



A Review of Nonparametric Time Series Analysis

Author(s): Wolfgang Hardle, Helmut Lutkepohl, Rong Chen

Source: *International Statistical Review / Revue Internationale de Statistique*, Vol. 65, No. 1 (Apr., 1997), pp. 49-72

Published by: International Statistical Institute (ISI)

Stable URL: <http://www.jstor.org/stable/1403432>

Accessed: 01/10/2008 14:20

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/action/showPublisher?publisherCode=isi>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We work with the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact support@jstor.org.



International Statistical Institute (ISI) is collaborating with JSTOR to digitize, preserve and extend access to *International Statistical Review / Revue Internationale de Statistique*.

<http://www.jstor.org>

A Review of Nonparametric Time Series Analysis

Wolfgang Härdle¹ Helmut Lütkepohl² and Rong Chen³

¹ ²*Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin, 10178 Berlin, Germany*

³*Department of Statistics, Texas A&M University, College Station, TX 77843, USA*

Summary

Various features of a given time series may be analyzed by nonparametric techniques. Generally the characteristic of interest is allowed to have a general form which is approximated increasingly precisely when the sample size goes to infinity. We review nonparametric methods of this type for estimating the spectral density, the conditional mean, higher order conditional moments or conditional densities. Moreover, density estimation with correlated data, bootstrap methods for time series and nonparametric trend analysis are described.

Key words: Kernel estimators; Smoothing techniques; Dependent observations; Bootstrap; Hermite expansions.

1 Introduction

The use of nonparametric techniques has a long tradition in time series analysis. As early as the late 19th century Schuster (1898) introduced the periodogram which may be regarded as the origin of spectral analysis. By now the latter technique is a classical nonparametric tool for analyzing time series. The increased data availability especially in finance and the explosion of computing power have made it possible to use a wide range of other modern nonparametric techniques in time series analysis recently. In this article we review some of these developments.

For a given time series X_1, \dots, X_n , nonparametric techniques are used to analyze various features of interest. Generally, the idea underlying many of these techniques is that the characteristic of interest is allowed to have a general form which is approximated increasingly precisely with growing sample size. For example, if a process is assumed to be composed of periodic components, a general form of spectral density may be assumed which can be approximated with increasing precision when the sample size gets larger. Similarly, if the autocorrelation structure of a stationary process is of interest the spectral density may be estimated as a summary of the second moment properties. A brief review of this classical method of nonparametric time series analysis is given in Section 2.

Because the final objective of many time series analyses is prediction, it is often of interest to study the conditional means, conditional variances or complete conditional densities in some period, given the past of the process. When a point prediction is the final objective, an estimate of some conditional mean may be desired, while the conditional variances are needed if interval forecasts or assessments of future volatility are desired. Moreover, if higher order moments of a series are potentially important, the focus may be on estimating the complete conditional density.

In order to analyze the conditional mean nonparametrically one may, for instance, start from a

model of the form

$$X_t = f(X_{t-1}, X_{t-2}, \dots) + \varepsilon_t \quad (1.1)$$

where ε_t is a series of innovations which is independent of past X_t . In this case $f(\cdot)$ represents the conditional expectation in period t , given past observations X_{t-1}, X_{t-2}, \dots and it is the minimum mean squared error (MSE) 1-step predictor for X_t . In parametric time series analysis the function $f(\cdot)$ is chosen from some parametric class so that the specific candidate is obtained by specifying a fixed finite number of parameters. Nonparametric approaches on the other hand allow $f(\cdot)$ to be from some flexible class of functions and they approximate $f(\cdot)$ in such a way that the approximation precision increases with the sample size. For this purpose several different techniques and procedures are available. For instance, local approaches approximate $f(\cdot)$ in the neighborhood of any given argument by letting the neighborhood decrease and thereby increase the approximation precision with growing sample size. For this purpose the number of lagged X_t used in the model is usually limited. In other words, $f(X_{t-1}, X_{t-2}, \dots)$ is replaced by $f(X_{t-1}, \dots, X_{t-p})$ for some fixed p . Alternatively, global approximators use parametric functions $f_n(\cdot)$, where the number of parameters and thereby the flexibility of the function may increase with the sample size n . The functions $f_n(\cdot)$ are chosen such that they approach $f(\cdot)$ in a certain norm when the sample size increases. This way it is also possible to let the number of lagged X_t 's increase with the sample size n and thus avoid assuming a fixed number of lags at an early stage of the analysis. A number of methods for estimating the conditional mean function of a process are discussed in Section 3.

As mentioned earlier, in many situations point forecasting is too limited an objective and the future volatility and other higher order moments are of interest in addition to the conditional mean. Therefore the framework in (1.1) is often extended to a more general model

$$X_t = f(X_{t-1}, X_{t-2}, \dots) + g(X_{t-1}, X_{t-2}, \dots)\varepsilon_t \quad (1.2)$$

where $g(\cdot)$ is used to represent the conditional variance of the process in period t given the information from previous periods. Again various nonparametric approaches exist for joint estimation of $f(\cdot)$ and $g(\cdot)$. Of course, it is also possible to specify a parametric form of one of the two functions and treat the other one nonparametrically. Techniques for nonparametric analyses of model (1.2) are the subject of Section 4. More generally the complete predictive (conditional) density $h(X_t|X_{t-1}, X_{t-2}, \dots)$ may be of interest when the shape of the conditional distribution and higher order moments are relevant to the analysis. For this case a number of different nonparametric approaches have been proposed as well. Some of them are also sketched in Section 4.

There are numerous other nonparametric procedures and techniques that have been used in time series analysis. For instance, when a parametric time series model such as (1.2) with parametric functions $f(\cdot)$ and $g(\cdot)$ is specified it may be of interest to estimate the distribution of the residuals by nonparametric methods in order to improve the parameter estimators or to assess the statistical properties of the estimators. More precisely, density estimation for the residuals and bootstrap methods based on the residuals have been used in this context. These methods are reviewed in Section 5. Another important characteristic of a time series is its trending behaviour. Deterministic trend functions have also been analyzed nonparametrically. In addition, there are a number of nonparametric tests for stochastic trends. They are also presented in Section 5.

If very general assumptions are made, a rich data set is usually necessary to obtain a good idea about the features of interest. Therefore, many of the nonparametric techniques reviewed in this article are typically used when long time series are available. Therefore, these methods have, for instance, been used for analyzing financial time series which are observed with a high frequency and are consequently relatively long. Other fields of applications include survey of riverflow, the analysis of encephalographic data and of sleep states. Although we provide a fairly broad survey of many nonparametric analysis techniques for time series we are aware that such a survey is necessarily

limited neglecting many interesting and potentially promising facets of research in this area. In particular, we are unable to give a complete listing of related publications because of the recent explosion in the literature due to the increase in data availability and computing power. We apologize for any omissions of relevant related work. Further references may be found in Györfi, Härdle, Sarda & Vieu (1989), Tjøstheim (1994) and Hart (1996).

2 Spectral Analysis

Suppose $\{X_t\}$ is a zero mean univariate stationary stochastic process with autocovariances $\gamma_k = E(X_t X_{t+k})$. Then the spectral density of $\{X_t\}$ is

$$f_X(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_k e^{-i\omega k} = \frac{1}{2\pi} (\gamma_0 + 2 \sum_{k=1}^{\infty} \gamma_k \cos \omega k), \quad \omega \in [-\pi, \pi].$$

Here $i = +\sqrt{-1}$ as usual. Hence, the spectral density may be regarded as a weighted sum of cyclical components corresponding to frequencies ω in the interval $[-\pi, \pi]$. Since

$$\gamma_k = \int_{-\pi}^{\pi} e^{i\omega k} f_X(\omega) d\omega,$$

the second order characteristics of the process can be recovered if the spectral density is available. In particular, $\gamma_0 = \text{Var}(X_t) = \int_{-\pi}^{\pi} f_X(\omega) d\omega$ and thus the spectral density represents the contributions of the frequencies to the variance of the process. Hence, the spectral density may be regarded as a summary of the cyclical components of the process or alternatively as a representation of the second order moments or autocovariance structure of the process.

Given a time series X_1, \dots, X_n the autocovariances of the generating process may be estimated as

$$\tilde{\gamma}_k = \frac{1}{n} \sum_{t=1}^{n-k} (X_t - \bar{X})(X_{t+k} - \bar{X}),$$

or by

$$\hat{\gamma}_k = \frac{1}{n-k} \sum_{t=1}^{n-k} (X_t - \bar{X})(X_{t+k} - \bar{X}),$$

$k = 1, \dots, n-1$, where $\bar{X} = \sum_{t=1}^n X_t/n$ is the sample mean. An obvious estimator of the spectral density at frequency ω is the so called *periodogram*

$$\tilde{f}_X(\omega) = \frac{1}{2\pi} \sum_{k=-(n-1)}^{n-1} \tilde{\gamma}_k e^{-i\omega k},$$

or similarly with $\hat{\gamma}_k$ replacing $\tilde{\gamma}_k$. Unfortunately, this estimator is not consistent. The reason is that too many quantities are estimated from the sample.

To ensure consistency a smoothed estimator of the form

$$\hat{f}_X(\omega) = \frac{1}{2\pi} \sum_{k=-M}^M \lambda_k \tilde{\gamma}_k e^{-i\omega k}$$

is usually used. The weights $\lambda_{-M}, \dots, \lambda_M$ represent the *spectral window* and $M (< n-1)$ is the truncation point which depends on the sample size. A number of different windows has been

proposed in the literature. The following are examples:

$$\begin{aligned} \lambda_k &= 1 - |k|/M \quad (\text{Bartlett, 1950}) \\ \lambda_k &= 1 - 2a + 2a \cos\left(\frac{\pi k}{M}\right) \quad (\text{Tukey, 1949, Blackman \& Tukey, 1959}) \\ \lambda_k &= \frac{1}{2} \left[1 + \cos\left(\frac{\pi k}{M}\right) \right] \quad (\text{Tukey, 1949}) \\ \lambda_k &= \begin{cases} 1 - 6\left(\frac{k}{M}\right)^2 + 6\left(\frac{|k|}{M}\right)^3 & \text{for } |k| \leq \frac{M}{2} \\ 2\left(1 - \frac{|k|}{M}\right)^3 & \text{for } \frac{M}{2} \leq |k| \leq M \end{cases} \quad (\text{Parzen, 1961}). \end{aligned}$$

A number of other windows are discussed in Priestley (1981, Sec. 6.2.3.). It may be worth noting that, for frequencies $\omega_j = 2\pi j/n$, the resulting spectral density estimators may be obtained alternatively by averaging over the periodogram values of neighboring frequencies. Hence,

$$\hat{f}_X(\omega_j) = \frac{1}{2\pi} \sum_{m=-h}^h K(\omega_j, \omega_{j+m}) \tilde{f}_X(\omega_{j+m}),$$

where $K(\cdot, \cdot)$ is a suitable kernel function and h is the bandwidth of frequencies used in the weighted average. In other words, $\hat{f}_X(\omega_j)$ may be obtained by kernel smoothing techniques which are discussed in more detail in the context of estimating the conditional mean (see Section 3.1). These ideas extend directly to the multivariate case where X_t is a vector of variables.

