A Survey on Nonparametric Time Series Analysis

by Siegfried Heiler

1 Introduction

In this survey we discuss the application of some nonparametric techniques to time series. There is indeed a long tradition in applying nonparametric methods in time series analysis, and this holds not only true for certain test situations, as, e.g. runs tests for randomness of a stochastic sequence, permutation tests or certain rank tests.
An old and established technique in time series analysis is periodogramme analysis. Although the periodogramme is an asymptotically unbiased estimate of the spectral density of an underlying stationary process, it is well known that it is not consistent. Therefore already in the early fifties smoothing the periodogramme directly with a so-called spectral window or using a system of weights, according to a lag window with which the empirical autocovariances are multiplied in the calculation of the Fourier transform, was introduced. Quite a number of different windows were proposed and with respect to the window width similar rules hold for achieving consistent estimates as the ones we will shortly discuss in the context of nonparametric regression later in this text. Nonparametric spectral estimation is extensively treated in many textbooks on time series analysis to which the interested reader is refered. Hence it will not be treated further in this survey.
Another area, where nonparametric ideas are being applied since a long time is smoothing and decomposing seasonal time series. Local polynomial regression can be traced back to 1931 (R.R. Macaulay). A. Fisher (1937) and H.L. Jones (1943) discussed a local least squares fit under the side condition that a locally constant periodic function (for modelling seasonal fluctuations) be annihilated and already in 1960 J. Bongard developped a unified principle for treating the interior and the boundary part (with and without seasonal variations) of a time series derived from a local regression approach. These ideas will be taken up later again in section 8, since they represent an attractive alternative to smoothing and seasonal decomposition procedures based on linear time series models.
The aim of this survey is to present some basic concepts of nonparametric regression including locally weighted regression with the special emphasis on their application to time series. Nonparametric regression has become an area with an abundance in new methodological proposals and developments in recent years. It is not the intention of this paper to give a comprehensive overview on the subject. We rather want to concentrate on the basic ideas only. The reader interested in some different aspects may be refered to a survey paper by Härdle, Lütkepohl and Chen (1997), where more specific areas, proposals and further references can be found.
The ARMA model is a typical linear time series model. Threshold autoregression (TAR) models and its variates are specific types of nonlinear models. ARCH and GARCH type models are also of a very specific nonlinear type to capture volatility phenomena. In contrast to that in nonparametric regression no assumption is made about the form of the regression function. Only some smoothness conditions are required. The complexity of the model will be determined completely by the data. One lets the data speak for themselves.

Thereby one avoids subjectivity in selecting a specific parametric model. But the gain in flexibility has a price. One has to choose bandwidths. We come back to this later. Besides this, a higher complexity in the mathematical argumentation is involved. However, asymptotic considerations will not be discussed in detail in this survey.

Because of their flexibility nonparametric regression techniques may serve as a first step in the process of finding an adequate parametric model. If no such one can be found which describes the underlying structure adequately, then the results of nonparametric estimation may be used directly for forecasting or for describing the characteristics of the time series.

2 Nonparametric regression

Since forecasting is an important objective of many time series analyses, estimating the conditional distribution, or some of its characteristics play a considerable role. For point prediction the conditional mean or median is of particular interest. In order to obtain confidence or prediction intervals also estimates of conditional variances or conditional quantiles are needed. The latter ones are also of interest in studying volatility in financial time series.
The first step to go is therefore to look at nonparametric estimation of densities and conditional densities. Let $x \in \mathbb{R}$ be a random variable whose distribution has a density f and let $x_1, \ldots, x_n$ be a random sample from x. Then a *kernel density estimator* for f is given by

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right). \tag{2.1}$$

Here K is a so-called *kernel* function, i.e. a symmetric density assigning weights to the observations x_i which decrease with the distance between x and x_i. Some popular kernel functions are listed in Table 2.1 and exhibited in Figure 2.1. The first 5 have the interval $[-1, 1]$ as support, whereas the Gaussian kernel has infinite support. h_n is the *bandwidth* which drives the size of the local neighbourhood being included in the estimation of f at x. The bandwidth depends on the sample size n and has to fulfil $h_n \to 0$ and $nh_n \to \infty$ for $n \to \infty$ as necessary condition for consistency. But for practical applications this asymptotic condition is not very helpful. A very small bandwidth will lead to a wiggly course of the estimated density, whereas a large bandwidth yields a smooth course but will possibly flatten out interesting details. Bandwidth selection will

be dealt with in section 7.

A $k_n - nearest\ neighbour\ (k_n - NN)\ estimator$ of f is obtained by substituting the

Table 2.1: Selected kernel functions

Name	Kernel
Uniform	$\frac{1}{2}\ \mathbb{I}_{[-1,1]}(u)$
Triangle	$(1-\|u\|)\mathbb{I}_{[-1,1]}(u)$
Epanechnikov	$\frac{3}{4}\ (1-u^2)\mathbb{I}_{[-1,1]}(u)$
Bisquare	$\frac{15}{16}\ (1-2u^2+u^4)\mathbb{I}_{[-1,1]}(u)$
Triweight	$\frac{35}{32}\ (1-3u^2+3u^4-u^6)\mathbb{I}_{[-1,1]}(u)$
Gaussian	$\frac{1}{\sqrt{2\pi}}\ \exp(-\frac{1}{2}u^2)$

fixed bandwidth h_n in (2.1) by the random variable $H_{n,k_n}(x)$ measuring the distance between x and the k_n-nearest observation among the $x_i, i = 1, \ldots, n$.

Nearest neighbour estimators have the property that the number of observations used for the local approach is fixed. This is an advantage if the x-space shows a greatly unbalanced design. On the other hand the bias varies from point to point due to the variable local bandwidth.

For $x \in \mathbb{R}^p$ a kernel $K : \mathbb{R}^p \to \mathbb{R}$ is needed in (2.1). In this case either $product\ kernels$

$$K(u) = \prod_{j=1}^{d} K_j(u_j)$$

with kernels K_j and $K_j : \mathbb{R} \to \mathbb{R}$, bandwidth h_j in coordinate j, and $h_n = h_1 \cdot \ldots \cdot h_p$ or $norm\ kernels$

$$K(u) = K\left(||u||\right)$$

with a suitable norm on $\mathbb{R}^p$ are used. In connection with time series applications frequently product kernels are applied,

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{p}\frac{1}{h_j}K_j\left(\frac{x_{ij}-x_j}{h_j}\right) \tag{2.2}$$

and $h_j = \hat{\sigma}_j \cdot h$ with an estimated standard deviation in the j-th coordinate is a popular choice for the bandwidths.

Let now (y, x) with $y \in \mathbb{R}, x \in \mathbb{R}^p$ be a random vector with joint density $f(y, x)$ and let $f_X(x)$ be the marginal density of x. Then the conditional density $g(y|x) = f(y, x)/f_X(x)$ can be estimated by inserting a kernel density estimator or a corresponding

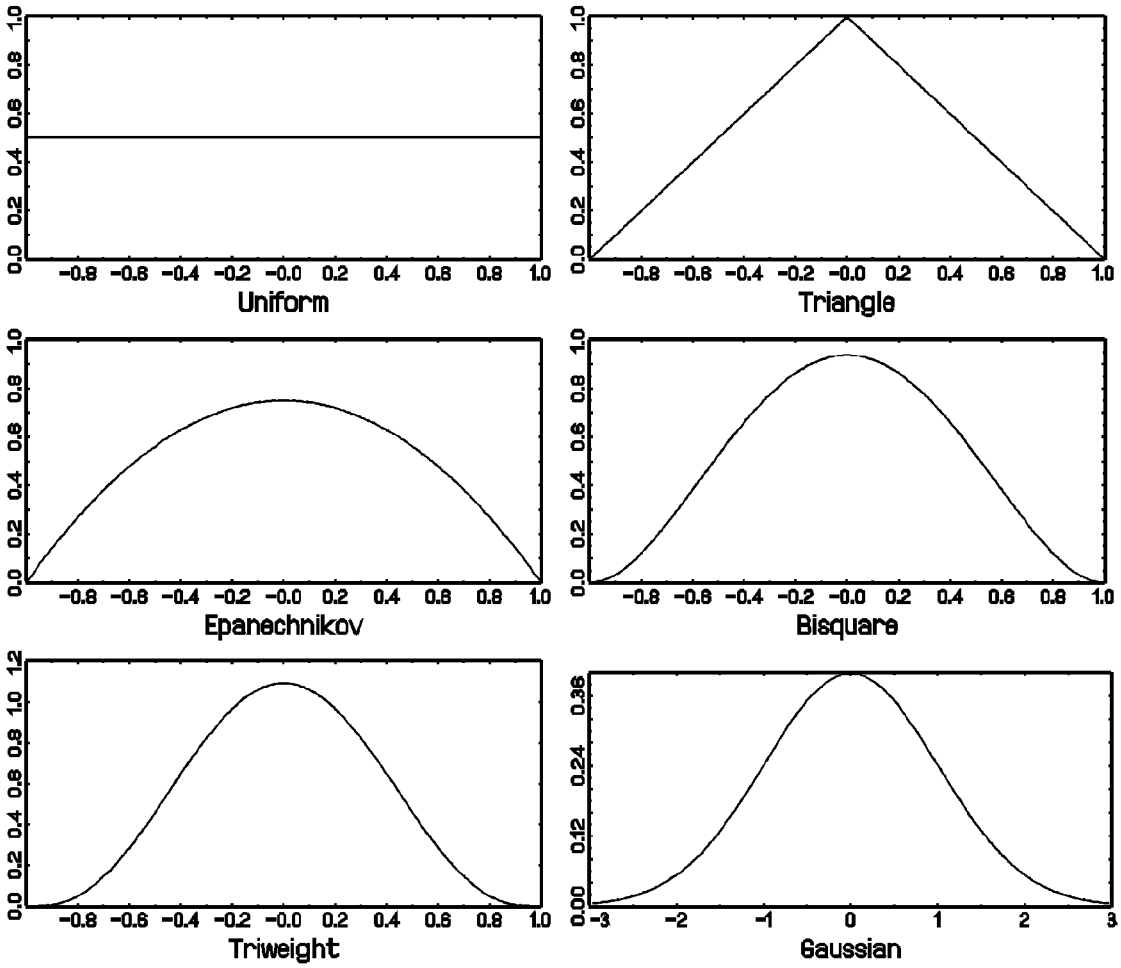

Figure 2.1. Some popular kernel functions in practice

nearest neighbourhood estimator in the nominator and denominator of $g(y|x)$. With the choice of a kernel function

$$K = \mathbb{R}^{p+1} \to \mathbb{R},\ K(y,x) = K_1(y)K(x)$$

and bandwidths h_1 resp. h we obtain the kernel estimator for the conditional density

$$g_n(y|x) = \frac{h_1^{-1} \sum_{i=1}^{n} K_1\left(\frac{y_i - y}{h_1}\right) K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}. \tag{2.3}$$

An estimator for the *conditional mean* $m(x) = \int_{-\infty}^{\infty} y g(y|x) dy$ is obtained when we replace

g in the integral by its estmator g_n. For K_1 being a symmetric density this immediately yields

$$m_n(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}. \tag{2.4}$$

This is the well-known *Nadaraya-Watson nonparametric regression estimator* (NW-estimator, Nadaraya, 1964; Watson, 1964). We see that it can be written as a weighted mean

$$m_n(x) = \sum_{i=1}^{n} y_i w_{n,i}(x; x_1, \ldots, x_n), \tag{2.5}$$

where the random weights depend on the point x and the random variables $x_1, \ldots, x_n$. Apart from conditional means also conditional quantiles are of interest in various time series applications. Let

$$F(y|x) = \int_{-\infty}^{y} g(y|x)dy \tag{2.6}$$

denote the conditional distribution function of y given x. Then the conditional α-quantile at x, $q_\alpha(x)$ is defined as

$$q_\alpha(x) = inf\{y \in \mathbb{R} | F(y|x) \geq \alpha\},\ 0 < \alpha < 1. \tag{2.7}$$

If $g(\cdot|x)$ is strictly positive, then of course $q_\alpha(x)$ is the unique solution of $F(y|x) = \alpha$, i.e. $q_\alpha(x) = F^{-1}(\alpha|x)$. One possible procedure for estimating q_α is to take the empirical α-quantile of an estimator $F_n = (\cdot|x)$ according to (2.7).
Let $F_1(z) = \int_{-\infty}^{z} K_1(u)du$ be the distribution function pertaining to the kernel K_1. Then the estimated conditional distribution, obtained by integrating $g_n(\cdot|x)$ from $-\infty$ to y, is given by

$$F_n(y|x) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i-x}{h}\right) F_1\left(\frac{y-y_i}{h_1}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i-x}{h}\right)}. \tag{2.8}$$

Let us assume that K_1 has support $[-1,1]$. Then we have

$$F_1\left(\frac{y-y_i}{h_1}\right) = \begin{cases} 1 & , \quad for \quad y_i \le y - h_1 \\ 0 & , \quad for \quad y_i \ge y + h_1 \end{cases},$$

so that in this case

$$\begin{aligned} F_n(y|x) &= \frac{1}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)} \left\{ \sum_{i=1}^{n} \mathbf{1}_{(-\infty, y-h_1]}(y_i) K\left(\frac{x_i - x}{h}\right) \right. \\ &\quad \left. + \sum_{i=1}^{n} \mathbf{1}_{(y-h_1, y+h_1)}(y_i) F_1\left(\frac{y-y_i}{h_1}\right) K\left(\frac{x_i - x}{h}\right) \right\}. \end{aligned} \tag{2.9}$$

One can see that the estimation contains only observations in the regressor space laying in a band around x. The first sum on the right hand side includes observations, whose y-values are less than or equal to $y - h_1$. The second sum contains observations with y_i-values in a neighbourhood of y. In contrast to a usual empirical distribution function here also observations greater than y obtain a positive weight.
Of particular interest may be the median regression function $q_{1/2}$ for asymmetric distributions as an alternative to ordinary regression based on the mean. Another interesting application may be the estimation of $q_{\alpha/2}$ and $q_{1-\alpha/2}$ in order to get predictive intervals. These can be compared with intervals obtained from parametric models, which lack the possibility to evaluate the bias due to mis-specification of the model.
Taking some boundary corrections into account, for a not too unbalanced design the second sum in (2.9) can be approximated by $\sum_{i=1}^{n} \mathbf{1}_{(y-h_1, y]} K\left(\frac{x_i - x}{h}\right)$, so that the conditional distribution function is estimated by

$$\tilde{F}_n(y|x) = \frac{\sum_{i=1}^{n} \mathbf{1}_{(-\infty, y]}(y_i) K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}. \tag{2.10}$$

This estimator was for $x \in \mathbb{R}$ considered by Horvath and Yandell (1988) who proved asymptotic results for the i.i.d. case. Abberger (1996) derives from (2.10) the empirical quantile function

$$q_{n,\alpha}(x) = \inf\{y \in \mathbb{R} | \tilde{F}_n(y|x) \ge \alpha\}, \quad 0 < \alpha < 1 \tag{2.11}$$

and investigates the behaviour of $\tilde{F}_n$ and $q_{n,\alpha}$ in applications to stationary time series.

