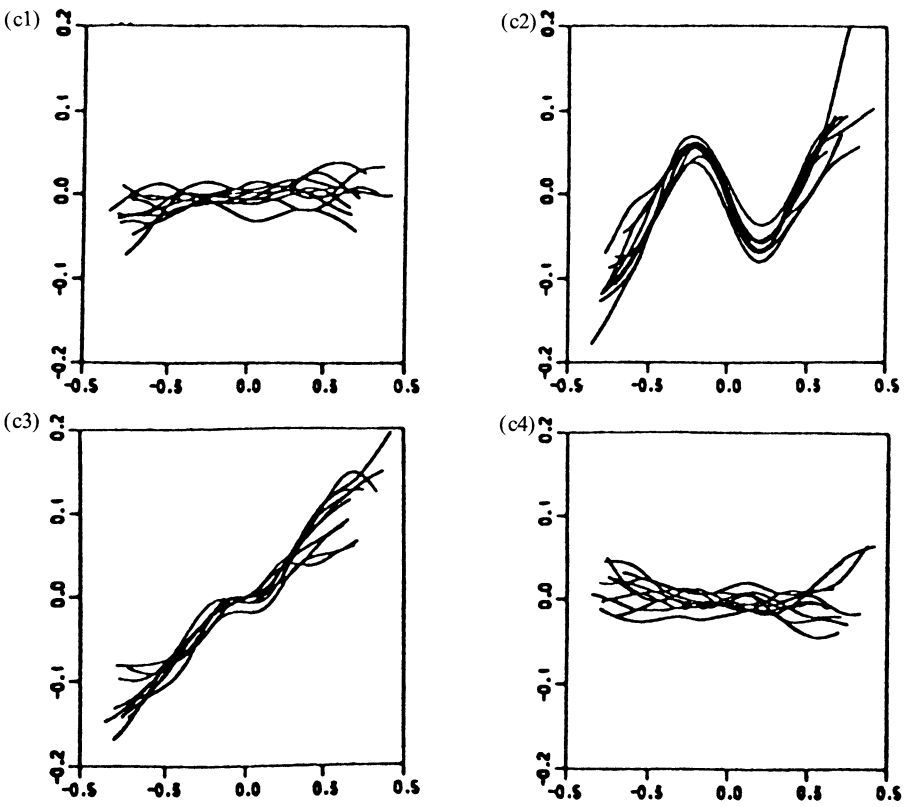


Fig. 2. (a1–a4) $\hat{P}_k(x)$ plotted against x for the linear model in (4.4a). From left to right the figures show $\hat{P}_k(x)$ for $k = 4, 6, 10, 12$, respectively. $\hat{P}_k(x)$ is estimated from 10 independent realizations each consisting of $n = 500$ observations, ($h = 0.05$). (b1–b4) $\hat{P}_k(x)$ plotted against x for the threshold model in (4.4b). From left to right the figures show $\hat{P}_k(x)$ for $k = 4, 6, 10, 12$, respectively. $\hat{P}_k(x)$ is estimated from 10 independent realizations each consisting of $n = 500$ observations ($h = 0.05$). (c1–c4) $\hat{P}_k(x)$ plotted against x for the exponential model in (4.4c). From left to right the figures show $\hat{P}_k(x)$ for $k = 4, 6, 10, 12$, respectively. $\hat{P}_k(x)$ is estimated from 10 independent realizations each consisting of $n = 500$ observations ($h = 0.05$). Note that the scale is different from (a) and (b). (a)–(c) are from Auestad & Tjøstheim (1994a).

where $\{e_t\}$ is Gaussian with var $(e_t) = 0.01$. Here the selection procedure of section 4.1 has been used to select the lags 4, 6, 10, 12 as, on the average, most significant among six ($L_2 = 6$) candidate lags. The figure displays the model features quite clearly, although there is some bias. Expressions for the bias in the linear case and a more detailed discussion are given in Tjøstheim & Auestad (1994a).

Projections can, of course, be applied to multivariate additive models as well, and some possible econometric applications in this case are discussed in Masry & Tjøstheim (1992). The procedure may also be extended to an additive conditional variance structure and to a hierarchy of interactions of additive type, but in the latter case one again has to be aware of the difficulties owing to dimensionality. It should be remarked that a similar projection technique can be used in general analysis of variance as indicated by Gu & Wabha (1992) in their paper on spline estimation.

Projectors are of little help in analysing regime models or in choosing between restricted non-parametric models. However, they may be of use as initial estimates in a backfitting algorithm such as BRUTO, which is used in additive modelling (cf. Hastie & Tibshirani, 1990; Chen & Tsay, 1993a, b).

5. Fitting and use of non-linear models

Once a model has been specified it must be fitted to data. We consider briefly the fitting of parametric models as well as those obtained using a restricted non-parametric approach.

5.1. Parametric models

The fitting in this case amounts to estimating the parameters of the model. The methods used are the traditional methods of least squares or conditional least squares, methods of moments, maximum likelihood or some quasi maximum likelihood method, often after adjustments for unknown thresholds or switch points. Which method is most appropriate depends on the model, but clearly exact maximum likelihood is very difficult to apply since marginal and joint distributions can be worked out only for very special non-linear models.

Concrete algorithms and in some cases software packages exist. For threshold processes a non-linear conditional least squares algorithm has been constructed. Examples run with this procedure can be found in Tong (1990). A recursive least squares algorithm for fitting bilinear models can be found in Subba Rao & Gabr (1984), where there are also examples. Granger & Teräsvirta (1992b) discuss and give examples of algorithms for fitting possibly multivariate data to smooth transition AR models. The fitting of residuals of linear AR processes to ARCH models is increasingly common, and estimation algorithms do exist (*Handbook of Econometrics*, Vol. 4., 1993, ch. on ARCH models).

For regime models and Kalman-type models, the EM algorithm can be used, sometimes in combination with a stochastic simulation algorithm such as the Gibbs sampler (Bølviken, 1993; Holst *et al.*, 1994). Finally, for neural net-type models the most popular algorithm is the back-propagation algorithm which uses or "passes through" the same data several times. It should be mentioned that its statistical properties are largely unknown. A treatment of neural network estimation can be found in White (1992).

Error limits on parameter estimates are supplied for most of the algorithms. Usually these are obtained by adapting asymptotic formulae to analogues of the Fisher information matrix. Sometimes such a procedure cannot be justified fully in the non-linear time series case. Another source of error is that for some models the likelihood function can be very flat in the area of interest. Such problems have been reported in particular for the parameter θ_3 in the exponential AR process of (2.1) and for the parameter θ_3 in the logistic AR process of (2.2) (cf. Chan *et al.*, 1991; Teräsvirta, 1990). It has been suggested that to some degree they can be avoided by a reparametrization of the model. On the other hand, the criterion function may be multimodal, and since the search algorithms are often non-linear and are sensitive to initial values, it may be advantageous to try several sets of initial values.

Examples of parametric model fitting and their use are given in the books by Granger & Teräsvirta (1992b), Subba Rao & Gabr (1984) and Tong (1990).