As mentioned in the introduction, spectral analysis of stationary processes is now a standard technique. It can be found in many time series textbooks and monographs. More recent developments in spectral analysis include nonstationary and nonlinear processes. For instance, Priestley (1981, Chapter 11) and Dahlhaus (1993) consider processes with time varying spectra. Priestley (1996) discusses the use of wavelets in this context. Nowadays spectral methods are used in various ways for analyzing time series both theoretically and empirically. Applications of these techniques include studies of seasonal behaviour of time series, approximation of the stationary part of more general processes, construction of testing and estimation procedures and examination of their properties (see, e.g., the chapters in Brillinger & Krishnaiah (1983) and in particular Robinson (1983a)). The related literature is too voluminous to be reviewed here. Hence, we regard our foregoing remarks on spectral analysis as a brief reminder that these techniques belong under the heading of this survey.

3 Estimation of the Conditional Mean

In this section we review some nonparametric methods for estimating the function $f(\cdot)$ in (1.1). We first present some smoothing approaches for locally approximating this function in the sense discussed in the introduction. For that purpose it is assumed that only a finite number of lagged X_t 's enters $f(\cdot)$, that is, $f(X_{t-1}, X_{t-2}, \dots) = f(X_{t-1}, \dots, X_{t-p})$. Some of the methods discussed in this section impose further restrictions on $f(\cdot)$ by assuming e.g. additivity of the lags (see Section 3.2). We also consider the problem of choosing the lag length p . Moreover, in Section 3.3 global approximations are reviewed which, in principle, allow an infinite number of lags of X_t in $f(\cdot)$.

The parametric approach to estimation of the conditional mean of a time series is to formulate a parametric model for $f(\cdot)$. Many parametric structures proposed for $f(\cdot)$ have been successful in practice and have provided parsimonious models that capture the linearity or nonlinearity of the underlying process. The most common nonlinear structures are the threshold autoregressive (TAR) models of Tong (1983), the exponential autoregressive (EXPAR) models of Haggan & Ozaki (1981), the smooth-transition autoregressive (STAR) models of Chan & Tong (1986) and Granger & Teräsvirta (1993). In these models the structure for $f(\cdot)$ is supposed to be of threshold type where

the threshold functions are modeled in different ways. Many other related references can be found in Tong (1990) and Priestley (1988).

The nonparametric approach has the advantage of letting the data speak for themselves. Hence, it avoids the subjectivity of choosing a specific parametric model before looking at the data. However, there is the cost of more complicated mathematical arguments and difficulties in practical implementation, such as the selection of smoothing parameters. Also there is the cost of poor performance in high dimensions, often referred to as the ‘curse of dimensionality’. Hence, the nonparametric approach often serves as a guidance for choosing appropriate lower dimensional parametric models and for deciding between competing classes of models. Powerful computers and easy-to-use interactive statistical and graphical softwares such as S (Becker, Chamber & Wilks, 1988) and XploRe (Härdle, Klinkle & Turlach, 1995) provide solid platforms for these operations.

3.1 Unrestricted Local Smoothing Methods

Model (1.1) has the format of a nonlinear regression problem for which many smoothing methods exist when the observations are independent. Hart (1996) demonstrates that these methods can be ‘borrowed’ for time series analysis where observations are correlated by making use of the ‘whitening by windowing principle’. This principle is introduced first. Then we list some common nonparametric smoothing methods for inference on the function $f(\cdot)$ in model (1.1).

The Whitening by Windowing Principle

Given an independent random sample X_1, \dots, X_n , which is drawn from a distribution with density function $p(x)$, a popular method of estimating $p(x)$ is based on the kernel estimator

$$\hat{p}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \tag{3.1}$$

where $h > 0$ is the so-called bandwidth and $K(\cdot)$ is a kernel function, typically with finite support. The bandwidth is taken as a sequence $h = h_n$ tending to zero as $n \rightarrow \infty$. Note that, if the kernel function has support on $[-1, 1]$, the estimator only uses the observations in the interval $[x - h, x + h]$. This is an important feature when we extend this method to dependent observations. When the estimator is applied to dependent observations, it is affected only by the dependency of the observations in a small window, not that of the whole data set. Hence, if the dependency between the observations is of ‘short memory’ which makes the observations in small windows *almost independent*, then most of the techniques developed for independent observations apply in this situation. Hart (1996) calls this feature *the whitening by windowing principle*.

Various *mixing conditions* are the main tools for proving asymptotic properties of the smoothing techniques for dependent data. Basically these conditions try to control the dependence between X_i and X_j as the time distance $i - j$ increases. For example, a sequence is said to be α -mixing (strong mixing) (Robinson 1983b) if

$$\sup_{A \in \mathcal{F}_1^n, B \in \mathcal{F}_{n+k}^\infty} |P(A \cap B) - P(A)P(B)| \leq \alpha_k$$

where $\alpha_k \rightarrow 0$ and \mathcal{F}_i^j is the σ -field generated by X_i, \dots, X_j . A stronger condition is the ϕ -mixing (uniformly mixing) condition (Billingsley 1968) where

$$|P(A \cap B) - P(A)P(B)| \leq \phi_k P(A)$$

for any $A \in \mathcal{F}_1^n$, and $B \in \mathcal{F}_{n+k}^\infty$ and ϕ_k tends to zero for $k \rightarrow \infty$. The rate at which α_k and ϕ_k go to zero plays an important role in showing asymptotic properties of the nonparametric smoothing

procedures. We note that generally these conditions are difficult to check. However, if the process follows a stationary Markov chain, then geometric ergodicity implies absolute regularity, which in turn implies strong mixing conditions. Techniques exist for checking the geometric ergodicity, see Tweedie (1975), Tjøstheim (1990), Pham (1985), Diebolt & Guegan (1990).

Local Conditional Mean and Median

Consider the general nonlinear autoregressive process of order p

$$X_t = f(X_{t-1}, \dots, X_{t-p}) + \varepsilon_t. \quad (3.2)$$

Let $Y_t = (X_{t-1}, \dots, X_{t-p})$, and choose $\delta_n > 0$ as a function of the sample size n . For any $\mathbf{y} = (x_1, \dots, x_p) \in \mathbb{R}^p$, let $I_n(\mathbf{y}) = \{i : 1 < i < n \text{ and } \|Y_i - \mathbf{y}\| < \delta_n\}$ and $N_n(\mathbf{y}) = \#I_n(\mathbf{y})$. Here $\|\cdot\|$ denotes the Euclidean norm. The local conditional mean function estimator is given by $\hat{f}(x_1, \dots, x_p) = \hat{f}_n(\mathbf{y}) = \{N_n(\mathbf{y})\}^{-1} \sum_{i \in I_n(\mathbf{y})} X_i$, that is, an average of all observations X_i corresponding to Y_i in a small neighborhood of the argument \mathbf{y} is used as the estimator. Alternatively, the local conditional median estimator given by $\tilde{f}(x_1, \dots, x_p) = \text{median}\{X_i, i \in I_n(\mathbf{y})\}$ may be used. Under strong mixing conditions, Truong (1993) proved strong consistency and asymptotic normality of these estimators, along with the optimal rate of convergence for suitable sequences $\delta_n \rightarrow 0$.

Nonparametric Kernel Estimation

Robinson (1983b), Auestad & Tjøstheim (1990), Härdle & Vieu (1992), and others used a kernel estimator (or robustified versions of it) to estimate the conditional mean function $f(X_{t-1}, \dots, X_{t-p})$. For this purpose the Nadaraya–Watson estimator with product kernels

$$\hat{f}(x_1, \dots, x_p) = \frac{\sum_{t=p+1}^n \prod_{i=1}^p K\{(x_i - X_{t-i})/h_i\} X_t}{\sum_{t=p+1}^n \prod_{i=1}^p K\{(x_i - X_{t-i})/h_i\}}, \quad (3.3)$$

is used where $K(\cdot)$ is again a kernel function with bounded support and the h_i 's are the bandwidths. In other words, a weighted average of the observations is used as an estimator of $f(\cdot)$.

Robinson (1983b) and Masry & Tjøstheim (1995a) show strong consistency and asymptotic normality for α -mixing observations. Bierens (1983, 1987) and Collomb & Härdle (1986) proved the uniform consistency of the estimator under the assumption of a ϕ -mixing process. Singh & Ullah (1985) extend this approach to multiple time series, where X_t is a vector rather than a scalar random variable.

Local Polynomial Regression

Local polynomial regression techniques offer yet another alternative for estimating the conditional mean of time series nonparametrically. In this approach polynomials of a prespecified degree, say $l - 1$, are fitted locally in the neighborhood of a given argument of $f(\cdot)$, where the size of the neighborhood shrinks with increasing sample size n . To state this estimator formally, suppose for simplicity that $p = 1$, that is, the model is $X_t = f(X_{t-1}) + \varepsilon_t$. We wish to estimate $f(x)$. In this case the estimator is obtained by minimization of

$$c_n(x) = \arg \min_{c \in \mathbb{R}^l} \sum_{t=1}^n (X_t - c^T U_{tn})^2 K\{(X_{t-1} - x)/h\}.$$

where $K(\cdot)$ is a kernel function, h is a positive bandwidth sequence, and

$$U_{ln} = F(u_{ln}), \quad F(u) = (1, u, \dots, u^{l-1}/(l-1)!)^T, \quad u_{ln} = (X_{t-1} - x)/h.$$

The estimator $\hat{f}(x)$ is given by $\hat{f}(x) = c_n(x)^T F(0)$. This estimator was first developed by Stone (1977) and Katkovnik (1979). In the context of independent observations Fan (1993) studied minimax efficiency and made the technique popular to applied statisticians. Tsybakov (1986) and Härdle & Tsybakov (1997) proved asymptotic normality of these estimators under conditions satisfying the assumptions of Tweedie (1975) and Diebolt & Guegan (1990). A multivariate extension of this approach is given by Härdle, Tsybakov & Yang (1996).

Nonparametric Multi-step Prediction

All these methods estimate the conditional mean of a nonlinear AR process and thereby provide a one-step ahead predictor. Often forecasts for more than one step ahead are desired. Similar nonparametric techniques can be used for that purpose and we briefly mention some proposals here.