3 Kernel estimation in time series

When a kernel- or NN-estimator is applied to dependent data, as it is the case in time series, then it is effected only by the dependence among the observations in a small window and not by that between all data. This fact reduces the dependence between the estimates, so that many of the techniques developed for independent data can be applied in these cases as well. This fact was called *the whitening by windowing principle* by Hart (1996).
A typical situation for an application to a time series $\{z_t\}$ is that the regressor vector x consists of past time series values

$$x_t = (z_{t-1}, \ldots, z_{t-p}), \tag{3.1}$$

which leads to the very general nonparametric autoregression model

$$z_t = m(z_{t-1}, \ldots, z_{t-p}) + a_t, \quad t = p+1, p+2, \ldots \tag{3.2}$$

with $\{a_t\}$ a white noise sequence. Of course x_t might also include time series values of other predictive variables like leading indicators.
An indispensable requirement for proving asymptotic properties of kernel estimates in this and related situations is that the underlying processes are stationary. Another condition is that the memory of these underlying processes decreases with distance between events and that the rate of decay can be estimated from above by so-called *mixing conditions*. So-called strong mixing conditions are used by Robinson (1983, 1986). Collomb (1984, 1985) worked with so-called ϕ-or uniform mixing conditions.
We will not present these fairly complicated asymptotic considerations here. But we would like to remark that these mixing conditions are hard to check in practice.

In contrast to linear autoregressive models of the form $z_t = \phi_1 z_{t-1} + \ldots + \phi_p z_{t-p} + a_t$, and in a certain sense also to threshold autoregression where the autoregressive parameters vary according to some threshold variable the model (3.2) is more general and flexible and its estimation may lead to insights which can be helpful in choosing an appropriate parametric (possibly nonlinear) model afterwards.

For $x \in \mathbb{R}^p$, x_t as in (3.1) and weights

$$w_{n,t} = K\left(\frac{x_t - x}{h}\right) / \sum_{s=p+1}^{n} K\left(\frac{x_s - x}{h}\right)$$

the Nadaraya-Watson estimator in model (3.2) is given by

$$m_n(x) = \sum_{s=p+1}^{n} z_t w_{n,t}(x). \tag{3.3}$$

For x equal to the last observed pattern, $x = (z_n, z_{n-1}, \ldots, z_{n-p+1})'$ this provides a one-step ahead predictor for z_{n+1} which allows a very intuitive interpretation. Given the course of the time series observed over the last p instants, the predictor is a weighted mean of all those time series values in the past, which followed a course pattern that is similar to the last observed one. The weights depend on how close the pattern observed in the past comes to the pattern given by $(x_n, \ldots, x_{n-p+1})'$.
A k-step ahead predictor is given if z_t in (3.3) is replaced by z_{t-k+1}.

$$m_{n,k} = \sum_{t=p+1}^{n-k+1} z_{t+k-1} w_{n,t}(x) \quad , k = 1, 2, \ldots \ . \tag{3.4}$$

This predictor does not use the variables $z_{n+1}, \ldots, z_{n+k}$, which are unknown, but may contain information about the conditional expection $E(z_{n+k}|(z_n, \ldots, z_{n-p+1})')$. They might be replaced by estimates in a multistep procedure which consists in a succession of one-step ahead forecasts. This procedure can lead to a smaller mean squared error than the multistep procedure (3.4). For a different proposal see Chen (1996).

Up to now we have only considered the autoregressive case where the regressor vector contains past time series values. The case of vector autoregression, where for each individual (scalar) time series also past values of related time series or leading indicators are included in the regression vector, can be treated in a similar way as nonparametric autoregression, although the number of components in x is restricted due to the *"curse of dimensionality"*, to which we come back later.
If the regressor vector $x_t = (z_{t-1}, \ldots, z_{t-p})'$ is used in estimating conditional distribution functions and conditional quantiles, as e.g. in (2.10) and (2.11), then we arrive at *quantile autoregression*. The median autoregression $q_{n,1/2}$ may serve as an alternative to the mean autoregression (3.3). In financial data one is often interested in the behaviour of quantiles in the tails. For instance the *value at risk* of a certain asset is measured by looking at low quantiles ($\alpha = 0.01$ or $\alpha = 0.05$) of the conditional distribution of the corresponding series of returns.

Abberger (1996) applied quantile autoregression to time series of daily stock returns. In order to assess such models forecast error cannot serve as a criterion, since quantiles are not observable. Abberger proposed the criterion

$$\xi_\alpha = 1 - \sum_{t=1}^{n} \rho_\alpha \left(z_t - q_\alpha(x_t)\right) / \sum_{t=1}^{n} \rho_\alpha \left(z_t - q_\alpha\right), \tag{3.5}$$

where

$$\rho_\alpha(u) = \alpha \mathbf{1}_{[0,\infty)}(u)u + (\alpha - 1)\mathbf{1}_{(-\infty,0)}(u)u \tag{3.6}$$

is the loss function introduced by Koenker and Basset (1978) in their seminal paper on quantile regression and q_α is the unconditional α-quantile of the corresponding distribution.
ξ_α is constructed according to the R^2-criterion in ordinary regression. It assumes values between zero and one, where $\xi_\alpha = 0$ if $q_\alpha(x_t) = q_\alpha$ for all x_t and $\xi_\alpha = 1$ if $z_t = q_\alpha(x_t)$ for all t and all α , i.e. if the distribution of $\{z|x\}$ is a one-point distribution.
The following Figure 3.1 and Table 3.1 illustrate the behaviour of ξ_α with a simulated conical data set of 500 observations.
The observations are heteroscedastic and have mean zero. The correlation between x and y is -0.002. In Table 3.1 empirical ξ_α-values for different α are exhibited. They are calculated by replacing in (3.5) $q_\alpha(x_t)$ by its kernel estimator $q_{n,\alpha}(x_t)$ and q_α by the empirical unconditional quantile of the first $t-1$ data values $z_1, \ldots, z_{t-1}$. The latter can be interpreted as a naive forecast of $q_\alpha(x_t)$.
The findings of Abberger (1996, 1997) for several German stock returns were ξ_α -values close to zero for the median and increasing in a U-shaped form towards the boundary areas around $\alpha = 0.01$ respectively $\alpha = 0.99$.

ARCH- and GARCH models represent a very specific kind of parametric modeling for studying the phenomenon of volatility. A flexible alternative to the combination of an ARMA-

Table 3.1. ξ_α-values for the data in Figure 3.1

α	0.01	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.99
ξ_α	0.43	0.36	0.27	0.10	0.01	0.11	0.26	0.34	0.41

model with ARCH- or GARCH-residuals is given by the **c***onditional* **h***eteroscedastic* **a***uto***r***egressive*
) *model*

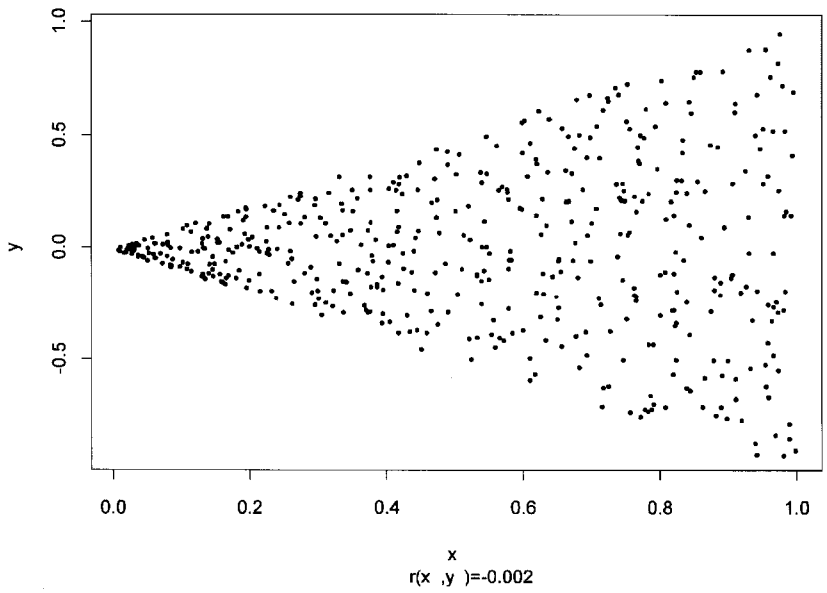

Figure 3.1. Simulated heteroskedastic data, n=500

$$z_t = m(x_t) + \sigma(x_t)\xi_t, \tag{3.7}$$

studied by Härdle and Yang (1996) or Härdle, Tsybakov and Yang (1997). Here $x_t = (z_{t-1}, \ldots, z_{t-p})'$ is again the autoregressive vector (3.1), ξ_t is a random variable with mean zero and variance one. $\sigma^2(x)$ is called the *volatility function*. Given an estimator for m , e. g. the NW-estimator m_n according to (3.3), it was suggested that $\sigma^2(x)$ can be estimated by

$$\sigma_n^2(x_t) = g_n(x_t) - m_n^2(x_t), \tag{3.8}$$

where

$$g_n(x) = \frac{\sum_{t=1}^{n} K\left(\frac{x_t - x}{h}\right) z_t^2}{\sum_{t=1}^{n} K\left(\frac{x_t - x}{h}\right)} = \sum_{t=1}^{n} z_t^2 w_{n,t}(x). \tag{3.9}$$

Since the estimator (3.8) is based on a difference, it can happen that from time to time a negative variance estimator results. This can be avoided if the volatility function is estimated on the basis of residuals. See (7.10), the discussion there and Feng and Heiler (1998a).

In the context of time series analysis not only past values of the time series itself or of related series may occur as regressor variables, but also the time index itself, in which case $x_t = t$, or some functions of the time index like polynomials or trigonometric functions. This leads to smoothing approaches. In the case $m(x_t) = m(t)$ the NW estimator at t consists in a weighted mean of the time series values in a neighbourhood $[t-h, t+h]$ of z_t with nonrandom weights. Polynomials and trigonometric functions in t are used in decomposing a seasonal time series into trend- cyclical and seasonal components according to an unobserved components model. This application will be studied in section 8 after the discussion of locally weighted regression.
In the area of quantile estimation the regressor $x_t = t$ leads to quantile smoothing. This technique was used by Abberger (1996, 1997) in order to compare the results of a nonparametric procedure for stock returns with those of a GARCH-model, evaluated with an *S–Plus* package under the standard assumption of an underlying Gaussian distribution. As an example we take daily discrete DAX returns, defined as $z_t = (price_t - price_{t-1})/price_{t-1}$, exhibited in Figure 3.2.

Since the Gaussian distribution is completely determined by mean and variance, conditional quantiles can easily be calculated from the outcomes of the GARCH model estimation. The results are depicted in Figure 3.3 and 3.4 for the lower and upper quartiles and for the 0.1 and 0.9 quantiles, respectively. Two messages can be learned from the results. The first is that the asymmetric behaviour of volatility, which is revealed by the nonparametric approach, will remain completely hidden by the choice of a wrong parametric model which is being offered as the default option by the package. In the presented example, which is not untypical for stock returns, volatility is a phenomenon which has mainly to do with movements in the lower tails of the conditional distributions. The second finding in the figures is that kernel smoothing is very robust towards aberrant and erratic observations in the course of the time series, whereas GARCH models react very sensitively to them.