5.2. Restricted non-parametric models

The algorithms used for such models are usually iterative and recursive, and the algorithm itself is really the central piece of the modelling effort, and often *defines* the modelling process. For example, the ACE model given in (2.6) is fitted using the alternating conditional expectations (ACE) algorithm. Additive models of type (2.5) are fitted using a recursive backfitting algorithm (BRUTO) described in Hastie & Tibshirani (1990). It requires initial estimates of the components. It can be extended to allow for additive interactions. A version of this modelling, essentially replacing $f_i(X_{t-i})$ by $g_i(X_{t-i})X_{t-i}$ has been used by Chen &

Tsay (1993a, b) both for univariate and multivariate data. In particular they have modelled the chicken pox data of Sugihara & May (1990) and obtained better forecasts.

In MARS models, a recursive fitting of products of left and right truncated linear splines is used. The technology appears very effective in dealing with high dimensional data and is described in detail in Friedman's (1991) paper. Lewis & Stevens (1991a) have applied the procedure to time series problems. In particular, they have compared prediction capability on sunspot data with the non-linear threshold and bilinear model as well as the linear model. A number of MARS models were fitted by varying the input parameters of the model, and many of them incorporated both bivariate and trivariate interaction terms, in addition to the purely additive linear truncated terms. For the best MARS model superior prediction performance was observed. In particular, the improvement was dramatic for forecast in the range from 6 to 12 time periods.

Most of the models discussed in this subsection contain a mixture of estimated parameters and functions. For some of them, MARS being the most prominent example, there are no methods currently of computing error limits for estimated quantities. To assess these models as a group is not easy. They are new, and only limited empirical evidence is available when it comes to comparison individually and as a general alternative to the group of non-linear parametric models. In particular, the MARS models appear very promising, and important new developments are under way (Breiman, 1992) in this general field. One should note that for restricted non-parametric models, convergence properties of the algorithms used could sometimes be a problem. Also, the danger of overfitting the model in general seems to be larger than for parametric models.

6. Theory of estimation, asymptotics

Asymptotics and estimation theory are quite well developed in the parametric and purely non-parametric case. To my knowledge there is little asymptotic theory for the restricted non-parametric case (see Hastie & Tibshirani, 1990, however).

6.1. Parametric models

Estimation theory is extensive for linear models, and a good account is given in Brockwell & Davis (1987). For non-linear models it is both more fragmentary, and it is more difficult to apply in practice.

A general theory for conditional least squares estimates in a non-linear situation is outlined in Klimko & Nelson (1978). It is carried over to the time series case in Tjøstheim (1986b) and Tong (1990, ch. 5). Both consistency, the asymptotic distribution and the law of the iterated logarithm are considered. The theory can be formulated in terms of a criterion function, e.g. a conditional least squares or a maximum likelihood criterion $Q_n(\boldsymbol{\theta})$ depending on the unknown parameter vector $\boldsymbol{\theta}$. Then to obtain consistency (Tjøstheim, 1986b) one can require ($\boldsymbol{\theta}_0$ is the true value)

A1: $n^{-1}\,\partial Q_n(\boldsymbol{\theta}_0)/\partial\theta_i \xrightarrow{\text{a.s.}} 0$,

A2: the matrix $\{\partial^2 Q_n(\boldsymbol{\theta}_0)/\partial\theta_i\,\partial\theta_j\}$ is almost surely non-negative definite and $\lim_{n\to\infty}\lambda^n_{\min}(\boldsymbol{\theta}_0) \overset{\text{a.s.}}{>} 0$, where $\lambda^n_{\min}(\boldsymbol{\theta}_0)$ is the smallest eigenvalue of $n^{-1}\,\partial^2 Q_n(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}^2$,

A3: $\lim_{n\to\infty}\sup_{\delta\downarrow 0}(n\delta)^{-1}\left|\partial^2 Q_n(\boldsymbol{\theta}^*)/\partial\theta_i\,\partial\theta_j - \partial^2 Q_n(\boldsymbol{\theta}_0)/\partial\theta_i\,\partial\theta_j\right| \overset{\text{a.s.}}{<} \infty$ for $|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0| < \delta$.

Then there *exists* an estimator $\hat{\boldsymbol{\theta}}_n \xrightarrow{\text{a.s.}} \boldsymbol{\theta}_0$.

A1 essentially guarantees that asymptotically $\mathbf{Q}_n$ has an extreme point at $\boldsymbol{\theta}_0$, A2 that this is a maximum, and A3 that the remainder term of the Taylor expansion can be ignored.

To obtain asymptotic normality a sufficient set of additional requirements is that

B1: $n^{-1}\,\partial^2 Q_n(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}_i\,\partial\boldsymbol{\theta}_j \xrightarrow{\text{a.s.}} V_{ij}$,

B2: $n^{-\frac{1}{2}}\,\partial Q_n(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta} \xrightarrow{d} \mathcal{N}(0, W)$

for some positive definite matrix V and W. Then if A1–A3 and B1–B2 are fulfilled,

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, V^{-1}WV). \tag{6.1}$$

In Tjøstheim (1986b) these general results are applied to a range of time series models. A similar approach is taken in Tong (1990). The theory is largely limited to the stationary case, but I would like to point out that a much more general theory can be constructed using extensions of LeCam theory, where in general non-ergodic processes are allowed, and asymptotic normality does not obtain; at least not without a random scaling. I refer to the monograph by Basawa & Scott (1983). General non-linear parametric dynamic systems in an econometric context have been considered by Gallant (1987) and Pötscher & Prucha (1991a, b).

There are some common difficulties with all of these approaches. Conditions such as A1–A3 and B1–B2 and the more general and complicated conditions in the above references often yield existence results only; i.e. the existence of a consistent estimator. Unless a unique solution of the minimization problem exists, reservations concerning the practical interpretation must be made (cf. Billingsley, 1961, p. 10). Moreover, for specific models the conditions are often difficult to verify (Tjøstheim, 1986b), and expressions for standard errors based on (6.1) may be hard to estimate and compute. An added difficulty is that sometimes standard errors are quite large, and proving consistency does not mean much for the sample sizes one is likely to encounter in practice (cf. Karlsen & Tjøstheim, 1988). In general it seems to be fair to state that there is a considerable gap between general estimation theory and practical applications.

It is perhaps somewhat easier to develop an estimation theory if the scope is restricted to particular classes of non-linear parametric time series. Thus, for RCA models, which admittedly is a rather small extension of the linear class, the estimation theory is close to the linear one and quite complete (Nicholls & Quinn, 1982). For threshold models it is more complicated. However, several contributions to the theory are available through the work of Chan (1990, 1991), Chan & Tong (1986, 1990) and others. Bilinear models are problematic to analyse primarily owing to difficulties in expressing the residuals in terms of previous Xs (the invertibility problem) and the problem of giving conditions guaranteeing the existence of moments of sufficiently high order. No complete theory exists for the most general model. For more specialized models there are contributions from Grahn (1992), Guegan & Pham (1989) and Pham (1985).

Econometricians are particularly interested in ARCH models, and an estimation theory has been developed. Recent references are Bera & Higgins (1993) and *Handbook of Econometrics*, Vol. 4 (1993, ch. on ARCH models). Special problems arise in the estimation of regime models because of difficulties in estimating the unknown random change points. Algorithms for this exist, but the theoretical properties of the corresponding estimates are difficult to pinpoint in general. Method of moments estimators were considered for a very special model in Tyssedal & Tjøstheim (1988). Consistency is shown, but the errors are very large, and exceedingly large sample sizes are required to get decent accuracy. Much better results were obtained using an informal stepwise least squares estimator, but we were unable to work out an asymptotic theory for it. For other and related advances in this field we refer to Holst *et al.* (1994).