Consider the nonlinear AR(1) model $X_t = f(X_{t-1}) + \varepsilon_t$. Since the conditional mean $m_k(x) = E(X_{t+k} | X_t = x)$ is the least squares predictor for k -step ahead prediction, Auestad & Tjøstheim (1990), Härdle & Vieu (1992) and Härdle (1990) proposed using the ordinary Nadaraya–Watson estimator

$$\hat{m}_{h,k}(x) = \frac{\sum_{t=1}^{n-k} K\{(x - X_t)/h\} X_{t+k}}{\sum_{t=1}^{n-k} K\{(x - X_t)/h\}} \tag{3.4}$$

to estimate $E(X_{t+k} | X_t = x)$ directly.

Note, however, that the variables $X_{t+1}, \dots, X_{t+k-1}$ may contain information about the conditional mean function $E(X_{t+k} | X_t)$. Therefore Chen (1996) and Chen & Hafner (1995) proposed a multistage kernel smoother which utilizes this information. For illustrative purposes consider two-step ahead forecasting. Due to the Markov property, we have

$$m_2(x) = E[X_{t+2} | X_t = x] = E[E(X_{t+2} | X_{t+1}, X_t) | X_t = x] = E[E(X_{t+2} | X_{t+1}) | X_t = x].$$

Define $f(y) = E(X_{t+2} | X_{t+1} = y)$. Ideally, if we knew $f(\cdot)$, we would use the pairs $(f(X_{t+1}), X_t)$, $t = 1, \dots, (n-1)$ in estimating $E(X_{t+2} | X_t)$, whereas the direct estimator (3.4) uses the pairs (X_{t+2}, X_t) . Since X_{t+2} is a noisy representative of $f(X_{t+1})$ with $O_p(1)$ error, we can improve the estimation by using an estimator $\hat{f}(X_{t+1})$ with $\hat{f}(X_{t+1}) - f(X_{t+1}) = o_p(1)$. This motivates the ‘multistage smoother’

$$\hat{m}_{h_1, h_2}(x) = \frac{\sum_{t=1}^{n-1} K\{(x - X_t)/h_2\} \hat{f}_{h_1}(X_{t+1})}{\sum_{t=1}^{n-1} K\{(x - X_t)/h_2\}}$$

where

$$\hat{f}_{h_1}(y) = \frac{\sum_{j=1}^{n-1} K\{(y - X_j)/h_1\} X_{j+1}}{\sum_{j=1}^{n-1} K\{(y - X_j)/h_1\}}.$$

It can be shown that the new smoother has a smaller mean squared error than (3.4).

Implementation Issues

One of the important implementation issues of the nonparametric smoothing tools is the bandwidth selection in finite samples. There are many data-driven methods proposed for independent data, e.g. the cross-validation method of Rudemo (1982) and Bowman (1994) and the plug-in rules of Sheather (1983), Park & Marron (1990) and Park & Turlach (1992).

Again, for simplicity we assume a nonlinear AR(1) model $X_t = f(X_{t-1}) + \varepsilon_t$. For dependent data, one of the criteria for selecting the bandwidth is to minimize the averaged squared error

$$d_A(h) = \frac{1}{n} \sum_{t=1}^n \{f(X_t) - \hat{f}_h(X_t)\}^2 w(X_t),$$

which is an approximation of the integrated squared error

$$d_I(h) = \int \{f(x) - \hat{f}_h(x)\}^2 \eta(x) w(x) dx.$$

Here $\eta(\cdot)$ denotes the density of the stationary distribution and $w(\cdot)$ is a weight function with compact support. The measure of accuracy $d_A(h)$ involves the unknown autoregression function $f(\cdot)$, so it cannot be estimated by a plug-in type approach. For the nonparametric kernel estimator, Härdle & Vieu (1992) and Härdle (1990) proposed to use the leave-on-out cross-validation function

$$CV(h) = \frac{1}{n-1} \sum_{t=2}^n \{X_t - \hat{f}_{h,t}(X_{t-1})\}^2 w(X_{t-1}),$$

where

$$\hat{f}_{h,t}(x) = \frac{n^{-1} \sum_{j \neq t} K\{(x - X_{j-1})/h\} X_j}{n^{-1} \sum_{j \neq t} K\{(x - X_{j-1})/h\}}, \quad (3.5)$$

to select the bandwidth. Let \hat{h} be the bandwidth that minimizes $CV(h)$. They proved that, under an α -mixing condition,

$$\frac{d_A(\hat{h})}{\inf_h d_A(h)} \rightarrow 1 \quad \text{in probability.}$$

Similar results for density estimation were obtained by Hart & Vieu (1990).

A Nonparametric Nonlinearity Test

Hjellvik & Tjøstheim (1995) proposed a nonlinearity test which may help in deciding whether to use a nonlinear model rather than a linear one. It is based on the distance between the best linear predictor $\rho_k X_{t-k}$ and the best nonlinear predictor $m_k(X_{t-k}) = E[X_t | X_{t-k}]$ of X_t based on X_{t-k} . The distance is defined as

$$L(m_k) = E[\{m_k(X_{t-k}) - \rho_k X_{t-k}\}^2 w(X_{t-k})]$$

where $w(x)$ is a weighting function with compact support and ρ_k is the autocorrelation between X_t and X_{t-k} , assuming X_t has zero mean. The function $m_k(\cdot)$ is estimated using the Nadaraya–Watson estimator.

Lag Selection and Order Determination

The lag selection and order determination problem is important for effective implementation of nonlinear time series modeling. Often the set of lagged variables and possibly additional exogenous variables is too large for an efficient application of nonparametric smoothing techniques. In that case one may wish to select the most significant components. For linear time series models, lag selection and order determination are usually done using information criteria as proposed by Akaike (1970, 1974), along with other model checking procedures such as residual analysis. In a fully nonparametric approach to time series analysis, Auestad & Tjøstheim (1990) and Tjøstheim & Auestad (1994b) proposed the FPE (final prediction error) criterion and Cheng & Tong (1992) suggested using cross

validation. More specifically, Tjøstheim & Auestad (1994b) proposed to use an estimated FPE criterion to select lag variables and to determine the model order of the general nonlinear AR model in (3.2). Let X_t be a stationary strong mixing nonlinear AR process and let $\mathbf{i} = (i_1, \dots, i_p)$ and $Y_t(\mathbf{i}) = (X_{t-i_1}, \dots, X_{t-i_p})^T$. Define

$$\widehat{FPE}(\mathbf{i}) = \frac{1}{n} \sum_t [X_t - \hat{f}\{Y_t(\mathbf{i})\}]^2 w\{Y_t(\mathbf{i})\} \frac{1 + (nh^p)^{-1} J^p B_p}{1 - (nh^p)^{-1} \{2K^p(0) - J^p\} B_p} \quad (3.6)$$

where

$$J = \int K(x)^2 dx, \quad B_p = n^{-1} \sum_t \frac{w\{Y_t(\mathbf{i})\}^2}{\hat{p}\{Y_t(\mathbf{i})\}}$$

and $\hat{f}\{Y_t(\mathbf{i})\}$ is the kernel conditional mean estimator in (3.3) based on the lags specified in \mathbf{i} and $\hat{p}\{Y_t(\mathbf{i})\}$ is a multivariate kernel density estimator defined as in (3.1). Note that the \widehat{FPE} is essentially a sum of squares of residuals (RSS) multiplied by a term in (3.6) that penalizes small bandwidths h and a large order p .

Cheng & Tong (1992) used a leave-one-out cross validation procedure to select the order of a general nonlinear AR model. Let $Y_t(p) = (X_{t-1}, \dots, X_{t-p})$ and

$$CV(p) = \frac{1}{n-r+1} \sum_t [X_t - \hat{f}_{h,t}\{Y_t(p)\}]^2 w\{Y_t(p)\}$$

where $\hat{f}_{h,t}$ is the kernel conditional mean estimator defined in (3.5) and $w(\cdot)$ is a weight function of finite support. They proved that, under regularity conditions,

$$CV(p) = RSS(p)\{1 + 2K(0)\gamma h^{-p}/n + o_p(1/h^p n)\}$$

where $\gamma = \int w(x)dx / \int w(x)p(x)dx$ and h is the bandwidth. Again, this can be viewed as a penalized sum of squares of residuals.

3.2 Restricted Autoregressive Approaches

Since the nonparametric general approach suffers from the ‘curse of dimensionality’, unless the AR order p is very small, restrictions on the function $f(\cdot)$ have been proposed. Common structural restrictions are additivity, single index restrictions and/or data dependent coefficients in a ‘linear’ model. These restrictions result in better convergence rates and are easier to interpret, especially with graphics supported from interactive statistical computing environments. This is important since nonparametric models are not the end of an analysis. They are rather an exploratory tool for a better understanding of the underlying dynamics of the process and a starting point for finding more parsimonious models.

Nonlinear Additive AR Models

A nonlinear additive autoregressive (NAAR) model is defined as

$$X_t = c + f_1(X_{t-i_1}) + f_2(X_{t-i_2}) + \dots + f_p(X_{t-i_p}) + \varepsilon_t. \quad (3.7)$$

Additive models have been studied extensively in the regression context by Hastie & Tibshirani (1990). The NAAR model in (3.7) is a generalization of the first-order nonlinear AR model of Jones (1978). It is very flexible as it encompasses linear AR models and many interesting nonlinear models as special cases. These models naturally generalize the linear regression models and allow interpretation of marginal changes, i.e. the effect of one variable (or lagged variable) on the mean function. They are also interesting from a theoretical point of view since they combine flexible

nonparametric modeling of many variables with statistical precision that is typical for just one explanatory variable. Accurate estimation can be achieved with moderate sample sizes. Here we introduce three procedures for estimating the NAAR model. Order determination and lag selection problems are addressed as well.

Chen & Tsay (1993a) use *backfitting algorithms* such as the Alternating Conditional Expectation (ACE) algorithm and the BRUTO algorithm of Hastie & Tibshirani (1990) to fit the additive model (3.7). Note that the AVAS algorithm of Tibshirani (1988) can also be used here. The main idea of backfitting is that if the additive model is correct, then for any k we have $f_k(X_{t-i_k}) = E\{X_t - c - \sum_{j \neq k} f_j(X_{t-i_j}) \mid X_{t-i_k}\}$. Consequently, we can treat $X_t - c - \sum_{j \neq k} f_j(X_{t-i_j})$ as the conditional response variable and use nonparametric smoothers to estimate $f_k(\cdot)$. In practice, all $f_k(\cdot)$'s are unknown so that the estimates are iterated until they all converge. The effective hat matrix of this algorithm is computed in Härdle & Hall (1993), showing that the iteration results depend on the starting index.