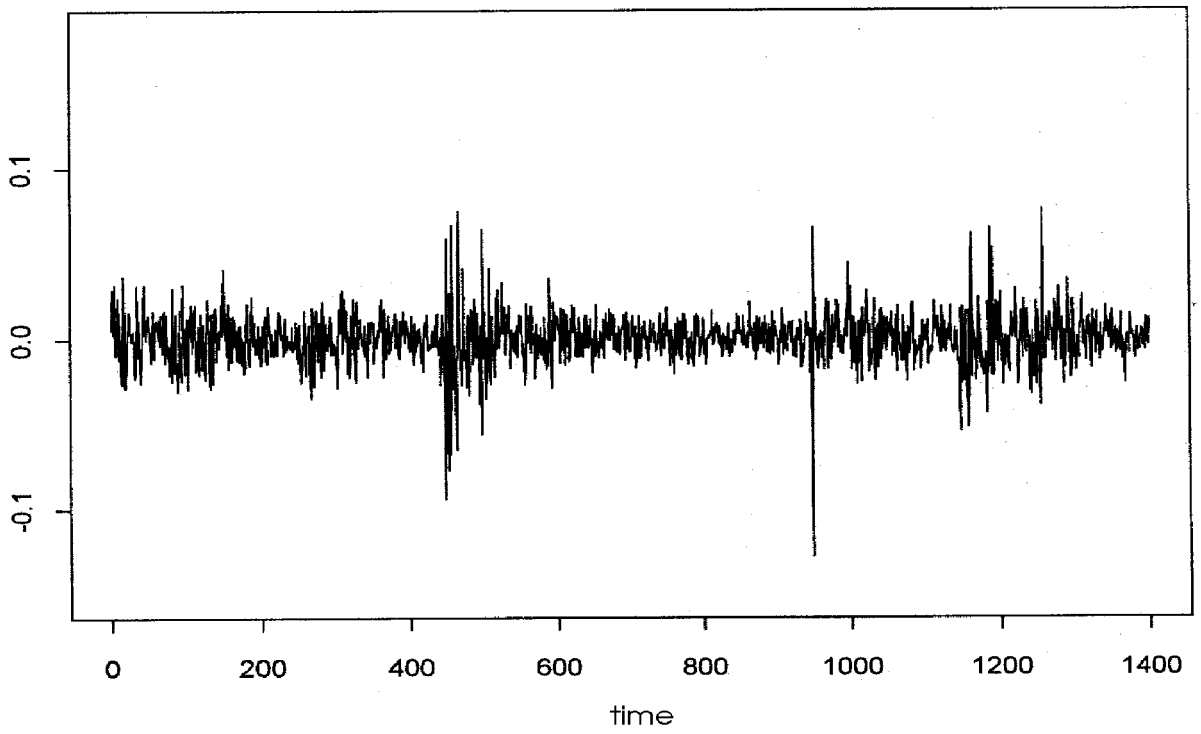

Figure 3.2 Time series of daily DAX returns from Jan. 2, 1986 to Aug. 13, 1991

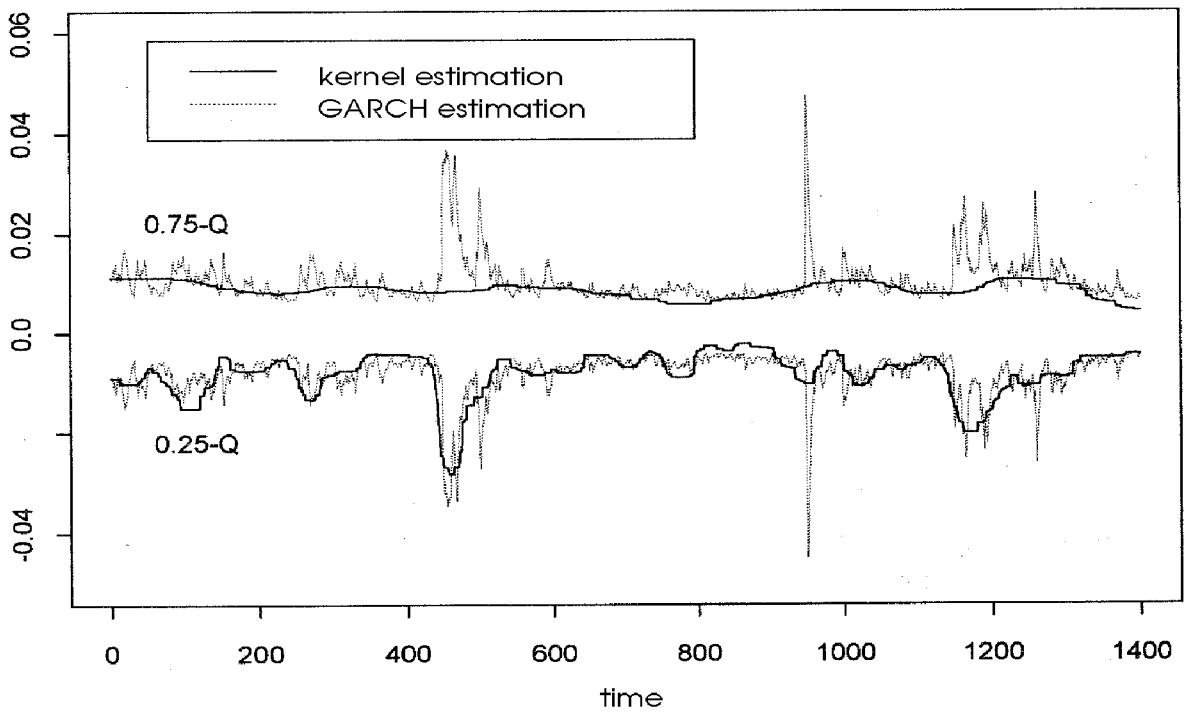

Figure 3.3 Estimation of 0.25- and 0.75-quantiles of daily DAX returns

4 Problems of simple kernel estimation and restricted approaches

The nonparametric approaches we have treated so far suffer from two drawbacks. One is the so-called *"curse of dimensionality"*, the other is increased bias in cases of a highly-clustered design density and particularly at the boundaries of the x-space. Curse of dimensionality describes the fact that in higher dimensional regression problems the subspace of $\mathbb{R}^{p+1}$ spanned by the data is rather empty, i.e., there are only few observations in the neighbourhood of a point $x \in \mathbb{R}^p$. In practice this happens to be the case already for $p > 2$.

Several proposals have been made to cope with the curse of dimensionality problem. We will describe only two of them very shortly. The first consists in decomposing $\mathbb{R}^p$ into a class of J disjoint course patterns, $A_j, j = 1, \ldots, J$ with the aid of a non-hierarchical cluster analysis. These J disjoint sets serve then as the states of a homogeneous Markov chain. In the model

$$m(x_t) = E[z_t | x_t \in A_j] \text{ for } x_t \in A_j, j = 1, \ldots, J$$

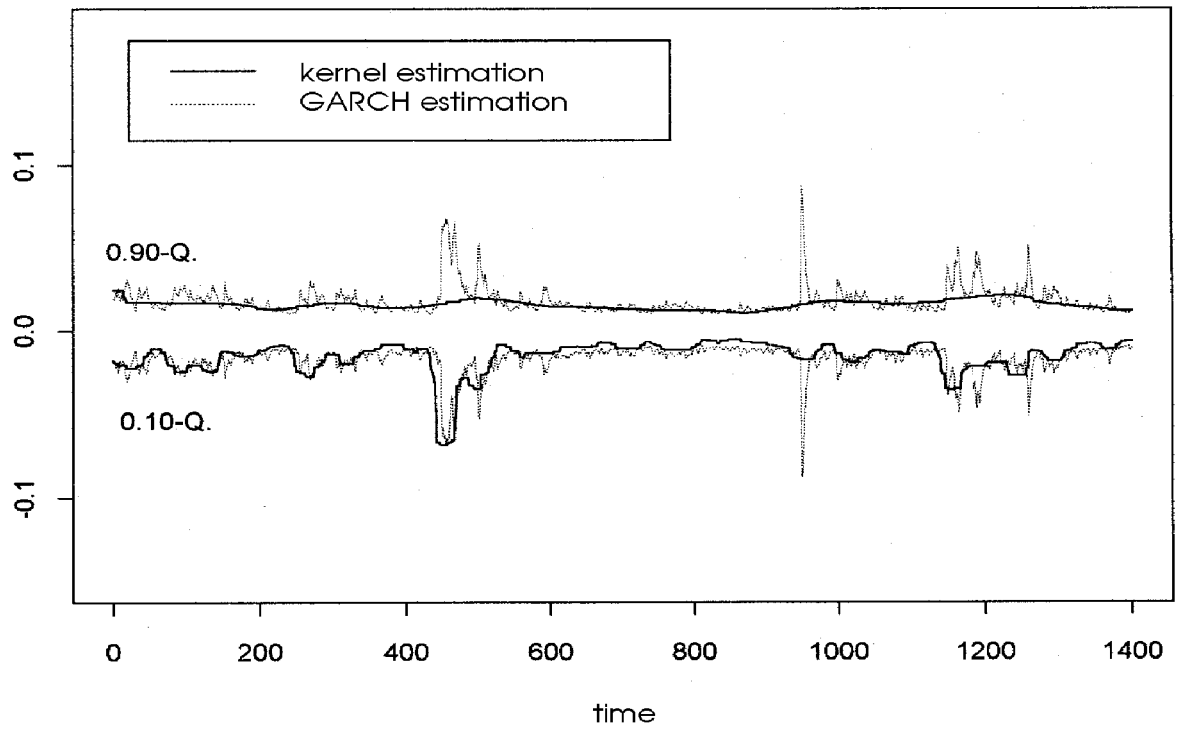

Figure 3.4 Estimation of 0.10- and 0.90-quantiles of daily DAX returns

with x_t being the autoregressive vector (3.1) m is estimated by

$$m_n(x_t) = N_j^{-1} \sum_{s=1}^{n} z_s \mathbf{1}_{A_j}(x_s),$$

where N_j is the number of course patterns of length p from the time series in A_j. Here the estimator is an unweighted mean of all values following courses in pattern class A_j. Markov chain models of this type were first used by S. Yakowitz (1979b) for analysing time series of water runoff in rivers. Asymptotic properties for this type of model are discussed by Collomb (1980, 1983).
Gouriroux and Montfort (1992) examined a corresponding model for economic time series by incorporating volatility. They called their model

$$z_t = \sum_{j=1}^{J} \alpha_j \mathbf{1}_{A_j}(x_t) + \sum_{j=1}^{J} \beta_j \mathbf{1}_{A_j}(x_t)\xi_t$$

a qualitative threshold ARCH model.
Another proposal in order to cope with the curse of dimensionality is given by the so-called generalized additive models, studied by Hastie and Tibshirani (1990), which are defined

as

$$z_t = m_0 + \sum_{j=1}^{p} m_j(z_{t-i_j}) + a_t.$$

The components m_j are again of a general form. For estimation so-called backfitting algorithms such as the alternating conditional expectation algorithm (ACE) of Breiman and Friedman (1985) or the BRUTO algorithm of Hastie and Tibshirani (1990) may be used. The main idea of backfitting goes as follows. In the above model $E[z_t - m_0 - \sum_{j \neq k} m_j(z_{t-i_j})] = m_k(z_{t-i_k})$. Hence the variable in square brackets can be used to obtain a nonparametric estimate for $m_k(z_{t-i_k})$. But of course, the other m_j are unknown as well, so that the estimation procedure has to be iterated until all the $m_{n,j}$ converge. For a more detailed study of generalized additive models the reader is refered to the book of Hastie and Tibshirani as well as to the two interesting papers by Chen and Tsay in JASA (1993). For further discussion and other approaches see also Härdle, Lütkepohl and Chen (1997).
Quite a few proposals can be found in the literature dealing with the bias problem of NW-estimators close to the boundary and in cases of an unbalanced design in the x-space.
Gasser and Müller (1979, 1984) suggested for the case $p = 1$ a system of variable weights, Gasser, Müller and Mammitzsch (1985) developed asymmetric boundary kernels and Messer and Goldstein (1993) suggested variable kernels which automatically get deformed and thus reduce the bias in the boundary area.
Yang (1981) and Stute (1984) suggested a symmetrized $k - NN$ estimator and Michels (1992) proposed boundary kernels for bias reduction which can be carried over to the case $p > 1$. We do not discuss the above mentioned proposals in more detail since the mentioned disadvantages can be repaired by using locally weighted regression.

5 Locally weighted regression

Locally weighted respectively local polynomial regression was introduced into the statistical literature by Stone (1977) and Cleveland (1979). The statistical properties were investigated since then in papers by Tsybakov (1986), Fan (1993), Fan and Gijbels (1992, 1995), Ruppert and Wand (1994) and many others. A detailed description may be found in the book of Fan and Gijbels (1996).
For the sake of simplicity we start with the assumption that the regressor x is a scalar. For a better understanding we regard the data as being generated by a location-scale model

$$y = m(x) + \sigma(x)\xi \tag{5.1}$$

akin to the one considered in (3.7), where the ξ are independent with $E(\xi) = 0, Var(\xi) = 1$ and $m(x_0) = E(y|x = x_0)$. m is assumed to be smooth in the sense that the $(p+1)$th derivative exists at x_0, so that it can be expanded in a Taylor series around x_0 .

$$m(x) = m(x_0) + (x - x_0)m'(x_0) + \ldots + (x - x_0)^r \frac{m^{(r)}(x_0)}{r!} + R_r(x) \tag{5.2}$$

with the remainder term

$$R_r(x) = (x - x_0)^{r+1} m^{(r+1)}\Big(x_0 + \theta(x - x_0)\Big)/(r-1)!,\ 0 < \theta < 1. \tag{5.3}$$

With

$$\beta_j(x_0) = m^{(j)}(x_0)/j!, \quad j = 0, 1, \ldots, r \tag{5.4}$$

we arrive at a local polynomial representation for m,

$$m(x) \approx \sum_{j=0}^{r} \beta_j(x_0)(x - x_0)^j. \tag{5.5}$$

This approach motivates the nonparametric estimation of m as a local polynomial by solving the least squares problem

$$min_{\beta \in \mathbb{R}^{r+1}} \left\{ \sum_{i=1}^{n} \left[y_i - \sum_{j=0}^{r} (x_i - x)^j \beta_j \right]^2 K\left(\frac{x_i - x}{h}\right) \right\}.$$

With the design matrix X_x having the n rows $[1, x_i - x, \ldots, (x_i - x)^r]$, the diagonal weight matrix $W_x = diag\left(K(\frac{x_i - x}{h})\right)$ and the vector $y = (y_1, \ldots, y_n)'$ the solutions at x is given by

$$\hat{\beta}(x) = (X_x' W_x X_x)^{-1} X_x' W_x y, \tag{5.6}$$

and with e_j being the j-th unit vector in $\mathbb{R}^{r+1}$ we see immediately that

$$\hat{m}(x) = \hat{\beta}_0 = e_1'(X_x'W_xX_x)^{-1}X_x'W_xy, \tag{5.7}$$

and that with

$$\hat{m}^{(j)}(x) = \hat{\beta}_j(x)j! = j!e'_{j+1}(X_xW_xX_x)^{-1}X_x'W_xy, \qquad j = 1,\ldots,r \tag{5.8}$$

an estimator for the j-th derivative of m is given.
The case $r = 0$ yields the Nadaraya-Watson estimator (3.3).
Let $u = (r_r(x_1))_{i=1}^n$ be the residual vector containing the remainder terms according to (5.3) at the data points. Then the conditional bias of $\hat{\beta}(x)$ is given by

$$B\Big(\hat{\beta}(x)\Big) = (X_x'W_xX_x)^{-1}X_x'W_xu,$$

and with $\Sigma_x = W(x)^2 diag\Big(\sigma^2(x_i)\Big)$ its conditional covariance matrix is

$$Var\Big(\hat{\beta}(x)\Big) = (X_x'W_xX_x)^{-1}(X_x'\Sigma_xX_x)(X_x'W_xX_x)^{-1}.$$

The above two expressions cannot be used directly since they contain the unknown vector u of remainder terms and the unknown diagonal matrix Σ_x.
A first order asymptotic expansion of the variance and the bias term uses the moments of K and K^2, denoted by

$$\mu_j = \int u^j K(u)du \qquad \text{and } \nu_j = \int u^j K^2(u)du,$$

which are contained in the matrices

$$S = (\mu_{j+l})_{0\le j,l\le r}, \qquad \tilde{S} = (\mu_{j+l+1})_{0\le j,l\le r}, \qquad S^* = (\nu_{j-l})_{0\le j,l\le r}$$

and the vectors $c_r = (\mu_{r+1},\ldots\mu_{2r+1})$, $\tilde{c_r} = (\mu_{r+2},\ldots,\mu_{2r+2})$. For an i. i. d. sample $(y_1,x_1),\ldots,(y_n,x_n)$ with the marginal density $f(x) > 0$ and with $f, m^{(r+1)}$ and

σ^2 continous in a neighbourhood of x we obtain for $h \longrightarrow 0$ and $nh_n \longrightarrow \infty$ the asymptotic conditional variance

$$Var\Big(\hat{m}^{(j)}(x)\Big) = e'_{j+1} S^{-1} S^* S^{-1} e_{j+1} \frac{(j!)^2 \sigma^2(x)}{f(x) n h^{1+2j}} + o_p\Big(\frac{1}{nh^{1+2j}}\Big). \tag{5.9}$$