Bootstrap estimates and bootstrap theory for non-linear parametric models are not well developed, although it may be better suited to assess error limits of parameter estimates than some of the procedures based on asymptotic theory. The parameter estimates will typically be based on a statistic whose components depend on a finite number of Xs jointly, so that the blockwise bootstrap (Künsch, 1989) can be applied. See also Bühlman (1992). For non-linear processes the blockwise bootstrap may be better suited than the boostrap based on innovations, because of the difficulties of computing these in a general case. A third possibility is a bootstrap based on the Markov property.

A good deal of work has been done for bootstrap estimates for linear models (Findley, 1986; Kreiss & Franke, 1992), but the comparison between procedures using asymptotic theory and those using the bootstrap to assess error limits is somewhat inconclusive. Owing to larger difficulties in applying asymptotic theory for non-linear models, I think that the potential of the bootstrap is greater here.

6.2. Restricted non-parametric and semi-parametric estimates

The difficulties in establishing asymptotic theory of some of the algorithms involved have been mentioned already. Possibly this could be used as a general argument against this type of modelling. It would certainly be desirable to establish theoretical properties, but from a pragmatic point of view they should be judged from the accumulated evidence obtained concerning their capability to model specific features of the data and in making forecasts.

The situation is different for the semi-parametric modelling option briefly mentioned in section 2.3. An asymptotic theory can usually be carried through both for the parametric and non-parametric part. Under quite general assumptions it is possible to obtain $\sqrt{n}$-consistency for the parametric part as demonstrated by Heckman (1986) and Robinson (1988). Powell *et al.* (1989) have developed the theory further and given econometric applications. Another recent contribution comes from Truong & Stone (1992).

6.3. Non-parametric estimates

Unlike the restricted non-parametric case, the asymptotic theory of general non-parametric estimation is well-developed. Actually, in a way the asymptotic theory of the non-parametric estimates is easier than the corresponding theory for the non-linear parametric case. This is due to the fact that non-parametric estimates depend on the data in a simpler manner.

A variety of regularity conditions exist for the validity of the asymptotic theory of the kernel estimated conditional mean, probability density and the conditional variance. Results exist for asymptotic normality, weak and strong consistency, rates and uniformity of convergence. Regularity conditions typically involve assumptions on mixing of $\{X_t\}$, which can be verified in the Markov case via the theory for geometric ergodicity as indicated in section 2.4, some smoothness assumptions on the estimated objects $M(x), p(x), V(x)$ and conditions on the rate at which the bandwidth parameter h tends to zero. For an explicit listing of conditions we refer to Robinson (1983). See also Masry & Tjøstheim (1992). Under these conditions and in the notation of section 3 we have

$$\sqrt{nh}(\hat{M}_k(x) - M_k(x)) \xrightarrow{d} \mathcal{N}\left(0, \frac{V_k(x)}{p(x)} \int K^2(z)\, dz\right) \tag{6.2}$$

and

$$\sqrt{nh}(\hat{V}_k(x) - V_k(x)) \xrightarrow{d} \mathcal{N}\left(0, \frac{W_k(x)}{p(x)} \int K^2(z)\, dz\right) \tag{6.3}$$

where $W_k(x) = E[\{X_t - M_k(x)\}^4 \mid X_{t-k} = x] - V_k^2(x)$ with similar expressions being true in the p-dimensional case, but with nh replaced by nh^p. In practice, one has to be careful with the application of (6.2) and (6.3) because of bias and boundary effects. We are also interested in properties of estimates of projections and other non-parametric functionals such as the one used in linearity testing and the one to be described for independence testing in section 7. The theory is less complete for such functionals, but U-statistic theory is often useful. References discussing asymptotic theory of U-statistics in the context of non-parametric estimation are Carlstein (1988), Denker & Keller (1983), Hall (1984), Skaug & Tjøstheim (1993b) and Yoshihara (1976). As indicated in section 3 such a theory may be of limited practical applicability.

A very important problem in non-parametric estimation and its application is the choice of bandwidth h. Unfortunately, there is no optimality theory available in the context we are discussing, and we have therefore adopted approaches whose properties have been established in a simpler framework. From a practical point of view there are basically three procedures available. The simplest solution is to compute estimates for several values of h and select one of them subjectively. In that manner we will also get an idea of how sensitive the statistic is to changes in bandwidth. Actually, the linearity test of section 3, the order determination procedure of section 4 and the independence test to be described in section 7 all seem to be fairly robust to changes in h-values over quite a wide region. A second possibility is to use asymptotic theory developed in a simpler situation. If we require variance and bias squared to be asymptotically balanced, then $n^{-1}h^{-p} \sim h^4$ or $h \sim n^{-1/(p+4)}$. An explicit form for the proportionality factor is given in Silverman (1986) for the special case of multidimensional density estimation with i.i.d. Gaussian observations. It seems to work reasonably well also in our framework, and we have used it systematically for the linearity and independence test. The third possibility, which is the most time consuming, but possibly also the one most used in practice, is to use some form of cross-validation or penalty factor. We have used this method in the order selection problem with an FPE penalty factor. The method shows considerable variation in the selection h for one and the same model.

6.4. Can asymptotic theory be used in practice?

For linear parametric models asymptotic theory generally works quite well in the range of 100 observations or so. For non-linear parametric models it is more difficult to use asymptotic theory to obtain confidence intervals and tests, but in many simpler cases (cf. Tong, 1990 and others) for a sample size of 200–500, sensible results can be obtained. In models containing several structures, which is typically the case for regime models, considerably more observations are needed (Karlsen & Tjøstheim, 1988). For non-parametric estimates and functionals the situation is even more problematic. For a high dimensional quantity the curse of dimensionality may require extremely large numbers of observations to come anywhere close to asymptotic results. For the functionals used in test linearity and independence the standard error typically is of the same order ($n^{-1/2}$) or even less than in the parametric case, but the smoothing enters in a complicated way so that unlike the parametric case, where the next term in the asymptotic expansion is of order n^{-1}, the next terms are almost of the same order as the leading term. This means that several terms of the asymptotic expansion should really be included, but except for simple functionals it is very difficult to obtain explicit expressions, not to mention estimates, for these.

This, in addition to our practical experiences for the linearity and independence test, points to the bootstrap or other resampling techniques as an alternative to purely asymptotic methods. Based on evidence thus far, it is tempting to claim that resampling is much more

essential in non-parametric time series settings than in parametric ones. Currently this claim cannot be backed up theoretically in terms of asymptotic theory (or perhaps one should rather look at finite sample properties) of the bootstrap. It is merely based on the fact that the bootstrap works better in several quite different situations studied by us.

7. Model verification

For a given model we may be interested in checking whether it is stationary, invertible and whether its residuals are i.i.d.

7.1. Stationarity and invertibility

Criteria for stationarity have already been mentioned in section 2.4. Formally these criteria can, of course, be applied to an estimated model as well, with unknown functions and parameters replaced by estimated ones. Conditional on the given realizations, if we feed "new" innovations into the model, it will settle in a stationary state if it satisfies one of the stationarity criteria. For example if $X_t = f(X_{t-1}) + e_t$ and f is estimated non-parametrically from $[X_1, \ldots, X_n]$, and we then consider

$$Y_t = \hat{f}_{[X_1, \ldots, X_n]}(Y_{t-1}) + e'_t, \quad t > 2, \quad Y_1 = y_1 \tag{7.1}$$

where $\{e'_t\}$ is an i.i.d. process independent of $\{e_t\}$, then the process $\{Y_t\}$ is geometrically ergodic conditionally on $[X_1, \ldots, X_n]$ being given if $\hat{f}_{[X_1, \ldots, X_n]}(\cdot)$ and $\{e'_t\}$ satisfy the criteria cited in section 2.4.