One of the problems associated with the backfitting algorithms is that with highly correlated observations, the convergence can be slow, as noted in Chen & Tsay (1993a). Linton & Nielson (1995) and Chen *et al.* (1996) proposed an integration estimator for estimating the functions in additive regression models without using backfitting. At the same time, Tjøstheim & Auestad (1994a) and Masry & Tjøstheim (1995b) proposed the same estimator for NAAR models. Specifically, the 'integration idea' is based on the following observation. If the model is of the additive form (3.7), and $f(x_1, \dots, x_p) = c + \sum_{j=1}^p f_j(x_j)$ is the conditional mean function, and $p_{-j}(\cdot)$ is the joint density of $X_{t-i_1}, \dots, X_{t-i_{j-1}}, X_{t-i_{j+1}}, \dots, X_{t-i_p}$, then for a fixed $x \in \mathbb{R}$,

$$f_j(x) + c = \int f(x_1, \dots, x, \dots, x_p) p_{-j}(x_1, \dots, x_p) \prod_{l \neq j} dx_l,$$

provided $E f_l(X_t) = 0, l = 1, \dots, p$. Using the Nadaraya–Watson estimator to estimate the mean function $f(\cdot)$, we average over the observations to obtain the following estimator.

Let $K_h(\cdot) = h^{-1}K(\cdot/h)$, where $K(\cdot)$ is a kernel function. For $1 \leq j \leq p$ and any x in the domain of $f_j(\cdot)$, define, for $h_n > 0, h'_n > 0$,

$$\begin{aligned} \hat{f}_j(x) &= \frac{1}{n} \sum_{t=1}^n \hat{f}(X_{t-i_1}, \dots, X_{t-i_{j-1}}, x, X_{t-i_{j+1}}, \dots, X_{t-i_p}) \\ &= \frac{1}{n} \sum_{t=i_p+1}^n \left[\frac{\sum_{s=i_p+1}^n [\prod_{l \neq j} K_{h'_n}(X_{s-i_l} - X_{t-i_l})] K_{h_n}(X_{s-i_j} - x) X_s}{\sum_{s=i_p+1}^n [\prod_{l \neq j} K_{h'_n}(X_{s-i_l} - X_{t-i_l})] K_{h_n}(X_{s-i_j} - x)} \right]. \end{aligned} \tag{3.8}$$

The asymptotic normality of this estimator was established by Chen *et al.* (1996) for independent observations and by Masry & Tjøstheim (1995b) under strong mixing conditions for time series observations. The rate of convergence for estimating $f(\cdot)$ is $n^{2/5}$ which is typical for regression smoothing with just one explanatory variable. Hence, the estimator does not suffer from the 'curse of dimensionality'.

Wong & Kohn (1996) use spline nonparametric regression to estimate the components of a NAAR model. They adopt an equivalent Bayesian formulation of the spline smoothing and use a Gibbs sampler to estimate the components and the parameters of the model, through Monte Carlo simulation of the posterior distributions.

Chen, Liu & Tsay (1995) propose three nonparametric procedures for testing additivity in nonlinear time series analysis. For lag selection, Chen & Tsay (1993a) propose a procedure that is similar to the best subset procedure in linear regression analysis.

Functional Coefficient AR Model

A functional coefficient autoregressive (FAR) model can be written as

$$X_t = f_1(X_{t-d})X_{t-1} + f_2(X_{t-d})X_{t-2} + \cdots + f_p(X_{t-d})X_{t-p} + \varepsilon_t.$$

The model generalizes the linear AR models by allowing the coefficients to change according to a threshold lag variable X_{t-d} . The model can be extended to allow for multiple threshold variables in the coefficient functions. The model is general enough to include the threshold AR (TAR) models of Tong (1983) and Tsay (1989) (when the coefficient functions are step functions) and the exponential AR (EXPAR) models proposed by Haggan & Ozaki (1981) (when the coefficient functions are exponential functions) along with many other models (e.g., the STAR models of Granger & Teräsvirta (1993) and Teräsvirta (1994) and sine function models). Chen & Tsay (1993b) use an arranged local regression (ALR) procedure to roughly identify the nonlinear functional forms. For $x \in \mathbb{R}$ and $\delta_n > 0$, let $I_n(x) = \{t : 1 < t < n, |X_{t-d} - x| < \delta_n\}$. If we regress X_t on X_{t-1}, \dots, X_{t-p} using all the observations X_t for which $t \in I_n(x)$, then the estimated coefficients can be used as estimates of $f_i(x)$, $i = 1, \dots, p$. One can then make inference directly or formulate parametric models based on the estimated nonlinear functional forms. Chen & Tsay (1993b) proved the consistency of the estimator under geometric ergodicity conditions. Note that the locally weighted regression of Cleveland & Devlin (1988) may be used for estimating FAR models as well.

Adaptive Spline Threshold AR Model

Lewis & Stevens (1991) propose the adaptive spline threshold autoregressive (ASTAR) model of the form $X_t = \sum_{j=1}^s c_j K_j(X_{t-1}, \dots, X_{t-p}) + \varepsilon_t$, where $\{K_j(x)\}_{j=1}^s$ are product basis functions of truncated splines $T^-(x) = (t-x)_+$ and $T^+(x) = (x-t)_+$ associated with the subregions $\{R_j\}_{j=1}^s$ in the domain of the lag variables $(X_{t-1}, \dots, X_{t-p})$. For example, Lewis & Stevens (1991) use the following ASTAR model for the famous sunspot numbers:

$$\begin{aligned} \hat{X}_t = & 2.711 + 0.96X_{t-1} + 0.332(47 - X_{t-5})_+ - 0.257(59.1 - X_{t-9})_+ \\ & - 0.003X_{t-1}(X_{t-2} - 26.0)_+ + 0.017X_{t-1}(44.0 - X_{t-3})_+ \\ & - 0.032X_{t-1}(17.1 - X_{t-4})_+ + 0.004X_{t-1}(26 - X_{t-2})_+(X_{t-5} - 41.0)_+ \end{aligned}$$

where $(u)_+ = u$ if $u > 0$ and $(u)_+ = 0$ if $u \leq 0$. The modeling and estimation procedures follow the Multivariate Adaptive Regression Splines (MARS) algorithm of Friedman (1988). It is basically a regression tree procedure using truncated regression splines.

Index Models

Bierens (1994) discusses another way of imposing constraints on the general model (1.1). He shows that for a rational valued process the conditional expectation can be written as a function of an index, i.e. $E(X_t | X_{t-1}, X_{t-2}, \dots) = f(\xi_t)$, where the index ξ_t is related to the past observations X_{t-1}, X_{t-2}, \dots . For instance, the index may be of the form $\xi_t = \sum_{i=1}^{\infty} \eta^{i-1} X_{t-i}$ for some $\eta \in (-1, 1)$. Obviously, in this case $f(\cdot)$ is one dimensional and is therefore relatively easy to estimate by kernel methods. For practical purposes, assuming that X_t is rational is not restrictive because on a computer only a finite number of digits can be stored so that all observed time series are actually rational.

Bierens shows that there is a wide range of indices to choose from and suggests the following procedure for applied work. In a first step the best fitting linear ARMA model should be constructed. The optimal linear one-step-ahead predictor from that model is then used as an index ξ_t . If especially designed specification tests indicate remaining nonlinearity the function $f(\cdot)$ may be chosen either

from some parametric family or by using nonparametric smoothing techniques. Of course, a linear model is maintained if no nonlinearity is detected.

3.3 Global Approximators

As mentioned previously, a sequence of parametric functions can be used as global approximators to approximate the conditional mean function $f(\cdot)$ in (1.1). As the sample size increases, the dimension of the parameter space also increases to achieve greater approximation accuracy. Thereby it is possible to allow $f(\cdot)$ to depend on infinitely many lagged variables although only a finite number of lags is considered for any given finite sample size. The approaches of this type differ in the class of parametric functions used. We begin with simple linear functions where just the number of lags in the model grows with the sample size. For this class it is particularly easy to discuss the assumptions usually made for deriving asymptotic properties of estimators. Then we consider neural networks as an important general class of nonlinear approximators.

Linear Functions

Suppose $\{X_t\}$ is a zero mean purely nondeterministic causal stationary process, then it has an AR representation of potentially infinite order,

$$\dot{X}_t = \sum_{i=1}^{\infty} \alpha_i X_{t-i} + \varepsilon_t.$$

If the second order moment properties of the process are of interest only it suffices to obtain the above representation which is linear in lagged X_t . Hence, the second order moment properties of the process may be estimated by approximating its infinite order AR representation. The simplest way to accomplish this is by fitting finite order AR(H_n) processes

$$X_t = \sum_{i=1}^{H_n} \alpha_i X_{t-i} + \varepsilon_{H_n,t},$$

where the order H_n is an increasing function of the sample size n . To obtain desirable properties of the resulting estimators and quantities derived from them we need to assume that the AR order H_n goes to infinity at a much smaller rate than n so that there is eventually enough information for estimating the parameters efficiently. On the other hand, the approximation quality must improve sufficiently rapidly so as to avoid large bias. Hence, there must be an appropriate lower bound on the rate of divergence of H_n . More precisely, it may be assumed that

$$(1.) H_n \text{ is } o(n^{1/3}), \text{ and}$$

$$(2.) \sqrt{n} \sum_{i>H_n} |\alpha_i| \rightarrow 0,$$

as $n \rightarrow \infty$. Here the two conditions are upper and lower bounds, respectively, on the rate at which the AR order goes to infinity with n . Under these conditions and mild assumptions for $\{\varepsilon_t\}$ the least squares estimators of the α_i are consistent and asymptotically normal. In fact, for consistency weaker conditions for H_n suffice.

Akaike (1969), Parzen (1974), Berk (1974) and Bhansali (1978) use this approach for spectral estimation and prediction of univariate processes. Parzen (1977), Lewis & Reinsel (1985), Lütkepohl (1991, Ch. 9) and Lütkepohl & Poskitt (1996) discuss multivariate extensions. They also consider estimation of other quantities derived from the autoregressive coefficients. Most of these results can be extended to nonstationary integrated and cointegrated processes (see Section 5.3).

Note that $\hat{X}_t = \sum_{i=1}^{\infty} \alpha_i X_{t-i}$ is the best (minimum MSE) *linear* 1-step predictor which may not be the conditional expectation and, hence, it may not be the optimal predictor in a more general class of nonlinear predictors. Consequently, it may be desirable to consider nonlinear functions $f_n(\cdot)$ to approximate the conditional mean function $f(\cdot)$. We will present one possible nonlinear approach next.