For the asymptotic conditional bias we have to distinguish between the cases where $r-j$ is odd and where $r-j$ is even.
For $r-j$ odd we have

$$Bias\Big(\hat{m}^{(j)}(x)\Big) = e'_{j+1} S^{-1} c_r \frac{j!}{(r+1)!} m^{(r+1)}(x) h^{r+1-j} + o_p(h^{r+1-j}). \tag{5.10}$$

For $(r-j)$ even the asymptotic bias is

$$\begin{aligned} Bias\Big(\hat{m}^{(j)}(x)\Big) &= e'_{j+1} S^{-1} \tilde{c}_r \frac{j!}{(r+2)!} \cdot \\ &\Big\{m^{(r+2)}(x) + (r+2) m^{(r+1)}(x) \frac{f'(x)}{f(x)}\Big\} h^{r+2-j} + o_p(h^{r+2-j}), \end{aligned} \tag{5.11}$$

provided that f' and $m^{(r+2)}$ are continuous in a neighbourhood of x and $nh^3 \longrightarrow \infty$. As a very interesting fact we notice the difference in asymptotic bias between $r-j$ odd and $r-j$ even. For instance we have for the NW-estimator $(r=0, j=0)$,

$$B(m_n(x)) = h^2 [m''(x)/2 + m' f'(x)/f(x)] \mu_2 + o_p(h^2),$$

whereas for the local linear approach we obtain

$$B(\hat{m}(x)) = h^2 m''(x) \mu_2 / 2 + o_p(h^2).$$

We see that the bias of the local linear estimator has a simpler structure. The linear term in the bias expansion vanishes, whereas the expression for the variance is the same in both cases and given by $\nu_0 \sigma^2(x)/nh$. The bias of the NW-estimator does not only depend on m' , but also on the score function $-f'/f$. This is the reason why an unbalanced design leads to an increased bias.
Similar considerations hold for higher order polynomials. In practice this means that for

estimating m it is sufficient to consider $r = 1$ or $r = 3$, and for m' only $r = 2$ or $r = 4$ should be considered. In many applications $r = j + 1$ suffices. Fitting a higher order polynomial will possibly reduce the bias, but on the other hand the variance will increase since more parameters have to be estimated locally.
If the regressor x is a vector rather than a scalar in most cases a local linear approach is chosen since in this case the step from $r = 1$ to $r = 3$ leads to strong increase of parameters to be estimated locally which entails an inacceptable increase in variance.
Since

$$\hat{\beta}_j(x) = e'_{j+1}\hat{\beta} = e'_{j+1}(X'_x W_x X_x)^{-1} X'_x W_x y = \sum_{i=1}^{n} w_{ni}^{j}\Big(\frac{x_i - x}{h}\Big) y_i \tag{5.12}$$

for estimating $\beta_j(x) = m^{(j)}(x)/j!$ we have a similar expression as a weighted mean like for the NW-estimator (3.3). The weights depend on the observations x_i and on the location of x in the design space.
It can be seen easily that the weights $w_{ni}^{j}(u_t) = w_{ni}^{j}\left(\frac{x_i - x_o}{nh}\right)$ satisfy the discrete moment conditions

$$\sum_{i=1}^{n}\Big(x_i - x\Big)^q w_{ni}^{j}\Big(\frac{x_i - x}{h}\Big) = \delta_{jq} \qquad \text{with } 0 \le j, q \le r.$$

As a consequence of this the sample bias for estimating a polynomial with degree less than or equal to r is zero.
The variance of $\hat{m}^{(j)}(x)$ is given by

$$Var\Big(\hat{m}^{(j)}(x)\Big) = \sum_{i=1}^{n} w_{ni}^{j}\Big(\frac{x_i - x}{h}\Big)^2 \sigma^2(x_i).$$

The kernel with the weights $w_{ni}^{j}(u_t)$ is called the *active kernel*.

A first order approximation to the w_{ni}^{j} is given if $(X'_x W_x X_x)$ is replaced by the kernel moments matrix S .
The according kernel

$$\tilde{K}^{(j)}(u) = e'_{j+1} S^{-1}(1, u, \ldots, u^r)' K(u) \tag{5.13}$$

is called the *equivalent kernel*. It satisfies the corresponding moment conditions

$$\int u^q \tilde{K}(j)(u)du = \delta_{jq} \qquad 0 \leq j,q \leq r. \tag{5.14}$$

For instance, for the case $r=1, j=0$ we have $\tilde{K}(u) = K(u)$, and for $r=2, j=1$ (estimation of m') $\tilde{K}^{(1)}(u) = \mu_2^{-1} u K(u)$. This means that for estimating m itself in the interior of the x-space the effective kernel is equal to the chosen symmetric kernel function itself whereas for estimating the first derivative $\tilde{K}^{(1)}$ is a skew function. As a general result $\tilde{K}^{(j)}$ is symmetric for j even and skew for j odd.
In terms of equivalent kernels the asymptotic conditional variance and the asymptotic conditional bias (for $r-j$ odd) are

$$Var\left(\hat{m}^{(j)}(x)\right) = \frac{(j!)^2 \sigma^2(x)}{f(x) n h^{1+2j}} \int \tilde{K}^{(j)2}(u)du + o_p(nh^{-1-2j}), \tag{5.15}$$

$$Bias\left(\hat{m}^{(j)}(x)\right) = \frac{j!}{(r+1)!} m^{(r+1)}(x) h^{r+1-j} \int u^{r+1} \tilde{K}^{(j)}(u)du + o_p(h^{-r-1+j}). \tag{5.16}$$

The big advantage of local polynomial regression over other smoothing methods consists in the automatic adaptation of the active resp. equivalent kernel to the estimation situation in the boundary area. If x is scalar and $x_* = min(x_i), x^* = max(x_i)$, then for a given bandwith h the interior of the x-space is given by all observations in the interval $\left[x_* + h, x^* - h\right]$. For all x in this interval the equivalent kernels $\tilde{K}^{(j)}$ have the above mentioned symmetry resp. asymmetry property. In the left boundary part $\left[x_*, x_* + h\right]$ the number of left neighbours in a local neighbourhood of a point x will be small compared to the number of right neighbours and for $x = x_*$ we have only right neighbours. Corresponding considerations hold for the right boundary part $\left[x^* - h, x^*\right]$. For $x \in \mathbb{R}^p, (p > 1)$ the boundary area will often cover an important part of the whole design space. For $(r-j)$ odd the active resp. equivalent kernels automatically adapt to the skew data situation in the boundary area. The situation in the right boundary area is illustrated in Figure 5.1 for the Epanechnikov kernel $K(u) = \frac{3}{4}(1-u^2)_+$ for a local linear estimation of m $(r=1, j=0)$ and a local quadratic estimation of m' $(r=2, j=1)$.

We see how the weighting systems get deformed towards the boundary. The pictures for the left boundary area are symmetric to those in Figure 5.1. Since the size of the local neighbourhood shrinks towards the boundary the bias part of the mean squared error (MSE) will be lower in the boundary area than in the interior. On the other hand the variance part will increase since less observations are included in the local estimation and

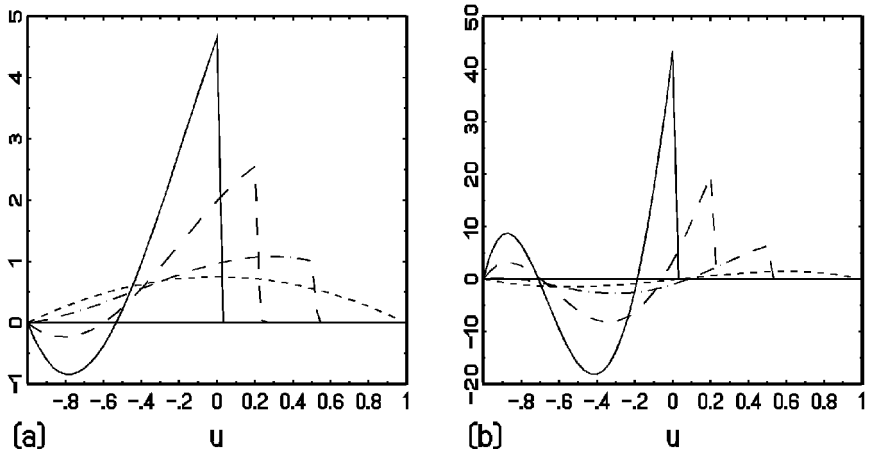

Figure 5.1 Active kernels derived from the Epanechnikov kernel with $nh = 30$ at the right boundary for (a) $r = 1, j = 0$ and (b) $r = 2, j = 1$. Estimation at interior points (short dashes), at $x = x^* - 15$ (dashes and points), at $x^* - 6$ (long dashes) and at the boundary point x^* (solid line).

also due to the increasing deformation of the weighting system towards the boundary. Usually, the increase in variance overcompensates the reduction of the bias, particularly if m'' remains roughly the same in the boundary area. As a conseqence, the MSE will increase towards the boundary. The increase will be even more pronounced for higher order polynomials.

For $x \in \mathbb{R}^p$ the local linear fit is given as the solution of the least squares criterion

$$\sum_{i=1}^{n} \Big[y_i - \beta_0 - \beta'(x_i - x) \Big]^2 K\Big(\frac{x_i - x}{h}\Big),$$

where K is a p–variate kernel. With the design matrix X_x with rows $\Big(1, (x_{i1} - x_1), \ldots, (x_{ip} - x_p)\Big)$ the solution has the same form as in (5.7).
Let K be a product kernel composed of the same univariate kernel and bandwidth h in each coordinate and let $H_m(x)$ be the Hessian matrix of the second derivatives of m. Then we get an asymptotic expression for the variance and the bias in the interior (see Ruppert and Wand, 1994)

$$Var\Big(\hat{m}(x)\Big) = \frac{\nu_0 \sigma^2(x)}{f(x) n h^p} + o_p(nh^p), \tag{5.17}$$

and

$$Bias\big(\hat{m}(x)\big) = \frac{h^2}{2}\mu_2 tr\{H_m(x)\} + o_p(ph^2). \tag{5.18}$$

The above considerations about the advantage of a local linear approach compared to the local constant estimation, about its design adaptation property and its automatic boundary adaptation hold for the multivariate case in a similar way.

Up to now we considered local least squares regression to estimate the mean function m. But the idea of locally weighted regression turns out to be a very versatile tool for estimation in a variety of situations.
Yu and Jones (1998) consider the estimation of the conditional distribution function $F(y|x)$. Let $F_1(u) = \int_{-\infty}^{u} K_1(v)dv$ be the distribution function pertaining to a symmetric kernel density K_1 and let h_2 be a bandwidth. Yu and Jones consider a local linear approach for $F(y|x)$ which is motivated by the approximations

$$E\Big[F_1(\frac{y-y_0}{h_2})|x_0\Big] \approx F(y_0|x_0)$$

and

$$F(y_0|x_0) \approx F(y_0|x) + \dot{F}(y_0|x)(x-x_0) = \beta_0 + \beta_1'(x-x_0),$$

where $\dot{F}(y_0|x) = \partial F(y_0|x)/\partial x$.
This suggests the least squares approach

$$\sum_{i=1}^{n}\Big[F_1(\frac{y_i-y}{h_2}) - \beta_0 - \beta'(x_i-x)\Big]^2 K(\frac{x_i-x}{h_1}),$$

where K is a second kernel with bandwidth h_1. The solution

$$\tilde{F}_{h_1,h_2}(y|x) = \hat{\beta}_0 = e_1'(X_x'W_xX_x)^{-1}X_x'W_x\tilde{y} \tag{5.19}$$

with $\tilde{y} = \Big(F_1(\frac{y_1-y}{h_2}), \ldots, F_1(\frac{y_n-y}{h_2})\Big)'$ is called a *local linear double-kernel smoothing* by the authors. The estimator is continuous and has zero as left boundary value (for $y \longrightarrow -\infty$) and 1 as right boundary value. It can happen that the estimator ranges outside $[0,1]$. But this does not, as the authors say, give problems estimating q_α by

$$\tilde{q}_\alpha(x) = \tilde{F}^{-1}_{h_1,h_2}(\alpha|x).$$

This estimator involes the problem that two bandwidths h_1 and h_2 have to be chosen. For a possible procedure with $h_2 < h_1$ we refer to the paper.

Fan, Yao and Tong (1996) considered a related idea for estimating the conditional density itself.

$$\begin{aligned} E\left[\frac{1}{h_2}K_1\left(\frac{y-y_0}{h_2}\right)\right] &\approx g(y_0|x) + \dot{g}(y_0|x)(x-x_0) \\ &= \beta_0 + \beta'(x-x_0) \end{aligned}$$

with $\dot{g}(y|x) = \partial g(y|x)/\partial x$ leads to the least squares criterion

$$\sum_{i=1}^{n}\left[\frac{1}{h_2}K_1\left(\frac{y_i-y}{h_2}\right) - \beta_0 - \beta'(x-x_0)\right]^2 K\left(\frac{x_i-x}{h_1}\right) \tag{5.20}$$

with the solution $\hat{g}(y|x) = \hat{\beta}_0$ as in (5.19), where now the vector $\tilde{y}$ is

$$\tilde{y} = \frac{1}{h_2}\left(K_1\left(\frac{y_1-y}{h_2}\right), \ldots, K_1\left(\frac{y_n-y}{h_2}\right)\right)'.$$

The local constant approach leads to the traditional estimator (2.3). Fan, Yao and Tong also consider the case of a local quadratic approach for estimating the first derivative. We will not pursue this case further here, since for the quadratic term $p(p+1)/2$ more parameters have to be estimated.

In all local regression approaches so far we used the least squares criterion. Let us now look at cases were instead of the square function another convex loss function $\rho : \mathbb{R} \to \mathbb{R}$ is used which has a unique minimum at zero and let $m_\rho(x) = argmin_{\beta_0} E\left[\rho(y-\beta_0)|x\right]$. $\rho(u) = u^2$ yields the conditional expectation which we analyzed mostly so far. $\rho(u) = |u|$ yields the conditional median. This is just a special case for $\alpha = 1/2$ of the loss function $\rho_\alpha(u) = |u| + (2\alpha - 1)u$, already mentioned in (3.6) . ρ_α was introduced by Koenker and Basset for parametric quantile estimation. The function $2\rho_\alpha(u)$ for various α is exhibited in Figure 5.2.