It should be noted that sometimes the estimation method forces a stationary structure on the process even though the process itself may not be stationary. Thus, for an arbitrary zero-mean process $\{X_t\}$, the Yule–Walker estimate

$$\hat{a} = \frac{\sum_{t=2}^{n} X_t X_{t-1}}{\sum_{t=1}^{n} X_t^2}$$

will always produce a stationary process $\{Y_t\}$ if $\{Y_t\}$ is generated as

$$Y_t = \hat{a}_{[X_1, \ldots, X_n]} Y_{t-1} + e'_t.$$

Hence, the stationarity may be a product of the estimation method rather than the structure of the process itself. It is well known that if the Yule–Walker estimates are replaced by least squares estimates, then the above property disappears.

Another way of testing for stationarity is just to feed i.i.d. innovations into the estimated model according to (7.1) and then observe whether the model blows up or not. Such a method is entirely informal and non-rigorous. For example, the process may be stationary but nevertheless produce bursts that may look like non-stationarities. As a "rough and ready" method the procedure is of considerable value.

The property of invertibility amounts to requiring that the innovations $\{e_t\}$ for each t should be measurable with respect to $\{X_s, s \leqslant t\}$. To be able to use the invertibility in practice in computing conditional likelihoods or forecasting, say, it must be required that e_t can be expressed in a convergent (in some sense) expansion of $\{X_s, s \leqslant t\}$. Usually moments of a certain order of such an expansion must also be required to exist. The problems associated with establishing such expressions and proving existence of moments can be very hard, as exemplified in particular by the bilinear model, where there are only scattered results

(Guegan & Pham, 1987; Quinn, 1982). There seems to be no general theoretical criterion for checking invertibility comparable to the Tweedie criterion. For non-linear pure AR models the problem of invertibility is, of course, absent.

7.2. *Test of model fit, residuals, independence tests*

One view of model fitting is that it consists in trying to find a transformation F of the data which reduces the transformed data points $e_t = F(X_t, X_{t-1}, \ldots, X_{t-p})$ to i.i.d. random variables. For ARMA models F is linear, and p is allowed to tend to infinity if there is an MA part. With such an approach one needs to test the residuals for the i.i.d. property.

For linear time series, such a test makes up the third stage in the specification–estimation–verification cycle, as it would for a general non-linear model. The most commonly used test to check independence of residuals is the Box–Ljung statistic, which for residuals $\{\hat{e}_t\}$ is given by the portmanteau statistic

$$Q_n = n(n+2) \sum_{j=1}^{k} \hat{\rho}_{\hat{e}}^2(j) \Big/ (n-j) \tag{7.2}$$

where

$$\hat{\rho}_{\hat{e}}(j) = \frac{\sum_t (\hat{e}_{t+j} - \bar{\hat{e}})(\hat{e}_t - \bar{\hat{e}})}{\sum_t (\hat{e}_t - \bar{\hat{e}})^2}$$

The test has been used to test model fit both for linear and non-linear models. Asymptotically it can be shown that if the estimated residuals stem from an ARMA (p, q) model with p and q known, and when the parameters are estimated by a maximum likelihood algorithm, then (Brockwell & Davis, 1987, ch. 9.4) asymptotically Q_n is distributed as a χ^2_{k-p-q}-variable.

This test is really not a test of independence, but rather of uncorrelatedness, and it is clear, therefore, that it will have very low power against residuals that are uncorrelated but dependent. There is at least one important class of processes where this is exactly what happens, namely the ARCH processes. Consider for example the ARCH process

$$X_t - \sum_{i=1}^{p} a_i X_{t-i} = e_t$$

$$e_t = \left(1 + \sum_{i=1}^{q} b_i e_{t-i}^2\right)^{1/2} \eta_t$$

where $b_i \geqslant 0, i = 1, \ldots, q$ and $\{\eta_t\}$ is a sequence of i.i.d. variables. Then the residuals $\{e_t\}$ are uncorrelated, but at the same time, if at least one $b_i > 0$, dependent; a fact which will be missed by the correlation test. An *ad hoc* procedure to correct for this is to consider instead squared residuals $\{\hat{e}_t^2\}$, which are correlated, and insert *them* in the statistic (7.2). This is essentially the idea of the McLeod–Li (1983) test of independence.

Motivated by these problems we have recently (Skaug & Tjøstheim 1993a, b, c) looked at other tests of serial independence, and we will close this paper with a review of those papers and some related results. We stress that thus far we have been mainly concerned with testing of independence as such. Possible modifications of our tests to apply them more specifically to independence tests of *estimated residuals* have not been undertaken, but we do believe that the tests will be appropriate in such situations by applying similar reasoning to that used for the BDS test (Brock *et al.*, 1991b). This test is also really intended as a general test of serial

independence, but is now in widespread use as a test of independence of residuals. The BDS test is based on the correlation integral of chaos theory; i.e.

$$C_{m,n}(\varepsilon) = \binom{n}{2}^{-1} \sum_{1 \leq s < t \leq n} 1(\|\mathbf{X}_s^m - \mathbf{X}_t^m\| < \varepsilon)$$

where $\mathbf{X}_s^m = [X_s, \ldots, X_{s+m-1}]$ and $\|\mathbf{X}_s^m - \mathbf{X}_t^m\| = \max_{0 \leq i \leq m-1} |X_{s+i} - X_{t+i}|$. Under the assumption that the X_ts are i.i.d., $C_m(\varepsilon) = C_1^m(\varepsilon)$, where $C_m(\varepsilon) = \lim_{n \to \infty} C_{m,n}(\varepsilon)$. This equality is the relationship that is tested. Properties of the test are treated in several publications (Brock *et al.*, 1987, 1991a, b; Brock & Potter, 1992).

The idea behind the papers of Skaug & Tjøstheim is that if X_t consists of i.i.d. variables, then X_t and X_{t-k} are independent for all $k \neq 0$. If a density function p exists this means

$$p_{X_t, X_{t-k}}(x_1, x_2) \doteq p_k(x_1, x_2) = p_{X_t}(x_1) p_{X_{t-k}}(x_2) \doteq p(x_1)p(x_2) \tag{7.3}$$

for all x_1 and x_2, whereas in general we have that the distribution function F satisfies

$$F_k(x_1, x_2) = F(x_1)F(x_2). \tag{7.4}$$

The task, then, is to construct functionals measuring the differences between $\hat{p}_k(x_1, x_2)$ and $\hat{p}(x_1)\hat{p}(x_2)$ or between $\hat{F}_k(x_1, x_2)$ and $\hat{F}(x_1)\hat{F}(x_2)$. In Skaug & Tjøstheim (1993a, c) density functions are considered, while we look at distribution functions in Skaug & Tjøstheim (1993b). It is obvious that this approach can be extended to higher order joint densities and empirical distribution functions, but it is also evident that the curse of dimensionality comes into play fairly rapidly, and to us it seems that a more practical approach will be to let k vary in (7.3) and (7.4) and consider combinations with various k-values. This is analogous to our approach to the test for linearity, and more detailed work on this aspect is now in progress for the independence test (Skaug & Tjøstheim, 1993c).