Neural Networks

Neural networks have been used in various fields to approximate complex nonlinear structures. Their name comes from the fact that they may be thought of as a network of neurons similar to (but of course much simpler as) the brain. The related computations may be extremely complex. Therefore neural network analysis nowadays represents a subfield of computer science or, more precisely, of artificial intelligence. Here we consider the *single hidden layer feedforward network* which may be best thought of as a class of flexible nonlinear functions of the form

$$f_n(X_{t-1}, \dots, X_{t-p}) = \beta_0 + \sum_{j=1}^q G(\gamma_{0j} + Y_t^T \gamma_j) \beta_j, \quad (3.9)$$

where $Y_t = (X_{t-1}, \dots, X_{t-p})^T$ and the $\gamma_j = (\gamma_{1j}, \dots, \gamma_{pj})^T$ are $(p \times 1)$ vectors for $j = 1, \dots, q$, and $\beta_0, \beta_1, \dots, \beta_q$ are scalar coefficients. The function $G: \mathbb{R} \rightarrow [0, 1]$ is a prespecified cumulative distribution function. Typical examples are the logistic function $G(x) = 1/(1 + e^{-x})$ and the hyperbolic function $G(x) = \tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$. Functions of the type (3.9) can approximate broad classes of functions if q is sufficiently large. Thus, if q increases with the sample size n , a good approximation of $f(X_{t-1}, \dots, X_{t-p})$ will eventually result. The function in (3.9) may also be estimated without specifying $G(\cdot)$ by using the projection pursuit regression of Hutchinson, Lo & Poggio (1994). In the following we will, however, assume a given specific form of $G(\cdot)$.

For practical purposes it will be advantageous to obtain a good approximation with small or moderate values of q . Therefore adding a linear AR term in (3.9) is often useful. Thus, in practice, a possible approximating function is

$$f_n(X_{t-1}, \dots, X_{t-p}) = \beta_0 + \sum_{i=1}^p \alpha_i X_{t-i} + \sum_{j=1}^q G(\gamma_{0j} + Y_t^T \gamma_j) \beta_j.$$

For given p and q , estimation of the parameters of this model is possible with LS procedures. Asymptotic properties of the resulting estimators are available both for fixed q and q increasing with the sample size. Kuan & White (1994) provide a comprehensive survey of neural network models and estimation results for the present situation. Also it is possible to let the number of lags p (i.e., the AR order) increase with the sample size. This, however, results in further complications of the asymptotic theory.

Since nonlinear optimization algorithms may be time consuming, it is undesirable to reestimate a model each time new observations become available. Therefore sequential estimation or learning procedures have been proposed which update the available estimates sequentially when new sample information becomes available. A prominent example is the backpropagation procedure (see Rumelhart, Hinton & Williams 1986). Kuan & White (1994) present asymptotic results for this procedure as well.

The network represented by (3.9) feeds the output of the neurons (the $G(\cdot)$) directly into the overall output and there is also no direct interaction between the neurons. There are various generalizations of this simple architecture. For instance, *multi-layer networks* may be considered. An example of a

2-layer network is

$$f_n(X_{t-1}, \dots, X_{t-p}) = \beta_0 + \sum_{j=1}^{q_2} G_2 \left(\sum_{l=1}^{q_1} G_1(\gamma_{0l} + Y_t^T \gamma_l) \beta_l \right) \delta_j,$$

where $G_1(\cdot)$ and $G_2(\cdot)$ are now prespecified cumulative distribution functions and the γ_{ij} , β_k and δ_j are unknown parameters which have to be estimated. Another possible extension would be to allow for feedback between the neurons. The following is an example of a *recurrent single hidden layer network*:

$$f_{n,t}(X_{t-1}, X_{t-2}, \dots, X_0) = \beta_0 + \sum_{j=1}^q \phi_{tj} \beta_j, \quad t = 0, 1, 2, \dots$$

where

$$\phi_{tj} = G(z_t^T \gamma_j + \sum_{l=1}^q \phi_{t-1,l} \delta_{lj}), \quad j = 1, 2, \dots, q.$$

Although the simpler single hidden layer feedforward networks have quite general approximation properties it may be useful in practice to consider more sophisticated architectures to obtain a good approximation with fewer terms (or neurons) than that in (3.9). Also there may be information on the structure of a data generation mechanism that suggests multi-layer or feedback architectures.

In practice there will often be uncertainty regarding the most suitable architecture for a given time series and regarding the number of lags and neurons that guarantee a good approximation of the actual generation mechanism. Therefore methods have been proposed for model selection and for deciding on restrictions that may be imposed on a given neural network model. For instance, Murata, Yoshizawa & Amari (1994) proposed a model selection criterion which extends the ideas underlying the AIC criterion to the present situation. Specification tests are also reviewed by Kuan & White (1994).

As mentioned earlier, neural networks establish a subfield of computer science and are applied in many areas. Therefore it is impossible to provide a complete survey of the literature in a limited review of this type. Those interested in this fascinating tool for nonparametric time series analysis may find the survey article by Kuan & White (1994) a useful point of departure for further studies.

4 Estimating Higher Order Conditional Moments and Densities

Techniques similar to those discussed for estimating the conditional expectation of a process may also be used for approximating higher order conditional moments which are often of interest, as we have argued earlier. Here we summarize some of these extensions. We begin with methods for estimating conditional variances in addition to conditional means. Then some possibilities for approximating the complete conditional density are presented.

4.1 Conditional Variances

Nonparametric Kernel Estimation

Auestad & Tjøstheim (1990) and Tjøstheim & Auestad (1994a,b) use kernel estimation techniques for analyzing models like (1.2) assuming that both the conditional mean and the conditional variance function depend on at most p lagged X_t . The function $f(\cdot)$ may again be estimated by the Nadaraya-Watson estimator with product kernels as in Section 3.1,

$$\hat{f}(x_1, \dots, x_p) = \frac{\sum_{t=p+1}^n \prod_{i=1}^p K\{(x_i - X_{t-i})/h_i\} X_t}{\sum_{t=p+1}^n \prod_{i=1}^p K\{(x_i - X_{t-i})/h_i\}},$$

and the conditional variance $g(\cdot)^2$ may be estimated by

$$\hat{g}(x_1, \dots, x_p)^2 = \frac{\sum_{t=p+1}^n \prod_{i=1}^p K\{(x_i - X_{t-i})/h_i\} X_t^2}{\sum_{t=p+1}^n \prod_{i=1}^p K\{(x_i - X_{t-i})/h_i\}} - \{\hat{f}(x_1, \dots, x_p)\}^2,$$

where again $K(\cdot)$ is a kernel function with bounded support and the h_i 's are the bandwidths.

Masry & Tjøstheim (1995a) show strong consistency and asymptotic normality of these estimators for α -mixing observations and Tjøstheim & Auestad (1994a,b) consider model specification and lag selection in models of the form (1.2).

Local Polynomial Regression and Other Techniques

Local polynomial nonparametric regression techniques can be used in an analogous fashion to estimate the conditional mean and variance functions. Assume $p = 1$ so that the functions $f(\cdot)$ and $g(\cdot)$ depend on X_{t-1} only. Then they may be estimated by minimization of

$$c_n(x) = \arg \min_{c \in \mathbb{R}^l} \sum_{t=1}^n (X_t - c^T U_{tn})^2 K\{(X_{t-1} - x)/h\}$$

as in Section 3.1, and

$$s_n(x) = \arg \min_{s \in \mathbb{R}^l} \sum_{t=1}^n (X_t^2 - s^T U_{tn})^2 K\{(X_{t-1} - x)/h\}$$

where h is again a positive bandwidth, and

$$U_{tn} = F(u_{tn}), \quad F(u) = (1, u, \dots, u^{l-1}/(l-1)!)^T, \quad u_{tn} = (X_{t-1} - x)/h.$$

Here the degree of the approximating polynomial is assumed to be $l - 1$. The estimators $\hat{f}(x)$ and $\hat{g}(x)$ are given by

$$\hat{f}(x) = c_n(x)^T F(0) \quad \text{and} \quad \hat{g}(x) = s_n(x)^T F(0) - \{c_n(x)^T F(0)\}^2.$$

Härdle & Tsybakov (1996) prove asymptotic normality of these estimators under similar conditions as in Section 3.1 where the conditional mean was estimated only.

An extension of this model to nonparametric vector autoregression is presented in Härdle, Tsybakov & Yang (1996) who consider the model

$$X_t = f(Y_t) + \Sigma^{1/2}(Y_t)\varepsilon_t, \quad t = p, p + 1, \dots,$$

where $X_t = (X_{t1}, X_{t2}, \dots, X_{td})^T \in \mathbb{R}^d$, $\varepsilon_t = (\varepsilon_{t1}, \varepsilon_{t2}, \dots, \varepsilon_{td})^T \in \mathbb{R}^d$ and $Y_t = (X_{t-1}, X_{t-2}, \dots, X_{t-p}) \in \mathbb{R}^{d \times p}$ is a matrix of lagged variables.

Alternatively, conditional heteroscedasticity can also be modeled with neural network methods (Weigend & Nix 1994).

4.2 Estimating the Predictive Density

Kernel Techniques

For a stationary time series, Robinson (1983b) proposed a kernel estimator to estimate the one-step-ahead transition density $h(y | x)$. Note that $h(y | x) = p(x, y)/p(x)$, where $p(x, y)$ is the joint density of (X_t, X_{t+1}) and $p(x)$ is the marginal density of X_t . Replacing the terms on the right-hand

side with corresponding kernel estimators, we have

$$\hat{p}(y | x) = \frac{(nh^2)^{-1} \sum_{i=1}^{n-1} K_2[\{(x, y) - (X_i, X_{i+1})\}/h]}{(nh)^{-1} \sum_{i=1}^n K[(x - X_i)/h]}$$

where $K_2(\cdot)$ is a bivariate kernel function, commonly of the product form $K_2(u, v) = K(u)K(v)$. Note that the estimation of the transition density allows us to construct nonparametric multi-step-ahead prediction density functions as well. For extensions see Singh & Ullah (1985).

Hermite Expansion Approach

Gallant & Tauchen (1989) used Hermite expansions to approximate the one-step-ahead conditional density of the process given its past. This approach is based on the fact that a large class of density functions, $h(y)$ say, is proportional to $[P(z)]^2 \phi(z)$, where $z = (y - \mu_y)/\sigma_y$, with μ_y and σ_y location and scale parameters of the distribution, respectively, $P(z) = 1 + \psi_1 z + \dots + \psi_r z^r$ is a polynomial of possibly infinite degree r and $\phi(z) = (2\pi)^{-1} \exp(-z^2/2)$ is the standard normal density. Dividing $[P(z)]^2 \phi(z)$ by a normalizing constant this is just the Hermite expansion of $h(y)$. Hence, the density may be written as the product of a standard normal density and the square of a polynomial.