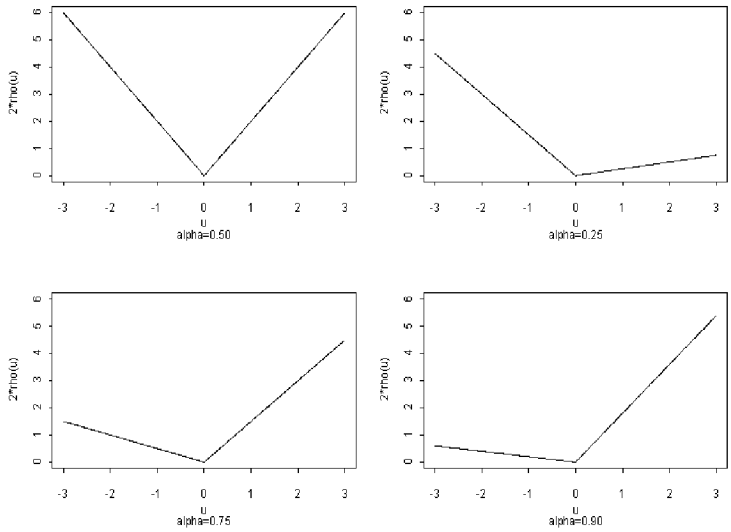

Figure 5.2 $2*\rho_\alpha(u)$ according to Koenker and Basset for several α

In robustness considerations ρ-functions were introduced which increase less rapidly than the square function and for which ρ' is the so-called ψ-function. See Huber (1981) or Hampel et al (1986).
A local constant estimator for m_ρ is

$$\hat{m}_\rho(x) = \text{arg}min_{\beta_0} \sum_{i=1}^{n} \rho(y_i - \beta_0) K(\frac{x_i - x}{h}).$$

The known drawbacks of a local constant approach is that it cannot adapt to unbalanced design situations and that it has adverse boundary effects which require boundary corrections.
This idea leads to the estimator

$$\hat{m}_\rho(x) = \hat{\beta}_0$$

where

$$(\hat{\beta}_0, \hat{\beta}) = \text{arg}min_{\beta_0,\beta} \sum_{i=1}^{n} \rho\big(y_i - \beta_0 - \beta'(x - x_0)\big) K\big(\frac{x - x_0}{h}\big). \tag{5.21}$$

For a ρ-function belonging to a robustness class, such as Huber's M-type estimators known methods for robust estimation can be applied in order to solve the minimum problem (5.21). We would like to remark that the use of kernels automatically safeguards against large deviations in the design space. For nonparametric robust M-, L- and R-estimation in a time series setting see Michels (1992).
For a local α-quantile regression with the ρ_α function (3.6) the local solution in (5.18) can be evaluated by solving a linear programming problem, as was shown in the paper of Koenker and Basset (1978). An algorithm for evaluating this can be found in Koenker and Dorey (1987).
For the case of a general convex ρ-function and i. i. d. observations asymptotic normality is proved in Fan, Hu and Truong (1994). The α-quantile estimation according to (5.21) is also considered by Yu and Jones (1998) and compared with the estimator (5.19). For reasons of practical performance the authors prefer the double smoothing approach (5.19). They also give an asymptotic expression for the mean squared error for x scalar, which for the solution of (5.21) is given by

$$\begin{aligned} MSE\big(\hat{q_\alpha}(x)\big) &= Bias^2\big(\hat{q_\alpha}(x)\big) + Var\big(\hat{q_\alpha}(x)\big) \\ &= \frac{1}{4}h^4\mu_2^2 q_\alpha''(x) + \frac{\nu_0\alpha(1-\alpha)}{nhf(x)f(q_\alpha(x)|x)^2}. \end{aligned}$$

These expressions are used for suggestions of bandwidth choice.
The cases of robust locally linear regression and of quantile regression are also considered in Fan and Gijbels (1996).

6 Applications of locally weighted regression to time series

Local linear or higher order polynomial regression, originally mainly considered for independent data, can be applied in the same way to stationary processes with certain memory restrictions. The reasons are the same as those mentioned at the beginning of section 3. Given two (dependent) random variables x_s and x_t and a point x in the design space, the random variables $\frac{1}{h}K(\frac{x_s-x}{h})$ and $\frac{1}{h}K(\frac{x_t-x}{h})$ are nearly uncorrelated as $h \to 0$. This is the *whitening by windowing principle* and it is worthwile mentionening that this property is not shared by parametric estimators. To handle memory restrictions in the proofs of consistency and asymptotic normality mixing conditions (strong mixing, uniform mixing or ϕ-mixing) are used. They give a bound to the maximal dependence between events being at least k instants apart from each other. Short term dependence does not have

much effect on local regression. But local polynomial techniques are also applicable under weak dependence in medium or long term. If suitable mixing conditions are fulfilled, local polynomial estimators for dependent data have the same asymptotic properties as for independent data. Of course the bias is not influenced by dependence, whereas the variance terms are affected. In proving asymptotic equivalence then the task consists in showing that the additional terms due to nonvanishing covariances between the variables are of smaller order asymptotically.
For a local linear estimation of $m(x) = m(x_1, \ldots, x_p)$ in the autoregressive model (3.2) the design matrix and the vector y have the form

$$X_x = \begin{pmatrix} z_p - x_1, & \ldots & z_1 - x_p \\ \vdots & & \\ z_{n-1} - x_1, & \ldots & z_{n-p} - x_p \end{pmatrix}, \qquad y = \begin{pmatrix} z_{p+1} \\ \vdots \\ z_{t-1} \end{pmatrix},$$

and with $(x_t - x)' = (z_{t-1} - x_1, \ldots, z_{t-p} - x_p)'$ the esimator can be evaluated as in (5.7). For $x = x_{n+1} = (z_n, \ldots, z_{n-p+1})'$,

$$\hat{m}(x_{n+1}) = \hat{\beta}_0$$

yields the one-step ahead predictor. A direct k-step ahead predictor is given if $y = (z_{p+k}, \ldots, z_n)'$ and if the last row of the X_x-matrix is $(z_{n-k} - z_n, \ldots, z_{n-k-p+1} - z_{n-p+1})$. But in this case a succession of one-step ahead predictions seems preferable, as already mentioned in section 3.

Asymptotic normality results for locally linear autoregression can be found in Härdle, Tsybakov an Yang (1997) and in Fan and Gijbels (1996).

For the CHARN model $z_t = m(x_t) + \sigma(x_t)\xi_t$ the function $g(x_t)$ according to (3.9) can be estimated in a similar way as above, where only in the vector y the time series values are replaced by the squares. Asymptotic normality for this case is shown in Härdle and Tsybakov (1997). For a residual based estimator of $\sigma^2(x)$ see (7.10) or Feng and Heiler (1998a).
The local linear estimation of a conditional density in a time series setting with the before mentioned double smoothing procedure as in (5.19) is considered in Fan, Yao, and Tong (1996) and in Fan and Gijbels (1996), where also asymptotic results can be found.

For the estimation of of the conditional distribution function according to the proposal of Yu and Jones (1998) as in (5.19) and for a general solution of (5.21) asymptotic

results are known for independent data. See the papers of Yu and Jones (1998), Härdle and Gasser (1984) and Tsybakov (1986). For dependent data, we have not found yet formally puplished proofs. But considering the *whitening by windowing* effect makes it clear that for these cases consistently results will hold under suitable mixing conditions.

7 Parameter selection

One of the first questions to be answered in the application of kernel smoothing is which type of kernel to use for different choices of r and j. It is well known that for $r-j$ odd in the interior of the x-space the Epanecknikov kernel $K(u) = \frac{3}{4}(1-u^2)_+$ is the one which minimizes the mean squared error in the class of all nonnegative, symmetric and Lipschitz continuous functions and that for the endpoints x_* and x^* the triangular kernels $(1-u)\mathbf{1}_{[0,1]}(u)$ resp. $(1+u)\mathbf{1}_{[-1,0]}$ are optimal. For other points in the boundary area optimal solutions are not known.
It is easy to see that when looking at variance only the uniform kernel $\frac{1}{2}\mathbf{1}_{[-1,1]}(u)$ is the one minimizing the variance.
It is well known that in practice the choice of the kernel is not very important compared to the choice of the bandwidth. The Epanecknikov kernel will therefore be a good choice in many cases. Nonetheless in practice often higher oder kernels like the Bisquare or the Triweight are prefered. This has to do with the degree of smoothness, since the kernel estimates inherit the smoothness properties of the kernel. According to the degree of smoothness as introduced by Müller (1984), the uniform kernel has degree zero (not continuous), the triangle and the Epanecknikov kernel have degree 1 (continuous, but first derivate not continuous), the Bisquare and the Triweight have degrees 2 and 3, respectively, and the Gaussion kernel has degree ∞.

The most crucial task in kernel smoothing is bandwidth selection. Much ink has been spoiled on papers concerning this problem. It is hence impossible to give a comprehensive survey here. Instead we will discuss only a few basic ideas. The aim is to choose bandwidths such that the conditional mean squared error, given by

$$MSE(\hat{m}^{(j)}(x)) = Bias^2(\hat{m}^{(j)}(x)) + Var(\hat{m}^{(j)}(x)) \tag{7.1}$$

becomes minimal. We have to distinguish between a locally optimal banwidth and a globally optimal, constant banwidth.
It is clear that a large bandwidth will lead to a low variance, but a high bias. Decreasing the bandwidth will increase the variance, but reduce the bias. An optimal bandwidth is

achieved when the changes in bias and variance balance.

Using the asymptotic expressions (5.15) and (5.16) for the conditional variance and bias, then minimizing (7.1) with respect to h yields for the (asymptotically) optimal bandwidth at x for a scalar x

$$h_n^* = C_{r,j}(K) \left[\frac{\sigma^2(x)}{(m^{(r+1)}(x))^2 f(x)} \cdot \frac{1}{n} \right]^{1/(2r+3)}, \tag{7.2}$$

where the constant

$$C_{r,j}(K) = \left[\frac{((r+1)!)^2 (2j+1) \int \tilde{K}^{(j)}(u)^2 du}{2(r+1-j) \left\{ \int u^{r+1} \tilde{K}^{(j)}(u) du \right\}^2} \right]^{1/(2r+3)} \tag{7.3}$$

depends only on r, j and the used kernel and can be calculated beforehand.
In time series applications we are mainly interested in a constant, global bandwidth, for which the integrated mean squared error $(IMSE)$

$$\int \left[Bias(\hat{m}^{(j)}(x))^2 + Var(\hat{m}^{(j)}(x)) \right] w(x) dx$$

is chosen as criterion, where w is a weight function going to zero at the bounderies to avoid boundery effects. Minimizing the $IMSE$ with respect to h yields the optimal global bandwidth

$$h_n^* = C_{r,j}(K) \left[\frac{\int \frac{\sigma^2(x)}{f(x)} w(x) dx}{\int \left\{ m^{(r+1)}(x) \right\}^2 w(x) dx} \cdot \frac{1}{n} \right]^{1/(2r+3)}. \tag{7.4}$$

For local linear estimation of m when x is a p-vector and the same bandwidth is chosen in each coordinate a similar expression can be derived (see Feng and Heiler, 1998a). Here

$$h_n^* = c_0 \left(\frac{p}{n} \right)^{\frac{1}{(p+4)}}$$

where

$$c_0 = \left[\frac{\nu_0}{\mu_2^2}\frac{\sigma^2(x)}{f(x)tr\{H_m(x)\}}\right]^{\frac{1}{(p+4)}}$$

and $H_m(x)$ is the matrix of second derivatives of m. All these expressions contain quantities which are unknown and are therefore not amenable in practice. So called *plug-in techniques* substitute these quantities by pilot estimates. For more details see Ruppert, Sheather and Wand (1995).

A simple procedure of bandwidth selection for independent data, firstly developed to find the smoothing parameter in spline smoothing, is *cross validation*. Let $\hat{m}_{h,i}(x_i)$ be the so-called *leave one out* estimator of m at x_i, where the observation (y_i, x_i) is not used in the estimation procedure. Then the criterion is

$$CV(h) = n^{-1}\sum_{i=1}^{n}[y_i - \hat{m}_{h,i}(x_i)]^2 \tag{7.5}$$

and $h_{CV} = argmin\ CV(h)$ is the cross validation bandwidth selector. The idea can also be used for $x \in \mathbb{R}^p$ and for estimating derivatives. See Härdle (1990) for details. It can be shown that it converges almost shurely to the $IMSE$ optimal bandwidth, but the convergence rate is with $n^{-1/10}$ very low. The cross validation idea was developed for independent data. In a time series setting it is suggested to replace the leave one out estimator by a "leave block out" estimator, where for estimating at x_i not only the i^{th} observation is omitted, but a whole block of data around (y_i, x_i). This idea was used by Abberger (1995, 1996) in smoothing the conditional α-quantile, where the square function is replaced by the ρ_α-function (3.6).