It should be noted that a different non-parametric approach to independence testing consists in using rank statistics. We refer to the survey paper by Hallin & Puri (1992).

7.2.1. Tests based on estimated densities

For ease of notation we take $k = 1$. Tests based on functionals measuring differences between $\hat{p}_1(x_1, x_2)$ and $\hat{p}(x_1)\hat{p}(x_2)$ have been considered very recently and several types of functionals have been studied:

The entropy functional

$$I_1 = \int_S \log \frac{p_1(x_1, x_2)}{p(x_1)p(x_2)} p(x_1, x_2)\, dx_1\, dx_2$$

where S is a suitably chosen region of integration, was considered by Joe (1989) and Robinson (1991), although Robinson modified it somewhat. The absolute value functional

$$I_2 = \int |p_1(x_1, x_2) - p(x_1)p(x_2)|\, dx_1\, dx_2$$

was studied by Chan & Tran (1992), and its power was investigated for some examples using simulations. Rosenblatt (1975) and Rosenblatt & Wahlen (1993) looked at the squared difference functional

$$I_3 = \int (p_1(x_1, x_2) - p(x_1)p(x_2))^2\, dx_1\, dx_2.$$

They derived asymptotic results in a situation where the objective was to test independence between two random variables X and Y given i.i.d. observations of both of them. In Skaug & Tjøstheim (1993a, c) we consider these functionals and the weighted difference functional

$$I_4 = \int (p_1(x_1, x_2) - p(x_1)p(x_2))p_1(x_1, x_2)\, dx_1\, dx_2. \tag{7.5}$$

At first sight I_4 looks rather silly. Whereas we have $I_i \geqslant 0$ for $i = 1$, 2 and 3 and $I_i = 0$ in the independent case, this does not hold for I_4. In fact, we can find examples (Skaug & Tjøstheim, 1993a) where we have dependence but $I_4 \leqslant 0$. Nevertheless, I_4 has performed well compared to the other functionals in a series of simulation experiments. An intuitive reason for this is that if $p_1 > p^2$, then the weight factor p_1 is large, whilst if $p_1 < p^2$, then the weight factor is small, so that we tend to end up with a positive I_4 in the dependent case. In the Gaussian case this can be formalized, and it can be shown that then $I_4 \geqslant 0$, and $I_4 = 0$ iff $\{X_t\}$ consists of i.i.d. variables.

By inserting kernel density estimates in (7.5) and introducing a weight function w having compact support and taking empirical averages, we have

$$\hat{I}_4 = n^{-1} \sum_t \{\hat{p}_1(X_t, X_{t-1}) - \hat{p}(X_t)\hat{p}(X_{t-1})\} w(X_t, X_{t-1}).$$

Under weak regularity conditions (Skaug & Tjøstheim, 1993a, c) $\hat{I}_4$ is consistent and asymptotically normal. In general we have used leave-one-out estimators for $p(X_t)$. Then under the null hypothesis of independence

$$E(\hat{I}_4) = n^{-1}\left\{2\left(\int p^2(x)w(x)\, dx\right)^2 - \int p^3(x)w(x)\, dx\right\} + O(n^{-1}h) + O(n^{-2}h^{-2})$$

and

$$\text{var}\,(\hat{I}_4) = n^{-1}\left(\int p^3(x)w(x)\, dx - \left(\int p^2(x)w(x)\, dx\right)^2\right)^2 + O(n^{-1}h^2) + O(n^{-2}h^{-2}).$$

It is seen that the leading term of the expansion for the variance is of the same order as in a parametric estimation problem, and this rate is the same under H_A. For the other functionals similar results are obtained, but the rates can be different under H_0 and H_A (cf. also the linearity test). For example the dominating term of the variance for $\hat{I}_3$ is of order $n^{-2}h^2$ under H_0 and n^{-1} under H_A.

An important difference between the asymptotic expansion of $\hat{I}_4$ and the corresponding standard parametric case is that the difference between the leading term and the next term is much smaller in order. A similar problem occurred for linearity testing and will also be true in general for the other functionals $\hat{I}_1 - \hat{I}_3$ (cf. section 6.4). This means that using a test depending only on the leading term of an asymptotic expansion for $\hat{I}_i$ may be inaccurate for moderate sample sizes, and indeed this was found to be the case. For simulations of Gaussian i.i.d. random variables the real (simulated) level of the test at a nominal level of 5% was almost twice as large. Considerably more accurate results were obtained by setting the mean of $\hat{I}_4$ equal to zero in the null situation and by bootstrapping the standard deviation in the asymptotic normal distribution. For 100 bootstrap replicas this leads typically to a computation time of a couple of minutes on an HP750 workstation. For more details see Skaug & Tjøstheim (1993a), where we also present simulation experiments comparing the correlation functional derived from (7.2) to several versions of $\hat{I}_1 - \hat{I}_4$. The correlation functional is optimal and outperforms the others for a Gaussian AR(1) process, but already for an AR(1) process with exponential residuals the difference is much smaller, and the correlation

functional is outperformed for a bilinear and uncorrelated, but dependent, example. For all of these examples we only consider lag 1 dependence. More general cases and a comparison with the BDS test are contained in Tjøstheim & Skaug (1993c).

As a standard in these examples we chose the bandwidth (the process is normalized so that its empirical variance is equal to one) as $h = n^{-1/6}$ to balance bias squared and variance in the bivariate case. There was evidence in Skaug & Tjøstheim (1993a) that this bandwidth is a bit low, but simulations indicated that once h gets large enough our results are fairly insensitive to changes in h.

7.2.2. Tests based on the empirical distribution function

In this case we have used a squared difference functional

$$I_F = \int (F_1(x_1, x_2) - F(x_1)F(x_2))^2 \, dF_1(x_1, x_2).$$

The idea of using such a functional for testing independence between two random variables X and Y in a non-time series setting goes back at least to Hoeffding (1948) and has been considered further by Blum *et al.* (1961) who developed the asymptotic theory in this framework. As far as I know, it has not been analysed, although a formal extension is entirely obvious, in a time series setting with testing of serial independence.

As a corresponding estimated functional at lag k we have used

$$\hat{I}_{F,k} = n^{-1} \sum_t \{F_n(X_{t-k}, X_t) - F_n(X_{t-k}, \infty)F_n(\infty, X_t)\}^2$$

where the two-dimensional empirical distribution function is estimated by

$$F_n(x_1, x_2) = (n-k)^{-1} \sum_{s=k+1}^{n} 1(X_{s-k} \leqslant x_1) 1(X_s \leqslant x_2).$$

In this case the asymptotic distribution is not normal, which is not surprising since the leading term in the asymptotic expansion is an n-independent degenerate U-statistic. Adapting the theory of Carlstein (1988) for strongly mixing processes in such a situation, instead

$$n\hat{I}_{F,k} \xrightarrow{d} \sum \lambda_{ij} W_{ij}^2$$

where the W_{ij} are i.i.d. $\mathcal{N}(0, 1)$ variables and the λ_{ij} are eigenvalues of the eigenvalue problem for the integral operator generated by the kernel function for the U-statistic. For low values of k in this case there is a good correspondence between asymptotic theory and simulations even for moderate sample sizes of $n = 50$. It is tempting to attribute this to the absence of a smoothing bandwidth. On the other hand, as k increases, the agreement between asymptotic theory and the simulations deteriorates. This also influences the functional

$$G_K = \sum_{k=1}^{K} I_{F,k} \tag{7.6}$$

as K increases, and thus for moderate sample sizes and moderately large values of K it may again be advisable to introduce the bootstrap. The BDS test is also based on empirical distribution functions, and in that case large sample sizes are required to get good correspondence between the nominal and simulated size of the test, when the asymptotic distribution (normal) for that case is used.