In the present situation we are interested in the conditional density $h(x_t | x_{t-1}, x_{t-2}, \dots)$. By the foregoing considerations we have

$$h(x_t | x_{t-1}, x_{t-2}, \dots) \propto [P(z_t)]^2 \phi(z_t)$$

where $z_t = (x_t - \mu_t)/\sigma_t$ with μ_t and σ_t being location and scale parameters, respectively, of the conditional distribution. The former is assumed to be a linear function of the past, $\mu_t = \nu + \alpha_1 X_{t-1} + \dots + \alpha_p X_{t-p}$, and the latter may be modeled as

$$\sigma_t = \rho_0 + \rho_1 |X_{t-1}| + \dots + \rho_q |X_{t-q}|.$$

The specification of the conditional scale parameter σ_t is similar but not identical to an ARCH process as originally proposed by Engle (1982). Alternative specifications may be used here. At any rate, the location and scale parameters μ_t and σ_t are modeled parametrically whereas higher order moment terms are captured by the polynomial. Letting the polynomial degree increase with the sample size makes this approach nonparametric. Overall the approach has been termed semi nonparametric (SNP) because it combines parametric with nonparametric elements.

To achieve a flexible adjustment of the model to higher order dynamics the polynomial coefficients ψ_1, \dots, ψ_r may be made dependent on the past, that is,

$$\begin{aligned} \psi_j(x_{t-1}, \dots, x_{t-K}) &= \psi_{j0} + \sum_{k=1}^K \psi_{j1}^{(k)} x_{t-k} + \sum_{k=1}^K \sum_{h=1}^K \psi_{j2}^{(kh)} x_{t-k} x_{t-h} + \dots \\ &+ \sum_{k=1}^K \dots \sum_{h=1}^K \psi_{jl}^{(\dots)} x_{t-k} \dots x_{t-h} \end{aligned}$$

where usually small values of K and l are sufficient to guarantee a rich dynamic structure. Of course, for $r = K = l = 0$ we get

$$h(x_t | x_{t-1}, x_{t-2}, \dots) \propto \phi \left(\frac{x_t - \nu - \alpha_1 x_{t-1} - \dots - \alpha_p x_{t-p}}{\sigma_t} \right)$$

so that we have a linear AR(p) process with conditionally heteroscedastic error term.

For given values of p, q, r, K and l the parameters of the model may be estimated by maximum

likelihood which is easily accomplished by minimizing the normalized negative log likelihood

$$L(\theta) = -\frac{1}{n} \sum_{t=1}^n \log h(X_t | X_{t-1}, \dots, X_{t-p}; \theta).$$

Asymptotic properties of this estimation procedure are given by Gallant & Nychka (1987) who allow the order of the Hermite expansion to increase with the sample size. In principle, an extension of this approach to the multivariate case is possible (see Gallant & Tauchen 1989).

5 Other Nonparametric Techniques for Time Series

5.1 Density Estimation with Correlated Observations

Kernel Methods

There is a rich literature on density estimation for independent observations, see Silverman (1986) and the references therein. A popular method is the kernel estimator of the form (3.1) where the kernel function $K(\cdot)$ is typically a probability density function. The key in density estimation is the bandwidth selection. A number of different methods have been proposed, including the cross-validation (Rudemo 1982, Bowman 1984) and the plug-in rules of Sheather (1983), Park & Marron (1990) and Park & Turlach (1992).

The earliest work on density estimation for stationary processes is that of Roussas (1969) and Rosenblatt (1970). The properties of the kernel estimator for dependent observations were investigated by Robinson (1983b) and Hall & Hart (1990a). They found that the bias of the estimator is not affected by the serial correlation. However, the variance is affected. The cross-validation method for dependent observations is studied by Hart & Vieu (1990), under certain regularity conditions. Detailed information and references can be found in Györfi, Härdle, Sarda & Vieu (1989), Prakasa Rao (1983) and Hart (1996). Density estimation for long range dependent data was studied by Hall, Lahiri & Truong (1994) and Csörgő & Mielniczuk (1995a).

Testing for Serial Dependence

Kernel density estimation techniques may also be used to test for independence, for instance, in checking the residual behavior of an estimated nonlinear time series model. Skaug & Tjøstheim (1993) proposed a nonparametric test for independence between two variables which is suitable in this situation. They propose to estimate the quantity

$$I = \int \{p(x, y) - p_1(x)p_2(y)\}^2 p(x, y) w(x, y) dx dy$$

where $p(x, y)$ is the joint density and $p_1(\cdot)$, $p_2(\cdot)$ are the marginal densities while $w(\cdot, \cdot)$ is a weight function with compact support. Using kernel density estimators, we obtain

$$\hat{I} = \frac{1}{n} \sum_t \{\hat{p}(X_t, Y_t) - \hat{p}_1(X_t) \hat{p}_2(Y_t)\}^2 w(X_t, Y_t).$$

which should be small under the null hypothesis that X and Y are independent and which can therefore be the basis for an independence test.

5.2 Bootstrap Methods

The bootstrap method is an important nonparametric tool which has also been used for time series analysis in a number of different ways. For instance, it may be used for assessing and improving

the properties of estimators and forecasts. Originally it was proposed for independent observations (Efron & Tibshirani 1993). Therefore an obvious extension to time series analysis is to bootstrap the residuals of some model. This approach has been used in many applications. Efron & Tibshirani (1993) discuss estimating the standard errors of linear autoregressive parameter estimates using this approach. Bose (1988) evaluates the distribution of the parameter estimator of an AR(1) model by the bootstrap and Kreiss & Franke (1992) discuss its extensions to ARMA(p, q) processes. Furthermore, Franke & Härdle (1992) propose a bootstrap method for spectral estimation.

It is also possible to apply a bootstrap directly to the time series observations by sampling blocks of observations rather than individual ones. This method is known as the moving blocks bootstrap. Specifically, given a time series X_1, \dots, X_n , all possible blocks of $l < n$ consecutive observations are considered and random samples of blocks are drawn and joint together to form a bootstrap time series of roughly length n . This process is repeated B times so that B bootstrap time series are obtained. These artificial series may be used to investigate the distributional properties of the original time series. The moving blocks bootstrap for time series was introduced by Künsch (1989) and Liu & Singh (1992). An introductory exposition is given by Efron & Tibshirani (1993, Sec. 8.6).

5.3 Trend Analysis

In much of the previous discussion we have assumed stationary processes. In practice many time series have trends and are therefore nonstationary. These trends may be removed prior to an analysis of the stationary part of the process if the trend function is known. In most cases it is unknown, however. In that situation nonparametric techniques may be used for trend estimation or trend elimination.

Estimating Trend Functions

Here we consider the case when the trend is characterized by a smooth deterministic function. Suppose X_1, \dots, X_n is a possibly nonstationary time series with trend $\mu(t) = E(X_t)$. Under the assumption that the trend is smooth, a traditional way of estimating the trend function is the running mean estimator described in Chatfield (1974). A more recent proposal is due to Hart (1991) who uses the kernel smoother of Gasser & Müller (1979) of the form

$$\hat{\mu}_{th} = \frac{1}{h} \sum_{i=1}^n X_i \int_{(i-1)/n}^{i/n} K\left(\frac{(t-0.5)/n-u}{h}\right) du$$

for trend estimation. Hart (1994) proposed a method called time series cross-validation for selecting the bandwidth h . He noted that the ordinary leave-one-out cross-validation tends to select a bandwidth many orders of magnitude too small, if the data are highly positively correlated.

Nonparametric Regression with Dependent Errors

Consider the fixed-design regression model

$$X_{in} = m(z_{in}) + \varepsilon_{in}$$

where $z_{in} = i/n$ and the errors $\{\varepsilon_{in}\}$ are correlated, both the Gasser & Müller (1979) estimator

$$\hat{m}_n(z) = \frac{1}{h_n} \sum_{i=1}^n X_i \left\{ \int_{z_{(i-1)n}}^{z_{in}} K\left(\frac{z-t}{h_n}\right) dt \right\}$$

and the Nadaraya–Watson type estimator

$$\hat{m}_n(z) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{z - z_{in}}{h_n}\right) X_i$$

have been proposed and studied. See Hart & Wehrly (1986) and Härdle (1990). Hall & Hart (1990b) and Csörgő & Mielniczuk (1995b) studied the same problem with long-range dependent errors.

Truong & Patil (1996) propose to use wavelet methods to estimate possibly discontinuous trends. Wavelet estimators have been shown to have extraordinary adaptability in handling discontinuity of the underlying function with independent observations (Donoho & Johnstone 1992, Donoho *et al.* 1995, and Hall & Patil 1995). They may be equally powerful in time series analysis.

Nonparametric Unit Root and Cointegration Tests

As an alternative to a deterministic trend, a time series may have a stochastic trend which can be removed by differencing. A process is said to be integrated of order d , $I(d)$, if a stochastic trend can be removed by differencing d times. For example, a random walk $X_t = X_{t-1} + \varepsilon_t$ with white noise error process ε_t is $I(1)$ because $X_t - X_{t-1} =: \Delta X_t = \varepsilon_t$. Nonparametric tests can be used for checking the order of integration of a process.

The random walk is the simplest version of a stochastic trend. Fuller (1976) and Dickey & Fuller (1979) therefore consider an AR(1) model

$$X_t = \rho X_{t-1} + \varepsilon_t \quad (5.1)$$

and test $H_0 : \rho = 1$ against $H_1 : \rho < 1$. An obvious test statistic is the t -ratio based on the LS estimator $\hat{\rho}$ of ρ :

$$t_{\hat{\rho}} = \frac{\hat{\rho} - 1}{s_{\hat{\rho}}}$$

where $s_{\hat{\rho}}$ is the usual estimator of the standard error of $\hat{\rho}$. Equivalently, this statistic may be obtained as the t -ratio of the parameter estimator in the model

$$\Delta X_t = \alpha X_{t-1} + \varepsilon_t$$

where $\alpha = \rho - 1$. The resulting test is also known as Dickey–Fuller (DF) test. The t -statistic does not have the usual standard normal limiting distribution but it has a nonstandard distribution for which the relevant critical values have been tabulated in Fuller (1976).