Let σ^2 be the variance of the residuals in an i.i.d. sample and in the time series case the unconditional variance of the stationary process. Rice (1983, 1984) proposed a criterion R which for a general linear smoother is given by

$$R(h) = RSS(h) - \hat{\sigma}^2 + 2\hat{\sigma}^2 n^{-1}\sum_{i=1}^{n} w_{ni}(x_i), \tag{7.6}$$

where the w_{ni} are the actual weights for estimating $m(x_i)$, $\hat{\sigma}^2$ is an estimate for σ^2 and

$$RSS(h) = n^{-1}\sum_{i=1}^{n}[y_i - \hat{m}_h(x_i)]^2 \tag{7.7}$$

is the mean *residual sum of squares*. Under the assumption that $\hat{\sigma}^2$ is a consistent estimator Rice (1984) showed that the proposed estimator $h_R = argmin\ R(h)$ is asymptotically optimal in the sense that $(h_R - h_0)/h_0 \to 0$ in probability, where h_0 is the minimizer of the *mean averaged squared error*

$$MASE(h) = n^{-1} E\left\{\sum_{i=1}^{n} [\hat{m}_h(x_i) - m(x_i)]^2\right\}.$$

The rate of convergences of h_R is the same low rate $n^{-1/10}$ as for the cross validation solution h_{CV}. The main differences between the two is that R involves an estimate of σ^2, whereas CV does not.
For $\hat{\sigma}^2$ Rice proposed an estimator based on first differences, whereas Gasser et al. (1986) suggested to take second differences (since they annihilate a local linear mean value function),

$$\hat{\sigma}_G^2 = \frac{2}{3(n-2)} \sum_{i=1}^{n-2} \left[y_{i+1} - \frac{1}{2}(y_i + y_{i+2})\right]^2. \tag{7.8}$$

An estimator based on a general *difference sequence* $D_m = \{d_0, d_1, \ldots, d_m\}$ such that $\sum_0^m d_j = 0$ and $\sum_0^m d_j^2 = 1$ was considered by Hall et al. (1990) . The variance estimator based on D_m is then

$$\hat{\sigma}_m^2 = (n-m)^{-1} \sum_{i=1}^{n-m} \left(\sum_{j=0}^{m} d_j y_{j+i}\right)^2. \tag{7.9}$$

Fan and Gijbels (1995) suggest the *residual sum ofsquares criterion* (RSC), which is based on a local estimator of the conditional variance derived under a local homogeneity assumption,

$$\hat{\sigma}^2(x) = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 K\left(\frac{x_i - x}{h}\right)}{tr\left[W_x - W_x(X_x' W_x X_x)^{-1} X_x' W_x\right]}. \tag{7.10}$$

With this the RSC is defined as

$$RSC(x;h) = \hat{\sigma}^2(x)\left[1 + (r+1)V\right], \tag{7.11}$$

where V is the first diagonal element of the matrix $(X_x'W_xX_x)^{-1}(X_x'W_x^2X_x)(X_x'W_xX_x)^{-1}$. V^{-1} reflects the effective number of local data points. RSC admits the following interpretation. If h is too large, then the bias is large and hence also $\hat{\sigma}^2(x)$. When the bandwidth is too small, then V will be large. Therefore RSC protects against extreme choices of h.
The minimizer of $E[RSC(x;h)]$ can be approximated by

$$h_{n0}(x) = \left[\frac{a_0\sigma^2(x)}{2C_r\beta_{r+1}^2 nf(x)}\right]^{1/(2r+3)}, \tag{7.12}$$

where a_0 denotes the first diagonal element of the matrix $S^{-1}S^*S^{-1}$, i.e. $a_0 = \int \tilde{K}^2(u)du$ and $C_r = \mu_{2r+2} - c_r'S^{-1}c_r$ with the definitions given in section 5 and $\beta_{r+1} = m^{(r+1)}(x)/(r+1)!$. $h_{n0}(x)$ differs from the optimal bandwidth in (7.3) by an adjusting constant which only depends on r, j, and the kernel used. Hence the latter one can be evaluated,

$$h_n^*(x) = Ad_{j,r}h_{n0}(x), \tag{7.13}$$

where

$$Ad_{j,r} = \left[\frac{(2j+1)C_r\int(\tilde{K}^{(j)}(u))^2du}{(r+1-j)\left\{\int u^{r+1}\tilde{K}^{(j)}(u)du\right\}^2\int\tilde{K}(u)^2du}\right]^{1/(2r+3)}.$$

For the Epanechnikov and the Gaussian kernel these constants are tabulated for various r and j in Fan and Gijbels (1996).

For a global bandwidth the minimizer $\hat{h}$ of the integrated RSC,

$$IRSC(h) = \int RSC(x;h)dx$$

is taken, which in practice breaks down to evaluating a mean over certain grid points $x_{i_1}, \ldots, x_{i_m}$. $\hat{h}$ is also selected from among a number of grid points in an interval $[h_{min}, h_{max}]$. The global bandwidth is then given by

$$\hat{h}_{j,r} = Ad_{j,r}\hat{h}. \tag{7.14}$$

The RS criterion suffers also from having a low convergence rate. Therefore the following refined bandwidth selection procedure is suggested. It is a *double smoothing* (DS) procedure. The pilot smoothing consists in fitting a polynomial of order $r+2$ and selecting $\hat{h}_{j,r}$ as above . With the bandwidth $\hat{h}_{r+1,r+2}$ estimates of $\hat{\beta}_{r+1}, \hat{\beta}_{r+2}$ and $\hat{\sigma}^2(x)$ are evaluated. With these pilot estimates in a second stage the $\widehat{MSE}_{(j,r)}(x;h) = \widehat{Bias}^2_{j,r}(x) + \widehat{Var}_{j,r}(x)$ is evaluated, where $\widehat{Bias}_{j,r}(x)$ denotes the $(j+1)^{th}$ element of the estimated bias vector and $\widehat{Var}_{(j,r)}(x)$ is the $(j+1)^{th}$ diagonal element of the matrix $(X_x'W_xX_x)^{-1}(X_x'W_x^2X_x)(X_x'W_xX_x)^{-1}\hat{\sigma}^2(x)$. With $S_{n,l} = \sum_{i=1}^{n} K\left(\frac{x_i-x}{h}\right)(x_i-x)^l$ the bias vector is estimated by

$$\hat{b}_r(x) = (X_x'W_xX_x)^{-1} \begin{pmatrix} \hat{\beta}_{r+1}S_{n,r+1} & + & \hat{\beta}_{r+2}S_{n,r+2} \\ \vdots & & \\ \hat{\beta}_{r+1}S_{n,2r+1} & + & \hat{\beta}_{r+2}S_{n,2r+2} \end{pmatrix}.$$

In order to avoid collinearity effects it is suggested to modify the vector on the right side by putting $S_{n,r+3} = \ldots = S_{n,2r+2} = 0$, which yields

$$\hat{b}_r(x) = (X_x'W_xX_x)^{-1} \begin{pmatrix} \hat{\beta}_{r+1}S_{n,r+1} & + & \hat{\beta}_{r+2}S_{n,r+2} \\ \hat{\beta}_{r+1}S_{n,r+2} & & \\ 0 & & \\ \vdots & & \\ 0 & & \end{pmatrix}.$$

The global refined bandwidth selector is then given by the minimizer $\hat{h}^R_{j,r}$ of

$$\int \hat{MSE}_{j,r}(x;h)dx. \tag{7.15}$$

This refined technique leads to an important improvement over the RSC bandwidth selector.

For a balanced design, i.e. for equally spaced x values, Heiler and Feng (1998) propose a simple double smoothing procedure, where in the pilot estimation step the R-criterion

is used. In Feng and Heiler (1998b) a further improvement of this proposal can be found, where a variance estimator based on the bootstrap idea is used. Equally spaced x values are for instance given in a time series setting where the regressor is the time index or a function of the time index. This kind of smoothing will be discussed in the next section.

For order selection in a time series autoregression model with $x_t = (z_{t-1}, \ldots, z_{t-p})$ and $\hat{m}_t(x)$ being the leave one out estimator according to (5.7), Cheng and Tong (1992) use the cross validation criterion

$$CV(p) = (n - r + 1)^{-1} \sum_t [z_t - \hat{m}_t(x_t)]^2 w(x_t). \tag{7.16}$$

where w is a weight function to avoid boundary effects.

Due to the curse of dimensionality problem it may be advisable not to take all lagged values $z_{t-1}, \ldots, z_{t-p}$ into account but to look for a subset of lagged values which yields the best forecasts. For a lag constellation $x_t(i) = (z_{t-i_1}, \ldots, z_{t-i_p})'$ Tiøstheim and Auestad (1994) propose to use the final prediction error

$$FPE(x_t(i)) = n^{-1} \sum_t [z_t - \hat{m}(x_t(i))]^2 f(i), \tag{7.17}$$

where the factor

$$f(i) = \frac{1 + (nh^p)^{-1} \nu_0 b_p(i)}{1 - (nh^p)^{-1} [2K^p(o) - \nu_o^p] b_p(i)},$$

and $\nu_o = \int K^2(u)du, \quad b_p(i) = n^{-1} \sum \frac{w^2(x_t(i))}{\hat{f}(x_t(i))},$

$\hat{f}(x_t(i))$ being a multivariate kernel density estimator. FPE in (7.16) is essentually a sum of squares of one-step ahead prediction errors multiplied with a factor that penalizes small bandwidths and a large order p.

8 Time series decomposition with locally weighted regression

As already mentioned in section 3, if x_t is the time index itself or a polynomial in t, then we arrive at trend smoothing. In a simple trend model

$$z_t = m(t) + a_t$$

the considerations at the beginning of section 5 deliver an estimator of the smooth trend function or its derivatives. Now the matrix X_t has the rows $(1, s-t, \ldots, (s-t)^r)$ for $s = 1, \ldots, n$ and $W_t = diag(K(\frac{s-t}{h}))$. As an interesting fact one can easily see that in the interior of the time series, i.e. for $h \leq t \leq n-h$ the weights given in (5.8),

$$w_{nt}^j(s) = e'_{j+1}(X_t'W_tX_t)^{-1}\,(1, s-t, \ldots, (s-t)^r)\,K\left(\frac{s-t}{h}\right),$$

are shift invariant in the sense $w_{n,t+1}^j(s+1) = w_{nt}^j(s)$. This means that in the interior of the time series the local polynomial fit works like a moving average. But the big advantage over other trend smoothing techniques lies in the automatic boundary adaptation of the procedure. This property makes the idea of extending the local regression approach to so-called unobserved components models very appealing.

Nonparametric estimation of trend-cyclical movements and of seasonal variations and their separation by local regression represents an interesting alternative to procedures based on parametric models like X–12 or TRAMO–SEATS. These involve extrapolation methods on either end of the time series in order to be able to estimate the components also in the boundary parts of a time series. This can lead to serious problems if unusual observations in the end parts of time series yield grossly erroneous forecasts. The latter problem will not appear with a local regression approach. Note also that with a data driven parameter selection the procedure works in a fully automatic way.
The decomposition of a time series into trend–cyclical and seasonal components by LOcally WEighted Scatterplot Smoothing (LOWESS) was suggested by Cleveland et al. (1990). The procedure discussed here is different from their procedure in essential features.

We consider the additive (unobserved) components model

$$z_t = T(t) + S(t) + a_t, \quad t = 1, 2, \ldots \tag{8.1}$$

For the sake of simplicity we assume that $\{a_t\}$ is a white noise sequence with mean zero and constant variance σ^2. $T(t)$ represents the trend cyclical and $S(t)$ the seasonal component. The usual assumption with respect to T is that it has certain smoothness properties so that the considerations at the beginning of section 5 apply, leading to a local polynomial representation of order r. With respect to the seasonal variations the usual assumption is that they show a similar pattern from one seasonal period to the next, but they are allowed to vary slowly in the course of time. Hence a natural assumption is that they can locally be approximated by a Fourier series, containing the seasonal frequency and its harmonics,

$$S(s) = \sum_{j=1}^{q} \left[\alpha_j(t) \cos 2\pi\lambda j(s-t) + \gamma_j(t) \sin 2\pi\lambda j(s-t)\right], \tag{8.2}$$

where λ is the seasonal frequency, $\lambda = 1/P$ and P is the period of the season. Of course $\lambda_q \leq 1/2$ (and for $\lambda_q = 1/2$ the last sine term has to be omitted).
Let

$$\begin{aligned} u_t(s) &= (\cos 2\pi\lambda(s-t), \sin 2\pi\lambda(s-t), \ldots, \cos 2q\pi\lambda(s-t), \sin 2q\pi\lambda(s-t))', \\ \alpha(t) &= (\alpha_1(t), \gamma_1(t), \ldots, \alpha_q(t), \gamma_q(t))'. \end{aligned}$$

Then $S(s) = \alpha(t)' u_t(s)$.
With the local polynomial representation for the trend-cyclical part

$$T(s) = \sum_{j=0}^{r} \beta_j(t)(s-t)^j = \beta(t)' x_t(s),$$

where $\beta(t) = (\beta_0(t), \ldots, \beta_r(t))'$, $x_t(s) = (1, s-t, \ldots, (s-t)^r)'$, the local least squares criterion is

$$\sum_{s=1}^{n} \left[z_t - \beta(t)' x_t(s) - \alpha(t)' u_t(s)\right]^2 K\left(\frac{s-t}{h}\right). \tag{8.3}$$

With the design matrices X_{1t} with rows $x_t(s)'$, X_{2t} with rows $u_t(s)'$, $X_t = (X_{1t} \vdots X_{2t})$, the composed vector $\gamma(t)' = (\beta(t)', \alpha(t)')$ and the weight matrix $W_t = diag\left(K\left(\frac{s-t}{h}\right)\right)$ the solution is

$$\hat{\gamma}(t) = (X_t' W_t X_t)^{-1} X_t' W_t y \tag{8.4}$$

$$\hat{T}(t) = e_1' (X_t' W_t X_t)^{-1} X_t' W_t y \tag{8.5}$$

$$\hat{S}(t) = (o', \phi_s') (X_t' W_t X_t)^{-1} X_t' W_t y, \tag{8.6}$$

where o' is a row of zeroes of length $r+1$ and ϕ_s' is a row vector of length $2q$ with entries $\phi_s' = (1\ 0\ 1\ 0 \ldots 1\ 0)$. It picks out the $\hat{\alpha}_j(t)$, pertaining to the cosine terms in $\hat{S}(t)$. The estimator for the j^{th} derivative $T^{(j)}$ of T is

$$\hat{T}^{(j)} = j! e_{j+1}' (X_t' W_t X_t)^{-1} X_t' W_t y. \tag{8.7}$$

All the above estimators work as moving averages in the interior part of the time series and have for $r-j$ odd the simple boundary adaptation property discussed in section 5. The decomposition $\hat{m}(t) = \hat{T}(t) + \hat{S}(t)$ is not unique, since the matrix $X_t' W_t X_t$ is not block diagonal. This could of course be achieved by an orthogonalization procedure but seems not to be compelling for practical purposes. We call the above decompostion a *natural decomposition.*

For parameter selection first a decision has to be made about the degree of the trend polynomial T and the trigonometric polynomial S. Since the seasonal variations are involved in the local approach the bandwidths should be such that at least three to five periods of the season are included. In order to achieve this, the modelization of T should be rather flexible. Hence for the interior part of the time series the polynomial degree $r = 3$ may be preferable to the choice $r = 1$. A data driven choice for a joint selection of r and bandwidth h is a very difficult task since the two parameters are highly correlated. A higher r allows a larger bandwidth and vice versa. In our experience collected so far a data driven procedure for the interior part always opted for the highest allowed degree r_{max} that was put beforehand even if the MSE criterion included a penalty term for overparameterization. As far as the trigonometric polynomial is concerned, all harmonic terms should be included, unless an inspection of the periodogramme or the estimated spectrum reveals that one or even more of the seasonal frequences can be ommitted.

After this preselection of parameters a procedure for bandwidth selection is needed. Since for an equidistant time series the "design density" f is a constant the procedure is somehow simpler than in the general situation discussed in section 7.

A variate of a double smoothing procedure is recommended. In the pilot stage a poly-

nomial of degree $r+2$ is fitted and the bandwidth is selected with the Rice criterion with respect to $\hat{m} = \hat{T} + \hat{S}$. But due to seasonal variations the difference based variance estimator (7.8) has to be altered. Heiler and Feng (1996) and Feng (1998) propose a seasonal difference based variance estimator of the form in (7.9), where not only a local linear function, but also a local periodic function is allowed for.
An example for monthly data $(P = 12)$ is

$$D_{26,12} = c^{-1}\{-1, 2, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, -4, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, -1, 2, -1\},$$

where c is determined such that $\sum_{j=0}^{m} d_j^2 = 1$.
$D_{26,12}$ annihilates a local linear trend and a local periodic function with periodicity $P = 12$. Similar sequences can easily be constructed.