We have compared the power of I_F to the correlation functional and to the "best" of the tests considered in 7.2.1 for three first order processes (AR, ARCH, and threshold AR). The

conclusion from these and similar experiments was (cf. Skaug & Tjøstheim, 1993b for more details) that for linear processes and weak linearities I_F comes close to the correlation functional in power properties, and it is better than the functionals of section 7.2.1, based on density function estimates. For stronger non-linearities, however, it was outperformed by the functionals of section 7.2.1, and sometimes even by the correlation functional. Generally, with the exception of the entropy functional, all of these functionals are *not* transformation invariant, and for some non-linearities their power increased considerably by computing them for $\{X_t^2\}$ instead of $\{X_t\}$, but this leads to a loss of power against a linear alternative.

We also conducted a few preliminary experiments for processes having their dependence structure at higher lags. As can be expected, the power is sensitive to the choice of K in (7.6). We tried a functional based on comparing the K-dimensional joint empirical distribution with one consisting of products of K one-dimensional factors. Its performance was inferior to that of G_K.

Acknowledgements

The original draft of this paper was written while I was visiting the Institut für Angewandte Mathematik, University of Heidelberg, in July 1992. I am very grateful to the institute and in particular to Professor Rainer Dahlhaus for providing an ideal working environment and for interesting discussions. I thank two referees and an associate editor for a number of constructive and very useful comments. I am indebted to Vidar Hjellvik for the plots of Fig. 1. Finally, I thank Professor Jostein Lillestøl for his invitation to give a talk at the Nordic Meeting at Røros 1992. The paper is based directly on that talk.

References

Akaike, H. (1969). Fitting autoregressions for predictions. *Ann. Inst. Statist. Math.* **21**, 243–247.

Ashley, R. A., Patterson, D. M. & Hinich, M. N. (1986). A diagnostic test for nonlinear serial dependence in time series fitting errors. *J. Time Ser. Anal.* **7**, 165–178.

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

Auestad, B. & Tjøstheim, D. (1991). Functional identification in nonlinear time series. In: *Nonparametric functional estimation and related topics* (ed. G. Roussas). Kluwer Academic, Amsterdam, pp. 493–507.

Basawa, I. V. & Scott, D. J. (1983). *Asymptotic optimal inference for non-ergodic models.* Lecture Notes in Statistics, **17,** Springer, New York.

Bera, A. K. & Higgins, M. L. (1993). A survey of ARCH models: properties, estimation and testing. *J. Econom. Surveys* (to appear).

Billingsley, P. (1961). *Statistical inference for Markov processes.* Chicago University Press, Chicago.

Blum, J. R., Kiefer, J. & Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *Ann. Math. Statist.* **32,** 485–498.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *J. Econom.* **31,** 307–327.

Bølviken, E., Storvik, G., Nilsen, D. E., Siring, E. & van der Wel, D. (1992). Automated prediction of sedimentary facies from wireline logs. In *Geological applications of wireline logs* II (eds A. Hurst, C. M. Griffiths & P. F. Worthington). Royal Geological Society, London, pp. 123–139.

Bølviken, E. (1993). Statistical methods using regime processes and other non-linear state space models. Unpublished manuscript, Department of Mathematics, University of Oslo.

Bougerol, P. & Picard, N. (1992). Stationarity of GARCH processes and some non-negative time series. *J. Econometrics* **52,** 115–127.

Box, G. E. P. & Jenkins, G. M. (1970). *Time series analysis, forecasting and control.* Holden Day, San Francisco.

Breidt, F. J. & Davis, R. A. (1991). Time-reversibility, identifiability and independence of innovations for stationary time series. Preprint, Department of Statistics, Colorado State University, Fort Collins.

Breiman, L. (1992). Hinging hyperplanes. Paper presented at the 24th Interface Symposium, Computing Science and Statistics, College Station, Texas.

Breiman, L. & Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.* **80,** 580–619.

Breiman, L., Friedman, J. H., Olshen, R. & Stone, C. J. (1984). *Classification and regression trees.* Wadsworth, Belmont CA.

Brock, W. A., Dechert, W. D. & Scheinkman, A. J. (1987). A test for independence based on the correlation dimension. Working paper, University of Wisconsin-Madison, Social Systems Research Institute.

Brock, W. A., Hsieh, D. A. & LeBaron, B. (1991a). *Nonlinear dynamics, chaos and instability.* MIT Press, Cambridge, MA.

Brock, W. A., Dechert, W. D., Scheinkman, J. A. & LeBaron, B. (1991b). A test for independence based on the correlation dimension. Manuscript.

Brock, W. A. & Potter, S. (1992). Diagnostic testing nonlinearity, chaos and general dependence in time series data. In: *Nonlinear modelling and forecasting* (eds M. Casdagli & S. Eubank). Addison Wesley, Redwood City, CA.

Brockett, P. L., Hinich, M. J. & Patterson, D. (1988). Bispectral-based tests for the detection of Gaussianity and linearity in time sieries. *J. Amer. Statist Assoc.* **83,** 657–664.

Brockwell, P. & Davis, R. A. (1987). *Time series. Theory and methods.* Springer, New York.

Bühlman, P. (1992). Weak convergence of the bootstrap multidimensional empirical process for stationary strong-mixing sequences. Research Report, Seminar für Statistik, Eidgenössische Technische Hochschule, Zürich.

Carlstein, E. (1988). Degenerate *U*-statistics based on non-independent observations. *Cal. Statist. Assoc. Bull.* **37,** 55–65.

Cartwright, P. A. & Newbold, P. (1983). A time series approach to the prediction of oil discoveries. In: *Time series analysis: theory and practice*, Vol. 4. Elsevier, Amsterdam.

Chan, N. H. & Tran, L. T. (1992). Nonparametric tests for serial dependence. *J. Time Ser. Anal.* **13,** 19–28.

Chan, K. S. (1990). Testing for threshold autoregression. *Ann. Statist.* **18,** 1886–1894.

Chan, K. S. (1991). Percentage points of likelihood ratio tests for threshold autoregression. *J. Roy. Statist Soc. Ser. B.* **53,** 691–696.

Chan, K. S. & Tong, H. (1986). On estimating thresholds in autoregressive models. *J. Time Ser. Anal.* **7,** 179–190.

Chan, K. S. & Tong, H. (1990). On likelihood ratio tests for threshold autoregression. *J. Roy. Statist. Soc. B,* **52,** 469–476.

Chan, K. S., Moeanaddin, R. & Tong, H. (1991). Some difficulties of non-linear time series modelling. *Proceedings of the Taipei symposium of statistics, June 28–30, 1990* (eds M. T. Chao & P. E. Chen). Inst. of Statistical Science, Academia Sinica, Taipei, Taiwan.