In practice, the model (5.1) is often too limited to be a reasonable approximation to the underlying data generating process. Therefore more general assumptions are often made for the error process $\{\varepsilon_t\}$. For instance, it may be assumed to be a stationary process. Ignoring the dependency of the ε_t in that case in constructing the test statistic may result in a badly biased test. Therefore nonparametric techniques are often used to model the dependence of the ε_t . One possible approach fits autoregressions

$$\Delta X_t = \alpha X_{t-1} + \pi_1 \Delta X_{t-1} + \cdots + \pi_H \Delta X_{t-H} + \varepsilon_t \quad (5.2)$$

where H goes to infinity with the sample size (see Said & Dickey 1984). Alternatively, a correction for the t -statistic based on spectral techniques has been proposed by Phillips & Perron (1988).

Tests of the foregoing type are often referred to as unit root tests. There is an extensive literature on these tests. Extensions allow also for deterministic terms such as intercepts and linear time trends (see Hamilton 1994, Chapter 17, for details). Also tests of the null hypothesis of a stationary process against the alternative of a unit root have been proposed (see Kwiatkowski, Phillips, Schmidt & Shin 1992). Again spectral techniques are used in the latter variant of a unit root test to account for higher order dynamics of the data generating process.

Multivariate extensions of the DF tests were proposed by Johansen (1989, 1991). In a multivariate AR process, unit roots indicate that some or all of the components are integrated variables. There may be linear combinations of the variables, however, which are stationary or integrated of lower order. This phenomenon is known as cointegration. Therefore unit root tests in multivariate processes are treated under the heading of testing for cointegration. Nonparametric variants of the Johansen tests are considered by Saikkonen & Luukkonen (1997) who approximate the stationary part of the process by autoregressions of growing order when the sample size increases analogously to (5.2). Cointegration tests based on spectral techniques are discussed by Stock & Watson (1988).

Further nonparametric generalizations of unit root tests are obtained by assuming that there may be an AR unit root in some unknown nonlinear monotone transformation of the original variables. To check the existence of such a unit root in the data generating process, DF or other unit root tests based on the ranks of X_t may be used (see Granger & Hallman 1991, Campbell & Dufour 1993, Breitung & Gouriéroux 1997).

5.4 Adaptive Estimation

In a model with finite dimensional parameter vector of interest θ , say, and an infinite dimensional nuisance parameter vector ψ , say, the latter is often taken care of with nonparametric methods. If that is done in such a way that the estimator for θ is asymptotically efficient, it is said to be estimated adaptively. In time series models the conditional mean and variance functions are often of foremost interest. They are therefore often parameterized in a specific way, for instance, as a linear function of the past. The remaining parts of the data generating process may then be estimated nonparametrically. A number of authors have discussed adaptive methods in this context (e.g., Linton 1993, Kreiss 1987, Robinson 1988, Steigerwald 1992, Engle & Gonzáles-Rivera 1991, Werker 1995, Drost, Klaassen & Werker 1994).

References

- Akaike, H. (1969). Power spectrum estimation through autoregressive model fitting. *Annals of the Institute of Statistical Mathematics*, **21**, 407–419.
- Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203–217.
- Akaike, H. (1974). A New Look at Statistical Model Identification, *IEEE Transactions on Automatic Control*, **AC-19**, 716–722.
- Auestad, B. & Tjøstheim, D. (1990). Identification of nonlinear time series: First order characterization and order estimation, *Biometrika*, **77**, 669–687.
- Bartlett, M.S. (1950). Periodogram analysis and continuous spectra. *Biometrika*, **37**, 1–16.
- Becker, R.A., Chambers, J. M. & Wilks, A. R. (1988). *The New S Language*. New York: Chapman and Hall.
- Berk, K.N. (1974). Consistent autoregressive spectral estimates. *Annals of Statistics*, **2**, 489–502.
- Bhansali, R.J. (1978). Linear prediction by autoregressive model fitting in the time domain. *Annals of Statistics*, **6**, 224–231.
- Bierens, H.J. (1983). Uniform consistency of kernel estimators of a regression function under generalized conditions. *Journal of the American Statistical Association*, **78**, 699–707.
- Bierens, H.J. (1987). Kernel estimators of regression functions, in T.F. Bewley (ed.) *Advances in Econometrics: Fifth World Congress*, Vol. 1. Cambridge: Cambridge University Press.
- Bierens, H.J. (1994). *Topics in Advanced Econometrics: Estimation, testing, and specification of cross-section and time series models*. Cambridge: Cambridge University Press.
- Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley.
- Blackman, R.B. & Tukey, J.W. (1959). *The Measurement of Power Spectrum from the Point of View of Communications Engineering*. New York: Dover.
- Bose, A. (1988). Edgeworth correction by bootstrap in autoregressions. *Annals of Statistics*, **16**, 1709–1722.
- Bowman, A.W. (1994). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**, 353–360.
- Breitung, J. & Gouriéroux, C. (1997). Rank tests for unit roots. *Journal of Econometrics*, forthcoming.
- Brillinger, D.R. & Krishnaiah, P.R. (Ed.) (1983). *Handbook of Statistics 3, Time Series in the Frequency Domain*. Amsterdam: North-Holland.
- Campbell, B. & Dufour, J.-M. (1993). Exact nonparametric orthogonality and random walk tests. Working paper, C.R.D.E., Montreal.

- Chan, K.S. & Tong, H. (1986). On estimating thresholds in autoregressive models. *Journal of Time Series Analysis*, **7**, 179–190.
- Chatfield, C. (1984). *The Analysis of Time Series: An Introduction*, 3rd ed. London: Chapman and Hall.
- Chen, R. (1996). A nonparametric multi-step prediction estimator in Markovian structures. *Statistica Sinica*, **6**, 603–615.
- Chen, R. & Hafner, C. (1995). Nonlinear time series analysis, in *XploRe, an Interactive Statistical Computing Environment*, (ed. Härdle, W., Klinke, S. & Turlach, B.). Heidelberg: Springer Verlag.
- Chen, R., Härdle, W., Linton, O.B. & Severance-Lossin, E. (1996). Estimation in additive nonparametric regression. In *COMPSTAT meeting Semmering*, Härdle, W. & Schimek, M. (eds), Physika Verlag.
- Chen, R., Liu, J.S. & Tsay, R.S. (1995). Additivity tests for nonlinear autoregressive models. *Biometrika*, **82**, 369–383.
- Chen, R. & Tsay, R. S. (1993a). Nonlinear additive ARX models. *Journal of the American Statistical Association*, **88**, 955–967.
- Chen, R. & Tsay, R. S. (1993b). Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, **88**, 298–308.
- Cheng, B. & Tong, H. (1992). On consistent non-parametric order determination and chaos (with discussion). *Journal of the Royal Statistical Society, Series B*, **54**, 427–474.
- Cleveland, W. S. & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**, 596–610.
- Collomb, G. & Härdle, W. (1986). Strong uniform convergence rates in robust nonparametric time series analysis and prediction: Kernel regression estimation from dependent observations. *Stochastic Processes and their Applications*, **23**, 77–89.
- Csörgő, S. & Mielniczuk, J. (1995a). Density estimation under long-range dependence. *Ann. Statist.* **23**, 990–999.
- Csörgő, S. & Mielniczuk, J. (1995b). Nonparametric regression under long-range dependent normal errors. *Ann. Statist.* **23**, 1000–1014.
- Dahlhans, R. (1993). *Fitting time series models to nonstationary processes*. Institut für Angewandte Mathematik, Universität Heidelberg.
- Dickey, D.A. & Fuller, W.A. (1979). Distribution of the estimators for autoregressive time series with unit root. *Journal of the American Statistical Association*, **74**, 427–431.
- Diebold, F. & Nason, J. (1990). Nonparametric exchange rate prediction. *Journal of International Economics*, **28**, 315–332.
- Diebolt, J. & Guegan, D. (1990). Probabilistic properties of the general nonlinear autoregressive process of order one. *Technical report, N° 128*, L.S.T.A., Université de Paris VI.
- Donoho, D.L. & Johnstone, I.M. (1992). Minimax estimation via wavelet shrinkage. *Technical Report 402*, Dept. Stat., Stanford University.
- Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. & Picard, D. (1995). Wavelet Shrinkage: Asymptopia? (with discussion). *Journal of the Royal Statistical Society, Series B*, **57**, 301–369.
- Drost, F.C., Klaassen, C.A.J. & Werker, B.J.M. (1994). Adaptive estimation in time series models. CentER Discussion Paper 9488, Tilburg University.
- Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica*, **50**, 987–1008.
- Engle, R.F. & Gonzáles-Rivera, G. (1991). Semiparametric ARCH models. *Journal of Business and Economic Statistics*, **9**, 345–359.
- Fan, J. (1993). Local linear regression and their minimax efficiency. *Annals of Statistics*, **21**, 196–216.
- Franke, J. & Härdle, W. (1992). On bootstrapping kernel spectral estimates. *Annals of Statistics*, **20**, 121–145.
- Friedman, J.H. (1988). Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, **19**, 1–141.
- Fuller, W.A. (1976). *Introduction to Statistical Time Series*. New York: Wiley.
- Gallant, A.R. & Nychka, D.W. (1987). Semiparametric maximum likelihood estimation. *Econometrica*, **55**, 363–390.
- Gallant, A.R. & Tauchen, G.E. (1989). Semiparametric estimation of conditional constrained heterogeneous processes: Asset pricing applications, *Econometrica*, **57**, 1091–1120.
- Gasser, T. & Müller, H.G. (1979). Kernel estimation of regression functions. In *Smoother Techniques for Curve Estimation*, eds. T. Gasser & M. Rosenblatt, 23–68.
- Granger, C.W.J. & Hallman, G. (1991). Nonlinear transformations of integrated time series. *Journal of Time Series Analysis*, **12**, 207–224.
- Granger, C.W.J. & Teräsvirta, T. (1993). *Modeling Nonlinear Economic Relationships*. Oxford: Oxford University Press.
- Györfi, L., Härdle, W., Sarda, P. & Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series*. Lecture Notes in Statistics 60. Heidelberg: Springer-Verlag.
- Haggan, V. & Ozaki, T. (1981). Modeling nonlinear vibrations using an amplitude-dependent autoregressive time series model. *Biometrika*, **68**, 189–196.
- Hall, P. & Hart, J.D. (1990a). Convergence rates in density estimation for data from infinite-order moving average processes. *Probability Theory and Related Fields*, **87**, 253–274.
- Hall, P. & Hart, J.D. (1990b). Nonparametric regression with long-range dependence. *Stochastic Processes and their Applications*, **36**, 339–351.
- Hall, P., Lihiri, S.N. & Truong, Y.K. (1994). On bandwidth choice for density estimation with dependent data. *Manuscript*.