Let $\hat{\sigma}_G^2$ be the resulting estimator and let g be the minimizer of the R-criterion(7.6). With $\hat{m}_g = \hat{T}_g + \hat{S}_g$ the resulting estimator is denoted.

For an arbitrary h the weights $w_t^h(s)$ for estimating $\hat{T}_h(t) + \hat{S}_h(t)$ are the components of the vector $(1\ 0, \ldots, 0, \phi_s')\,(X_t' W_t X_t)^{-1} X_t' W_t$, where for W_t a kernel with bandwidth h is taken.
Using the pilot estimates $\hat{m}_g(t)$ the bias part of the MSE at t for an estimator with bandwidth h is estimated by

$$\widehat{Bias}(\hat{m}_h(t)) = \sum_{s=1}^{n} w_t^h(s)\hat{m}_g(s) - \hat{m}_g(t)$$

which yields for the bias part of the *mean averaged squared error* $MASE(h)$

$$\begin{aligned} B(h) &= n^{-1}\sum_{t=1}^{n} \widehat{Bias}^2(\hat{m}_h(t)) \\ &= n^{-1}\sum_{t=1}^{n} \left\{\sum_{s=1}^{n} w_t^h(s)\hat{m}_g(s) - \hat{m}_g(t)\right\}^2. \end{aligned} \tag{8.8}$$

The variance is estimated by

$$V(h) = n^{-1}\hat{\sigma}^2 \sum_{t=1}^{n}\sum_{s=1}^{n} w_t^h(s)^2, \tag{8.9}$$

where $\hat{\sigma}^2$ should be a suitable root-n consistent estimator of $\hat{\sigma}^2$.
After the first pilot step a minimizer $\tilde{h}$ of the criterion

$$MASE(h) = B(h) + V(h) \tag{8.10}$$

is evaluated over a grid, where in the second step the estimator $\hat{\sigma}_G^2$ is used in $V(h)$. This second step leads already to a considerable improvement over the simple R-criterion, but the estimator $\hat{\sigma}_G^2$ is still not very good. Hence an improved estimation with a lower polynomial degree and a bandwith g_v larger than g is proposed. For details see Feng and Heiler (1998). According to considerations therein an estimator for g_v can easily be found by multiplying the minimizer $\tilde{h}$ of (8.10) with a *correction factor*. This factor only depends on the used kernel and on the polynomial degree r, $\hat{g}_v = CF_r\tilde{h}$.
For instance, we get for the Epanechnikov kernel $CF_1 = 1.431$, $CF_3 = 1.291$, for the Bi-square kernel $CF_1 = 1.451$, $CF_3 = 1.300$ and for the Gaussian kernel $CF_1 = 1.489$ and $CF_3 = 1.305$. See Table 5.1 in Müller (1988) or Table 1 in Feng and Heiler (1998).

Let now $\hat{m}_{g_v} = \hat{T}_{g_v} + \hat{S}_{g_v}$ be an estimator with bandwidth g_v . Then an improved variance estimator is obtained by taking the mean squared residuals

$$\hat{\sigma}_B^2 = n^{-1}\sum_{t=1}^{n}[z_t - \hat{m}_{g_v}(t)]^2. \tag{8.11}$$

In a third step this variance estimator is plugged into (8.9) for $\hat{\sigma}^2$ and with this again a minimizer h^* of the MASE (8.10) is evaluated.
In principle this procedure can be iterated several times, where in the next step with a polynomial of degree $r+2$ a new bias estimator is evaluated.

The above described procedure yields a bandwidth h^* for the interior part of the time series, where after the selection of h^* the interior is given by $[h^*+1, n-h^*]$. As described in section 5 the procedure automatically adapts towards the boundaries. But as also described there due to increasing variance the MSE will increase as well, particularly if $r = 3$ is chosen, as was recommended at the beginning of this section.
One possibility to at least partly compensate for that is to switch to a nearest neighbour estimator in the boundary area, that is, to keep the total bandwidth $h_T = 2h^*+1$ constant at both ends of the time series. This means that for estimating from $t = n - h^* + 1$ to $t = n$ the same local neighbourhood is used (and similarly for the left boundary).
Instead or in addition to that a switch from a local polynomial of order 3 to a local linear

approach (for T) may be recommended whenever the MSE for $r = 1$ becomes smaller than that for $r = 3$. In order to do that, for the given bandwidth and the asymmetric neighbourhood situation at each time point in the boundary area with the corresponding active weighting systems the $MSE's$ for $r = 3$ and $r = 1$ have to be evaluated according to the procedure described above. As soon as $MSE_1 < MSE_3$, a local linear approach is chosen for T and maintained to the end point. According to practical experiences collected so far such a switch happened to come to effect close to the end points in almost all cases.

In Figures 8.1 and 8.2 we present two examples where the discussed decomposition procedure is applied. The first time series is the quarterly series of the German GDP from 1968 to 1994. In the top panel in Figure 8.1 the time series itself and the estimated trend-cyclical component are exhibited. In the middle the estimated seasonal component is shown and in the bottom panel the first derivative of the trend-cyclical is exhibited. This latter picture shows clearly the temporary boom after German reunification. The double smoothing procedure with bootstrap variance estimator selected $h = 11$ as bandwidth. The polynomial degree was two for estimating the first derivative and three for the other estimations.
The second example presented in Figure 8.2 shows corresponding results for the monthly series of the German unemployment rates (in per cent) from January 1977 to April 1995. Here the selected bandwidth is $h = 21$. The polynomial degrees are the same as in the previous example.

Cleveland (1979) proposed an iterative robust locally weighted regression in a general regression context and in Cleveland et al. (1990) this idea is also used in time series decomposition. It can easily be adapted to the procedure discussed here, although in their proposal the subseries of equal weeks, month, quarters etc. are treated separately.
The idea consists in looking at the residuals $r_t = z_t - \hat{m}(t)$ of a first, nonrobust procedure and to evaluate a robust scale measure δ for the residuals. Cleveland suggests to take the median of the $|r_t|$. Since in many time series variability is different for different periods within the season depending on the size of the seasonal component, it seems reasonable to evaluate different scale measures for the different periods of the season.
For $t = 1, \ldots, n$ let $j = \left[\frac{t-1}{P}\right] + 1$ be the year index, $j = 1 \ldots, J = \left[\frac{n-1}{P}\right] + 1$, where [.] denotes the integer part and let $i = t - P(j-1)$ be the season index, i.e. $z_t \longrightarrow z_{ij}$. Then for all $i = 1, \ldots, P$ a robust scale measure

$$\delta_i = median_j \left(|r_{ij}|\right)$$

is evaluated. From this so-called robustness weights are derived, which according to Cleveland's proposal are given by

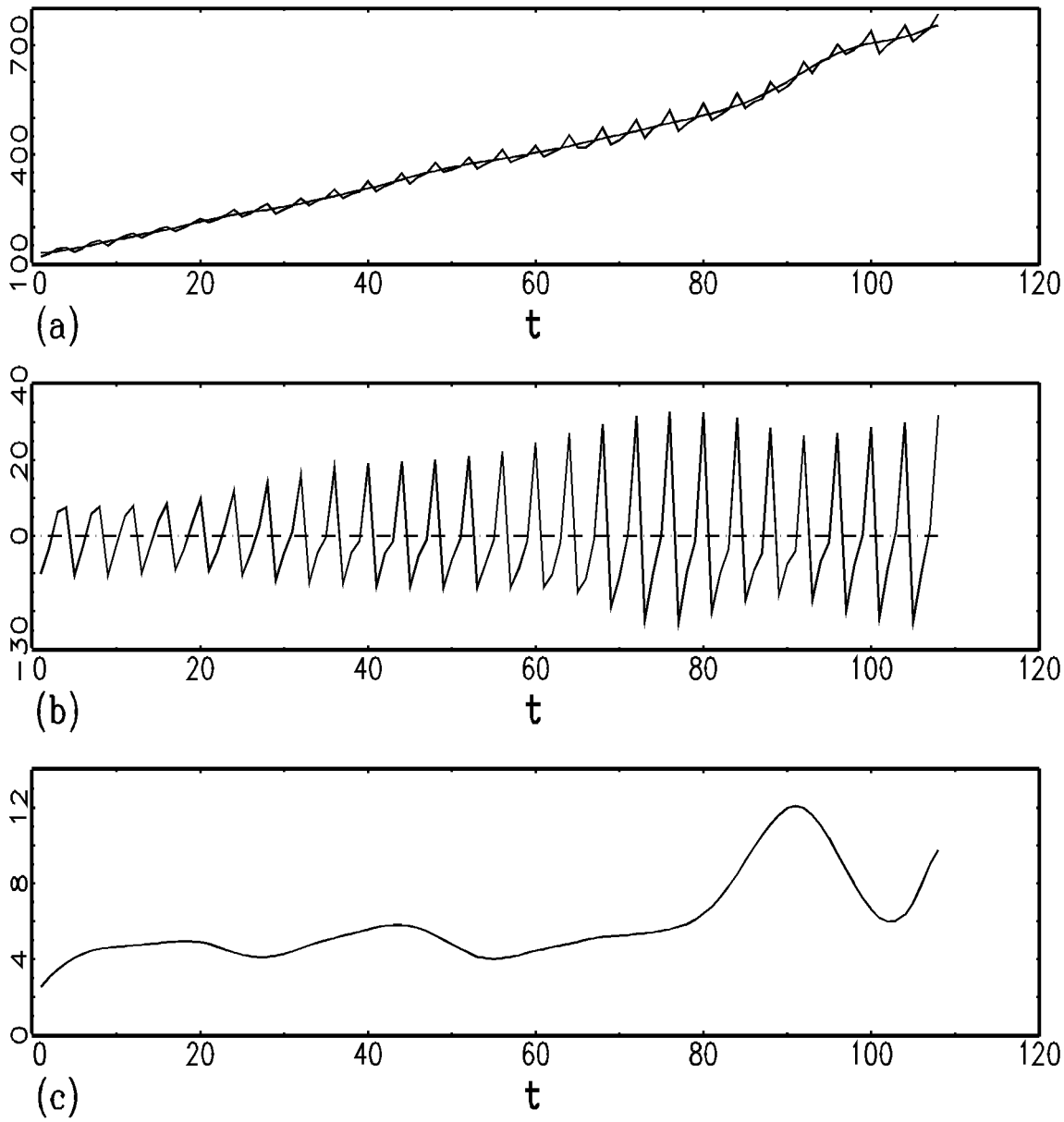

Figure 8.1. Decomposition results for the time series of the German GDP from 1968 to 1994. (a) The data and $\hat{T}$, (b) $\hat{S}$ and (c) $\hat{T}'$

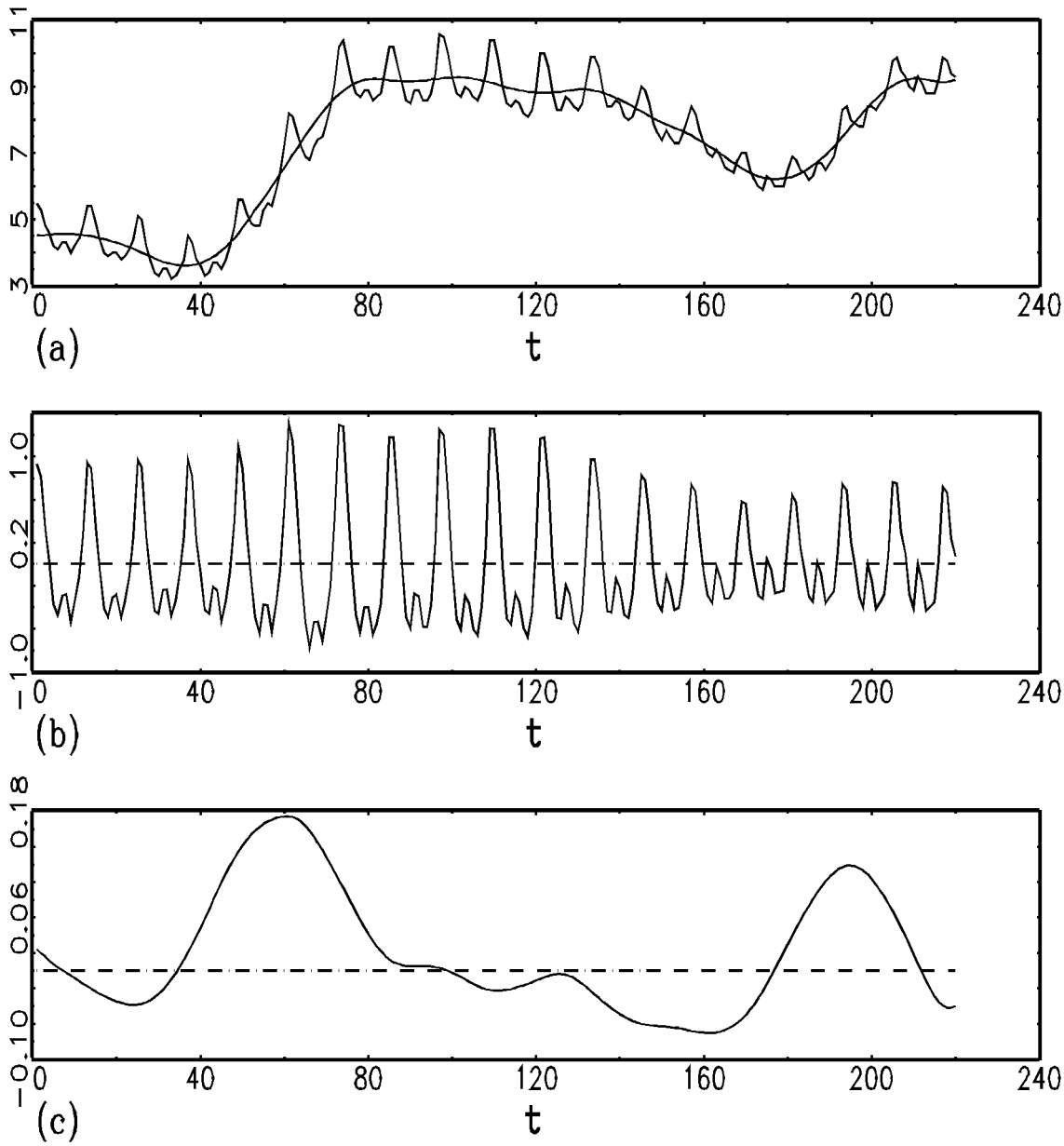

Figure 8.2. Decomposition results for the time series of the German unemployment rates (in %) from January 1977 to April 1995. (a) The data and $\hat{T}$, (b) $\hat{S}$ and (c) $\hat{T}'$

$$\beta_{ij} = K\left(\frac{r_{ij}}{6\delta_i}\right),$$

where K is a kernel function (the bisquare kernel is being suggested).
In a second step the local estimation procedure is repeated, where the neighbourhood weights $k_{st} = K\left(\frac{s-t}{h}\right)$ in the diagonal weight matrices W_t are multiplied with the corresponding robustness weights β_{ij}, where i and j are the season- and year index corresponding to s. Of course with the time dependent robustness weights the procedure is no more shift invariant, so that the least squares solution has to be evaluated for each t explicitely.
Starting with the new residuals the procedure can be iterated until the estimates stabilize. Since the robustness weights will change the active kernels, different bandwidths should be used in each iteration step. Cleveland (1979) claimed that two robust iterations should be adequate for almost all situations. In Feng (1998) with a stability criterion a higher number of iteration steps occured in most cases.