Chen, R. & Tsay, R. (1993a). Functional coefficient autoregressive models. *J. Amer. Statist. Assoc.* **88,** 298–308.

Chen, R. & Tsay, R. (1993b). Nonlinear additive ARX models. *J. Amer. Statist. Assoc.* **88,** 955–967.

Cheng, B. & Tong, H. (1992). On consistent non-parametric order determination and chaos (with discussion). *J. Roy Statist. Soc. Ser. B* **54,** 427–474.

Dahlhaus, R. (1989). Efficient parameter estimation for self similar processes. *Ann. Statist.* **17,** 1749–1766.

Denker, M. & Keller, G. (1983). *U*-statistics and von Mises' statistics for weakly dependent processes. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **64,** 505–522.

Diebolt, J. & Guegan, D. (1991). Probabilistic properties of the general non-linear Markovian process of order one and applications to time series modelling. Technical Report, Laboratoire de Statistique, Université Paris XIII.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of variance of U.K. inflation. *Econometrica* **50,** 987–1008.

Engle, R. F. & Granger, C. W. J. (1987). Cointegration and error correction: representation, estimation and testing. *Econometrica* **55,** 251–276.

Engle, R. F. & Ng, V. K. (1991). Measuring and testing the impact of news on volatility. Discussion Paper, Department of Economics, University of California, San Diego.

Engle, R. F. & Granger, C. W. J., Rice, J. & Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81,** 310–320.

Findley, D. (1986). On bootstrap estimates of forecast mean square errors for autoregressive processes. In *Computer science and statistics: the interface* (ed. D. M. Allen). North-Holland, Amsterdam, pp. 11–17.

Franke, J. & Wendel, M. (1990). A bootstrap approach for nonlinear autoregressions. Some preliminary results. Preprint, to appear in *Proceedings of the international conference on bootstrapping and related techniques, Trier, June 1990.*

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19,** 1–141.

Friedman, J. H. & Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76,** 817–823.

Gallant, A. R. (1987). *Nonlinear statistical models.* Wiley, New York.

Gallant, A. R. & Tauchen, G. (1990). A nonparametric approach to nonlinear time series analysis. estimation and simulation. Preprint, to appear in *IMA volumes on mathematics and its applications.* Springer Verlag, Berlin.

Gallant, A. R., Rossi, P. E. & Tauchen, G. (1990). Stock prices and volume. Graduate School of Buisness, University of Chicago, Working Paper.

Grahn, T. (1992). Ph.D. thesis on bilinear processes. Institut für Angewandte Mathematik, University of Heidelberg.

Granger, C. W. J. & Hallman, J. J. (1991). Long-memory processes with attractors. *Oxford Bull. Econom. Statist.* **53,** 11–26.

Granger, C. W. J. & Teräsvirta, T. (1992a). Experiments in modeling nonlinear relationships between time series. In *Nonlinear modeling and forecasting* (eds M. Casdagli, S. Eubank & S. Eubank). Addison Wesley, Redwood City, CA.

Granger, C. W. J. & Teräsvirta, T. (1992b). *Modelling nonlinear dynamic relationships.* Oxford University Press, Oxford.

Gu, C. & Wahba, G. (1992). Smoothing spline ANOVA with component-wise Bayesian "confidence intervals". Technical Report, Department of Statistics, University of Wisconsin.

Guegan, D. & Pham, D. T. (1987). Minimalité et invertibilité des modèles bilinèaires á temps discret. *C. R. Acad. Sci. Paris* **304,** 159–162.

Guegan, D. & Pham, D. T. (1989). A note on the estimation of the parameters of the diagonal bilinear model by the method of least squares. *Scand. J. Statist.* **16,** 129–136.

Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Multivar. Anal.* **14,** 1–16.

Hallin, M. & Puri, M. L. (1992). Rank tests for time series analysis. A survey. Discussion paper, Centre d'Economie Mathématique et d'Econometrie, Université Libre de Bruxelles.

Handbook of econometrics, Vol. 4 (1993) eds. R. F. Engle & D. McFadden. North-Holland, Amsterdam.

Hannan, E. J. & Deistler, M. (1988). *The statistical theory of linear systems.* Wiley, New York.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models.* Chapman & Hall, London.

Heckman, N. (1986). Spline smoothing in a partly linear model. *J. Roy. Statist. Soc. Ser. B* **48,** 244–248.

Hinich, M. (1982). Testing for Gaussianity and linearity of a stationary time series. *J. Time Ser. Anal.* **3,** 169–176.

Hjellvik, V. & Tjøstheim, D. (1994). Non-parametric tests of linearity for time series. *Biometrika* (to appear).

Hoeffding, W. (1948). A non-parametric test of independence. *Ann. Math. Statist.* **19,** 546–557.

Holst, U., Lindgren, G., Holst, J. & Thuresholmen, M. (1994). Recursive estimation in switching autoregressions with Markov regime. *J. Time Ser. Anal.* (to appear).

Joe, H. (1989). Estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Statist. Math.* **41,** 683–697.

Karlsen, H. A. (1990a). Existence of moments in a stationary stochastic difference equation. *Adv. Appl. Prob.* **22,** 129–146.

Karlsen, H. A. (1990b). A class of nonlinear time series models. Doctoral Thesis, Department of Mathematics, University of Bergen.

Karlsen, H. A. & Tjøstheim, D. (1988). Consistent estimates for the NEAR(2) and the NLAR(2) time series models. *J. Roy. Statist. Soc. Ser B* **50,** 313–320.

Karlsen, H. A. & Tjøstheim, D. (1990). Autoregressive segmentation of signal traces with applications to dipmeter oil well measurements. *IEEE Trans. Geosci. Remote Sensing* **28,** 171–181.

Keenan, D. M. (1985). A Tukey nonadditivity-type test for time series nonlinearity. *Biometrika* **72,** 39–44.

Klimko, L. A. & Nelson, P. I. (1978). On conditional least squares estimation for stochastic processes. *Ann. Statist.*, **6,** 629–642.

Kreiss, J.-P. & Franke, J. (1992). Bootstrapping stationary autoregressive moving-average models. *J. Time Ser. Anal.* **13,** 297–318.

Künsch. H. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17,** 1217–1241.

Lawrance, A. J. (1991). Directionality and reversibility in time series. *Int. Statist. Rev.* **59,** 67–79.

Lawrance, A. J. & Lewis, P. A. W. (1985). Modelling and residual analysis of nonlinear autoregressive time series in exponential variables. *J. Roy. Statist. Soc. Ser B* **47,** 165–202.

Lewis, P. A. W. & Stevens, J. G. (1991a). Nonlinear modeling of time series using multivariate adaptive regression splines (MARS). *J. Amer. Statist. Assoc.* **86,** 864–877.

Lewis, P. A. W. & Stevens, J. G. (1991b). Semi-multivariate nonlinear modeling of time series using multivariate adaptive regression splines (MARS). Preprint, Naval Postgraduate School.

Luukkonen, R. P., Saikkonen, P. & Teräsvirta, T. (1988a). Testing linearity in univariate time series. *Scand. J. Statist.* **15,** 161–175.

Luukkonen, R. P., Saikkonen, P. & Teräsvirta, T. (1988b). Testing linearity against smooth transition autoregression. *Biometrika* **75,** 491–499.