- Hall, P. & Patil, P. (1995). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Ann. Statist.* **23**, 905–928.
- Hamilton, J.D. (1994). *Time Series Analysis*. Princeton: Princeton University Press.
- Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.
- Härdle, W. & Hall, P. (1993). On the backfitting algorithm for additive regression models. *Statistica Neerlandica*, **47**, 43–57.
- Härdle, W., Klinke, S. & Turlach, B. (1995). *XploRe, an Interactive Statistical Computing Environment*. Heidelberg: Springer-Verlag.
- Härdle, W. & Tsybakov, A.B. (1997). Local polynomial estimators of the volatility function. *J. Econometrics*, to appear.
- Härdle, W. & Tsybakov, A.B. & Yang, L. (1996). Nonparametric vector autoregression. *Journal of Statistical Planning and Inference*, to appear.
- Härdle, W. & Vieu, P. (1992). Kernel regression smoothing of time series. *Journal of Time Series Analysis*, **13**, 209–232.
- Hart, J.D. (1991). Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society, Series B*, **53**, 173–187.
- Hart, J.D. (1994). Automated kernel smoothing of dependent data by using time series cross-validation. *Journal of the Royal Statistical Society, Series B*, **56**, 529–542.
- Hart, J.D. (1996). Some automated methods of smoothing time-dependent data. *J. Nonparametric Statistics*, **6**, 115–142.
- Hart, J.D. & Vieu, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Annals of Statistics*, **18**, 873–890.
- Hart, J.D. & Wehrly, T.E. (1986). Kernel regression estimation using repeated measurement data. *Journal of the American Statistical Association*, **81**, 1080–1088.
- Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*, Vol. 43 of *Monographs on Statistics and Applied Probability*. London: Chapman and Hall.
- Hjellvik, V. & Tjøstheim, D. (1995). Nonparametric tests of linearity for time series. *Biometrika*, **82**, 351–368.
- Hutchinson, J.M., Lo, A.W. & Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. *Journal of Finance*, **49**, 851–889.
- Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, **12**, 231–254.
- Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica*, **59**, 1551–1580.
- Jones, D. A. (1978). Non-linear autoregressive processes. *Journal of the Royal Statistical Society, Series A*, **360**, 71–95.
- Katkovnik, V. Y. (1979). *Linear and nonlinear methods for nonparametric regression analysis (in Russian)*. Avtomatika i Telemekhanika, 35–46.
- Kreiss, J.-P. (1987). On adaptive estimation in stationary ARMA processes. *Annals of Statistics*, **15**, 112–133.
- Kreiss, J.-P. and Franke, J. (1992). Bootstrapping stationary autoregressive moving-average models. *Journal of Time Series Analysis*, **13**, 297–317.
- Kuan, C.-M. & White, H. (1994). Artificial neural networks: An econometric perspective. *Econometric Reviews*, **13**, 1–91.
- Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, **17**, 1217–1241.
- Kwiatkowski, D., Phillips, P.C.B., Schmidt, P. & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, **54**, 159–178.
- Lewis, R. & Reinsel, G.C. (1985). Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis*, **16**, 393–411.
- Lewis, P.A.W. & Stevens, G. (1991). Nonlinear modeling of time series using multivariate adaptive regression splines (MARS). *Journal of the American Statistical Association*, **87**, 864–877.
- Linton, O. (1993). Adaptive estimation in ARCH models. *Econometric Theory*, **9**, 539–569.
- Linton, O. & Nielsen, J.P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, **82**, 93–100.
- Liu, R.Y. and Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the Limits of Bootstrap*. R. Lepage & L. Billard (eds.). New York: Wiley, 225–248.
- Lütkepohl, H. (1991). *Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.
- Lütkepohl, H. & Poskitt, D.S. (1996). Testing for causation using infinite order vector autoregressive processes. *Econometric Theory*, **12**, 61–87.
- Masry, E. & Tjøstheim, D. (1995a). Nonparametric estimation and identification of nonlinear ARCH time series: strong convergence and asymptotic normality. *Econometric Theory*, **11**, 258–289.
- Masry, E. & Tjøstheim, D. (1995b). Additive nonlinear ARX time series and projection estimates. *Econometric Theory*, to appear.
- Murata, N., Yoshizawa, S. & Amari, S. (1994). Network information criterion – determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, **5**, 865–871.
- Park, B.U. & Marron, J.S. (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, **85**, 66–72.
- Park, B.U. & Turlach, B. (1992). Practical performance of several data driven bandwidth selectors (with discussion). *Computational Statistics*, **7**, 251–270.
- Parzen, E. (1961). Mathematical considerations in the estimation of spectra. *Technometrics*, **3**, 167–190.
- Parzen, E. (1974). Some recent advances in time series modeling. *IEEE Transactions on Automatic Control*, **AC-19**, 723–730.
- Parzen, E. (1977). Multiple time series: Determining the order of approximating autoregressive schemes. In *Multivariate Analysis-IV*, P.R. Krishnaiah (ed.). Amsterdam: North-Holland, 389–409.
- Pham, D.T. (1985). Bilinear Markovian representations and bilinear models. *Stochastic Process. Appl.* **20**, 295–306.
- Phillips, P.C.B. & Perron, P. (1988). Testing for a unit root in time series analysis. *Biometrika*, **75**, 335–346.

- Prakasa Rao, B.L.S. (1983). *Nonparametric Functional Estimation*. Orlando, FL: Academic Press.
- Priestley, M.B. (1981). *Spectral Analysis and Time Series*. London: Academic Press.
- Priestley, M. B. (1988). *Non-linear and Non-stationary Time Series Analysis*, New York: Academic Press.
- Priestley, M.B. (1996). Wavelets and time-dependent spectral analysis. *Journal of Time Series Analysis*, **17**, 85–103.
- Robinson, P.M. (1983a). Review of various approaches to power spectrum estimation. In *Handbook of Statistics, Vol. 3* (D.R. Brillinger and P.R. Krishnaiah eds.), pp. 343–368. Amsterdam: North-Holland.
- Robinson, P.M. (1983b). Non-parametric estimation for time series models. *Journal of Time Series Analysis*, **4**, 185–208.
- Robinson, P.M. (1988). Semiparametric econometrics: A survey. *Journal of Applied Econometrics*, **3**, 35–51.
- Rosenblatt, M. (1970). Density estimation and Markov sequences. In *Nonparametric Techniques in Statistical Inference*, (M.L. Puri, ed.) 199–213. Cambridge University Press.
- Roussas, G.G. (1969). Nonparametric estimation in Markov process. *Annals of the Institute of Statistical Mathematics*, **21**, 73–87.
- Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian J. of Statist.*, **9**, 65–78.
- Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986). Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*. D.E. Rumelhart & J.L. McClelland eds. Cambridge: M.I.T. Press, 1, pp. 318–362.
- Said, S.E. & Dickey, D.A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, **71**, 599–607.
- Saikkonen, P. & Luukkonen, R. (1997). Testing cointegration in infinite order vector autoregressive processes. *Journal of Econometrics*, forthcoming.
- Schuster, A. (1898). On the investigation of hidden periodicities with application to a supposed 26-day period of meteorological phenomena. *Terr. Mag. Atmos. Elect.*, **3**, 13–41.
- Sheather, S.J. (1983). A data-based algorithm for choosing the window width when estimating the density at a point. *Computational Statistics and Data Analysis*, **1**, 229–238.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
- Singh, R.S. & Ullah, A. (1985). Nonparametric time series estimation of joint DGP, conditional DGP and vector autoregression. *Econometric Theory*, **1**, 27–52.
- Skaug, H.J. & Tjøstheim, D. (1993). Non-parametric tests of serial independence. In *The M. Priestley Birthday Volume* (ed. T. Subba Rao), pp. 207–229.
- Steigerwald, D.G. (1992). Adaptive estimation in time series regression models. *Journal of Econometrics*, **54**, 251–275.
- Stock, J.H. & M.W. Watson (1988). Testing for common trends. *Journal of the American Statistical Association*, **83**, 1097–1107.
- Stone, C.J. (1977). Consistent nonparametric regression. *Annals of Statistics* **5**, 595–635.
- Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, **89**, 208–218.
- Tibshirani, R. (1988). Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, **83**, 194–405.
- Tjøstheim, D. (1990). Nonlinear time series and Markov chains. *Advances in Applied Probability*, **22**, 587–611.
- Tjøstheim, D. (1994). Nonlinear time series, a selective review. *Scand. J. Statist.* **21**, 97–130.
- Tjøstheim, D. & Auestad, B. (1994a). Non-parametric identification of non-linear time series: projection. *Journal of the American Statistical Association*, **89**, 1398–1409.
- Tjøstheim, D. & Auestad, B. (1994b). Non-parametric identification of non-linear time series: selecting significant lags. *Journal of the American Statistical Association*, **89**, 1410–1419.
- Tong, H. (1983). *Threshold Models in Nonlinear Time Series Analysis. Lecture Notes in Statistics. Vol. 21*, Heidelberg: Springer.
- Tong, H. (1990). *Nonlinear Time Series Analysis: A Dynamic Approach*. Oxford: Oxford University Press.
- Truong, Y.K. (1993). A nonparametric framework for time series analysis. *New Directions in Time Series Analysis*. New York: Springer.
- Truong, Y.K. & Patil, P. (1996). On estimating possibly discontinuous regression involving stationary time series. *Manuscript*.
- Tsay, R. (1989). Testing and modeling threshold autoregressive processes. *Journal of the American Statistical Association*, **84**, 231–240.
- Tsybakov, A.B. (1986). Robust reconstruction of functions by the local approximation method. *Problems of Information Transmission*, **22**, 133–146.
- Tukey, J.W. (1949). The sampling theory of power spectrum estimates. *Proceedings of the Symposium on Applications of Autocorrelation Analysis to Physical Problems*, NAVEXOS-P-735, 47–67, Washington: Office of Naval Research.
- Tweedie, R. L. (1975). Sufficient Conditions for Ergodicity and Recurrence of Markov Chain on a General State Space. *Stochastic Processes and their Applications*, **3**, 385–403.
- Weigend, A.S. & Nix, D. (1994). Predictions with confidence intervals (local error bars). Discussion Paper No. 34, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.
- Werker, B.J.M. (1995). *Statistical Methods in Financial Econometrics*. CentER, Tilburg University.
- Wong, C.-M. & Kohn, R. (1996). A Bayesian approach to estimating and forecasting additive nonparametric autoregressive models. *Journal of Time Series Analysis*, **17**, 203–220.

Résumé

Beaucoup des éléments des séries temporelles sont analysables par des méthodes non-paramétriques. L'objet d'intérêt a une forme générale qui est approximée plus et plus précisément le nombre d'observations augmente. Cet article présente un survey des procédures non paramétriques en analyse des séries temporelles. Nous illustrons au moyen d'exemples portant sur l'estimation de densité, sur le bootstrap et l'estimation de tendance.

[Received August 1996, accepted November 1996]