Masry, E. & Tjøstheim, D. (1992). Non-parametric estimation and identification of ARCH and ARX nonlinear time series: convergence properties and rates. Preprint, Department of Mathematics, University of Bergen.

McLeod, A. I. & Li, W. K. (1983). Diagnostic checking ARMA time series models using squared-residuals autocorrelations. *J. Time Ser. Anal.* **4,** 269–273.

Nicholls, D. F. & Quinn, B. G. (1982). *Random coefficient autoregressive models: an introduction.* Lecture Notes in Statistics, **11,** Springer, New York.

Nummelin, E. (1984). *General irreducible Markov chains and non-negative operators.* Cambridge University Press, Cambridge.

Pham, D. T. (1985). Bilinear Markovian representations and bilinear models. *Stochastic Process. Appl.* **20,** 295–306.

Pötscher, B. M. & Prucha, I. R. (1991a). Basic structure of the asymptotic theory in dynamic nonlinear econometric models, Part I: consistency and approximation concepts. *Econometric Rev.* **10,** 125–216.

Pötscher, B. M. & Prucha, I. R. (1991b). Basic structure of the asymptotic theory in dynamic nonlinear econometric models, Part II: asymptotic normality (with discussion). *Econometric Rev.* **10,** 253–357.

Powell, J. L., Stock, J. H. & Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57,** 1403–1430.

Priestley, M. (1988). *Non-linear and non-stationary time series analysis.* Academic Press, London and San Diego.

Quinn, B. G. (1982). A note on the existence of strictly stationary solutions to bilinear equations. *J. Time Ser. Anal.* **3,** 249–252.

Rabiner. L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77,** 257–285.

Robinson, P. M. (1983). Non-parametric estimation for time series models. *J. Time Ser. Anal.* **4,** 185–208.

Robinson, P. M. (1988). Root-N-consistent semi-parametric regression. *Econometrica* **56,** 931–954.

Robinson, P. M. (1991). Consistent nonparametric entropy-based testing. *Rev. Econom. Stud.* **58,** 437–453.

Robinson, P. M. (1992). Time series with strong dependence. Invited paper at the 16th World Congress of the Econometric Society, Barcelona, 1990.

Rosenblatt, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Ann. Statist.* **3,** 1–14.

Rosenblatt, M. & Wahlen, B. (1993) A nonparametric measure of dependence under a hypothesis of independent components. *Statist. Probab. Lett.* (to appear).

Saikkonen, P. & Luukkonen, R. (1988). Lagrange multiplier tests for testing non-linearities in time series models. *Scand. J. Stat.* **15,** 55–68.

Shumway, R. H., Azari, A. S. & Pawitan, Y. (1988). Modeling mortality fluctuations in Los Angeles as functions of pollution and weather effects. *Environ. Res.* **45,** 224–241.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall, London.
Skaug, H. J. & Tjøstheim, D. (1993a) Non-parametric tests of serial independence. *The M. Priestley Birthday Volume* (ed. T. Subba Rao), pp. 207–229.
Skaug, H. J. & D. Tjøstheim, D. (1993b). A non-parametric test of serial independence based on the empirical distribution function. *Biometrika* **80,** 591–602.
Skaug, H. J. & Tjøstheim, D. (1993c). Measures of distance between densities with application to testing for serial independence. Preprint, Department of Mathematics, University of Bergen.
Stinchcombe, M. & White, H. (1989). Universal approximations using feedforward networks with non-sigmoid hidden layer activation functions. *Proceedings of the international joint conference on neural networks, Washington DC.*, SOS Printing, I. San Diego, pp. 613–618.
Subba Rao, T. & Gabr, M. M. (1980). A test for linearity of stationary time series. *J. Time Ser. Anal.* **1,** 145–158.
Subba Rao, T. & Gabr, M. M. (1984). *An introduction to bispectral analysis and bilinear time series models*. Lecture Notes in Statistics, **24,** Springer, New York.
Sugihara, G. & May, R. M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurements errors in time series. *Nature* **344,** 734–741.
Teräsvirta, T. (1990). Specification, estimation and evaluation of smooth transition autoregressive models. University of California, San Diego, Department of Economics, Discussion Paper No. 90-15.
Teräsvirta, T., Tjøstheim, D. & Granger, C. W. J. (1993a). Aspects of modelling nonlinear time series. Preprint, Department of Statistics, University of Gothenburg, to appear in *Handbook of econometrics*, Vol. 4 (eds R. F. Engle & D. McFadden). North-Holland, Amsterdam.
Teräsvirta, T., Lin, C.-E. & Granger, C. W. J. (1993b). Power of the neural network linearity test. *J. Time Ser. Anal.* **14,** 209–220.
Tjøstheim, D. (1986a). Some doubly stochastic time series models. *J. Time Ser. Anal.* **7,** 51–72.
Tjøstheim, D. (1986b). Estimation in nonlinear time series models. *Stochastic Process. Appl.* **21,** 251–273.
Tjøstheim, D. (1990). Non-linear time series and Markov chains. *Adv. Appl. Prob* **22,** 587–611.
Tjøstheim, D. & Auestad, B. (1994a). Non-parametric identification of non-linear time series: projections. *J. Amer. Statist. Assoc.* (to appear).
Tjøstheim, D. & Auestad, B (1994b). Non-parametric identification of non-linear time series: selecting significant lags. *J. Amer. Statist. Assoc.* (to appear).
Tong, H. (1983). *Threshold models in non-linear time series analysis*. Lecture Notes in Statistics, **21,** Springer, New York.
Tong, H. (1990). *Non-linear time series. A dynamical system approach*. Oxford University Press, Oxford.
Tong, H., Thanoon, B. & Gudmundson, G. (1985). Threshold time series modelling of the Icelandic riverflow systems. In *Time series analysis in water resources* (ed. K. W. Hipel). American Water Resource Association, Washington, DC.
Truong, Y. K. & Stone, C. J. (1992). Semi-parametric time series regression. Preprint, Department of Statistics, University of North Carolina.
Tsay, R. (1986). Nonlinearity tests for time series. *Biometrika* **73,** 461–466.
Tsay, R. (1989). Testing and modeling threshold autoregressive processes. *J. Amer. Statist. Assoc.* **84,** 231–240.
Tyssedal, J. S. & Tjøstheim, D. (1988). An autoregressive model with suddenly changing parameters and an application to stock market prices. *J. Roy. Statist. Soc. Ser. C (Appl. Statist.)* **37,** 353–369.
Tweedie, R. L. (1975). Sufficient conditions for ergodicity and recurrence of Markov chains on a general state space. *Stochastic Process. Appl.* **3,** 385–403.
Tweedie, R. L. (1983). Criteria for rates of convergence of Markov chains with applications to queueing theory. In *Papers in probability, statistics and analysis* (eds J. F. C. Kingman & G. E. H. Reuter). Cambridge University Press, Cambridge.
White, H. (1992). Parametric statistical estimation with artificial neural networks. Preprint, Department of Economics, University of California, San Diego.
Yao, Q. & Tong, H. (1992). On subset selection in non-parametric stochastic regression. Preprint, Institute of Statistics, University of Kent.
Yoshihara, K. (1976). Limiting behavior of U-statistics for stationary, absolutely regular processes. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **35,** 237–252.

Received October 1992, in final form May 1993

Dag Tjøstheim, Department of Mathematics, University of Bergen, 5007 Bergen, Norway.