References

[1] ABBERGER, K. (1996). *Nichtparametrische Schätzung bedingter Quantile in Zeitreihen – Mit Anwendungen auf Finanzmarktdaten.* Hartung-Gorre Verlag, Konstanz.

[2] ABBERGER, K. (1997). Quantile Smoothing in Financial Time Series. *Statistical Papers*, 38, 125–148.

[3] BONGARD, J. (1960). Some Remarks on Moving Averages. In: O.E.C.D. (editor), Seasonal Adjustment on Electronic Computers. *Proceedings of an international conference held in Paris*, 361-387.

[4] CHEN, R. (1996). A Nonparametric Multi-step Prediction Estimator in Markovian Structures. *Statistica Sinica*, 6, 603-615.

[5] CHEN, R. and TSAY, R.S. (1993). Functional-coefficient Autoregressive Models. *Journal Amer. Statist. Assoc.*, 88, 298-308.

[6] CHEN, R. and TSAY, R.S. (1993). Nonlinear Additive ARX Models. *Journal Amer. Statist. Assoc.*, 88, 955-967.

[7] CHENG, B. and TONG, H. (1992). On Consistent Non-parametric Order Determination and Chaos (with discussion). *Journal Royal Statist. Soc., Series B*, 54, 427-474.

[8] CLEVELAND, R.B., CLEVELAND, W.S., McRAE, I.E. and TERPENNING, I. (1990). STL: A Seasonal-trend Decomposition Procedure Based on LOWESS (mit Diskussion). *Journal of Official Statistics*, 6, 3–73.

[9] CLEVELAND, W.S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74, 829–836.

[10] COLLOMB, G. (1980). Estimation Nonparamétrique de Probabilités Conditionelles. *Comptes Rendus à l'Académie des Sciences de Paris*, 291, Série A, 427–430.

[11] COLLOMB, G. (1983). From Nonparametric Regression to Nonparametric Prediction: Survey of the Mean Square Error and Original Results on the Predictogram. *Lecture Notes in Statistics*, 16, 182–204.

[12] COLLOMB, G. (1984). Propriétés de Convergence Presque Complète du Prédicteur à Noyau. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 66, 441–460.

[13] COLLOMB, G. (1985). Nonparametric Time Series Analysis and Prediction: Uniform Almost Sure Convergence of the k-NN Autoregression Estimates. *Statistics*, 16, 297–307.

[14] EUBANK, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York.

[15] FAN, J. (1993). Local Linear Regression Smoothers and Their Minimax Efficiencies. *Annals of Statistics*, 21, 196–216.

[16] FAN, J. and GIJBELS, I. (1992). Variable Bandwidth and Local Linear Regression Smoothers. *Annals of Statistics*, 20, 2008–2036.

[17] FAN, J. and GIJBELS, I. (1995). Data-driven Bandwidth Selection in Local Polynomial Fitting: Variable bandwidth and Spatial Adaptation. *Journal of the Royal Statistical Society*, series B, 57, 371–394.

[18] FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and its Applications.* Chapman & Hall, London.

[19] FAN, J., HU, T-CH. and TRUONG, Y.K. (1994). Robust Non-parametric Function Estimsation. *Scandinavian Journal of Statistics*, 21, 433-446.

[20] FAN, J., YAO, Q. and TONG, H. (1996). Estimation of Conditional Densities and Sensitivity Measures in Nonlinear Dynamic Systems. *Biometrika*, 83, 189-216.+

[21] FENG, Y. (1998). *Kernel- and Locally Weighted Regression with Application to Time Series Decomposition.* Ph.D. Thesis. University of Konstanz.

[22] FENG, Y. and HEILER, S. (1998a). Locally Weighted Autoregression. In: R. Galata and H. Küchenhoff (editors), Econometrics in Theory and Practice. Festschrift for Hans Schneeweiß, 101-117.

[23] FENG, Y. and HEILER, S. (1998b). Bandwidth Selection Based on Bootstrap. *Discussion Paper*. University of Konstanz.

[24] FISHER, A. (1937). A Brief Note on Seasonal Variations. *Journal of Accountancy*, 64, 174.

[25] FRIEDMAN, J.H (1991). Multivariate Adaptive Regression Splines (mit Diskussion). *Annals of Statistics*, 19, 1–141.

[26] GASSER, T., KNEIP, A. and KÖHLER, W. (1991). A Flexible and Fast Method for Automatic Smoothing. *J. Amer. Statist. Assoc.*, 86, 643–652.

[27] GASSER, T. and MÜLLER, H.G. (1979). Kernel Estimation of Regression Functions. In: Gasser and Rosenblatt (editors), *Smoothing Thecniques for Curve Estimation*, Spring-Verlag, Heidelberg, 23–68.

[28] GASSER, T. and MÜLLER, H.G. (1984). Estimating Regression Functions and Their Derivatives by the Kernel Method. *Scandinavian Journal of Statistics*, 11, 171–185.

[29] GASSER, T., MÜLLER, H.G. and MAMMITZSCH V. (1985). Kernels for Nonparametric Curve Estimation. *Journal of the Royal Statistical Society*, series B, 47, 238–252.

[30] GASSER, T., SROKA, L. and JENNEN-STEINMETZ, C. (1986). Residual Variance and Residual Pattern in Nonlinear Regression. *Biometrika*, 73, 625–633.

[31] GOURIÉROUX, CH. and MONFORT, A. (1992). Qualitative Threshold ARCH Models. *Journal of Econometrics*, 52, 159–199.

[32] HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.

[33] HÄRDLE, W., HALL, P. and MARRON, J.S. (1992). Regression Smoothing Parameters That are not far from Their Optimum. *J. Amer. Statist. Assoc.*, 87, 227–233.

[34] HÄRDLE, W. and GASSER, T. (1984). Robust Non-parametric Function Fitting. *Journal Royal. Statist. Soc., Series B*, 46, 42-51.

[35] HÄRDLE, W., LÜTKEPOHL, H. and CHEN, R. (1997). A Review of Nonparametric Time Series Analysis. *International Statistical Review*, 65, 49-72.

[36] HÄRDLE, W. and TSYBAKOV, A.B. (1988). Robust Nonparametric Regression with Simultaneous Scale Curve Estimation. *Annals of Statistics*, 16, 120–135.

[37] HÄRDLE, W. and TSYBAKOV, A.B. (1998). Local polynomial Estimators of the Volatility Function. To appear in it Journal of Econometrics.

[38] HÄRDLE, W., TSYBAKOV, A.B. and YANG, L. (1997). Nonparametric Vector Autoregression. To appear in *Journal of Statistical Planning and Inference.*

[39] HÄRDLE, W. and YANG, L. (1996). Nonparametric Time Series Model Selection. Discussion paper, Humboldt-Universität zu Berlin.

[40] HALL, P., KAY, J.W. and TITTERINGTON, D.M. (1990). Asymptotically Optimal Difference-based Estimation of Variance in Nonparametric Regression. *Biometrika*, 77, 521-528.

[41] HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J. ans STAHEL, W.A. (1986). *Robust Statistics: The Approach Based on the Influence Function.* Wiley, new York.

[42] HART, J.D. (1996).Some Automated Methods of Smoothing Time-dependent Data. *Journal of Nonparametric Statistics*, 6, 115-142.

[43] HASTIE, T.J. and TIBSHIRANI, R.J. (1990). *Generalized Additive Models.* Monographs on Statistics and Apllied Probability, 43, Chapman and Hall, London.

[44] HEILER, S. (1995). Zur Glättung Saisonaler Zeitreihen. In: Rinne, H., Rüger, B. and Strecker, H. (editors). *Grundlagen der Statistik und Ihre Anwendungen.* Festschrift für Kurt Weichselberger, Physika-Verlag, Heidelberg, 128–148.

[45] HEILER, S. and FENG, Y. (1996). Datengesteuerte Zerlegung Saisonaler Zeitreihen. *ifo Studien*, 41–73.

[46] HEILER, S. and FENG, Y. (1998). A Simple Root n Bandwidth Selector for Nonparametric Regression. *Journal of Nonparametric Statistics*9, 1-21.

[47] HEILER, S. and FENG, Y. (1997). A Bootstrap Bandwidth Selector for Local Polynomial Fitting. Discussionpaper, SFB178, II–344, University of Konstanz.

[48] HEILER, S. and MICHELS, P. (1994). *Deskriptive und Explorative Datenanalyse.* Oldenbourg-Verlag, München.

[49] HORWATH; L. and YANDELL; B.S. (1988). Asymptotics of Conditional Empirical Porcesses. *Journal of Multivariate Analysis*, 26, 184-206.

[50] HUBER, P.J. (1981). *Robust Statistics.* Wiley, New York.

[51] JONES, H.L. (1943). Fitting of Polynomial Trends to Seasonal Data by the Method of Least Squares. *Journal Amer. Statist. Assoc.*, 38, 453

[52] JONES, M.C. and HALL, P. (1990). Mean Squared Error Properties of Kernel Estimates of Regression Quantiles. *Statistics & Probability Letters*, 10, 283–289.

[53] KOENKER, R. and BASSETT, G. (1978). Regression Quantiles. *Econometrica*, 46, 33–50.

[54] KOENKER, R. and DOREY, V. (1987). Computing Regression Quantiles. *Applied Statistics*, 36, 383-393.

[55] KOENKER, R., PORTNOY, S. and Ng, P. (1992). Nonparametric Estimation of Conditional Quantile Functions. In: L_1-*Statistical Analysis and Related Methods* (ed. Y. Dodge), North-Holland, New York.

[56] MACAULAY, R.R. (1931). *The smoothing of time series.* Natinal Bureau of Economic Research, New York.

[57] MESSER, K. and GOLDSTEIN, L. (1993). A new Class of Kernels for Nonparametric Curve Estimation. *Annals of Statistics*, 21, 179–195.

[58] MICHELS, P. (1992). *Nichtparametrische Analyse und Prognose von Zeitreihen.* Physica-Verlag, Heidelberg.

[59] MÜLLER, H.-G. (1985). Empirical Bandwidth Choice for Nonparametric Kernel Regression by Means of Pilot Estimators. *Statist. Decisions*, Supp. Issue 2, 193–206.

[60] MÜLLER, H.-G. (1988). *Nonparametric Analysis of Longitudinal Data.* Springer-Verlag, Berlin.

[61] NADARAYA, E.A. (1964). On Estimating Regression. *Theory of Probability and Its Applications,* 9, 141–142.

[62] PRIESTLEY, M.B. and CHAO, M.T. (1972). Nonparametric Function Fitting. *Journal of the Royal Statistical Society*, series B, 34, 385–392.

[63] RICE, J. (1983). Methods for Bandwidth Choice in Nonparametric Kernel Regression. In: J.E. Gentle (editor), *Computer Science and Statistics: The Interface.* North Holland, Amsterdam, 186-190.

[64] RICE, J. (1984). Bandwidth Choice for Nonparametric Regression. *Annal of Statistics*, 12, 1215–1230.

[65] ROBINSON, P.M. (1983). Nonparametric Estimators for Time Series. *Journal of Time Series Analysis*, 4, 185–207.

[66] ROBINSON, P.M. (1986). On the Consistency and Finite-sample Properties of Nonparametric Kernel Time Series Regression, Autoregression and Density Estimators. *Annals of the Institute of Statistical Mathematics*, 38, A, 539–549.

[67] RUPPERT, D., SHEATHER, S.J. and WAND, M.P. (1995). An Effective Bandwidth Selector for Local Least Squares Regression. *J. Amer. Statist. Assoc.*, 90, 1257–1270.

[68] RUPPERT, D. and WAND, M.P. (1994). Multivariate Locally Weighted Least Squares Regression, *Annals of Statistics*, 22, 1346–1370.

[69] SILVERMAN, B.W. (1984). Spline Smoothing: The Equivalent Variable Kernel Method. *Annals of Statistics*, 12, 898–916.

[70] SILVERMAN, B.W. (1985). Some Aspects of the Spline Smoothing Approach to Nonparametric Regression Curve Fitting (with discussion). *Journal of the Royal Statistical Society*, series B, 47, 1–52.

[71] STONE, C.J. (1977). Consistent Nonparametric Regression (with discussion). *Annals of Statistics*, 5, 595–620.

[72] STUTE, W. (1984). Asymptotic Normality of Nearest Neighbor Regression Function Estimates. *Annals of Statistics*, 12, 917–926.

[73] STUTE, W. (1986). Conditional Empirical Processes. *Annals of Statistics*, 14, 638–647.

[74] TJØSTHEIM, D. and AUESTAD; B. (1994a). Nonparametric Identifixation of Nonlinear Time Series: Projection. *J. Amer. Statist. Assoc.*, 89, 1398-1409.

[75] TJØSTHEIM, D. and AUESTAD; B. (1994b). Nonparametric Identifixation of Nonlinear Time Series: Selecting Significant Lags. *J. Amer. Statist. Assoc.*, 89, 1410-1419.

[76] TSYBAKOV, A.B. (1986). Robust Reconstruction of Function by the Local Approximation Method. *Problems of Information Transmission*, 22, 133–146.

[77] WAHBA, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

[78] WAND, M.P. and JONES, M.C. (1995). *Kernel Smoothing.* Chapman & Hall, London.

[79] WATSON, G.S. (1964). Smooth Regression Analysis. *Sankhyā*, Ser. A, 26, 359–372.

[80] YAKOWITZ, S. (1979a). Nonparametric estimation of Markov transition functions. *Annals of Statistics*, 7, 671-679.

[81] YAKOWITZ, S. (1979b). A nonparametric Markov model for daily river flow. *Water Resour. Research*, 15, 1035-1043.

[82] YAKOWITZ, S. (1985). Markov Flow Models and the Flood Warning Problem. *Water Resources Research*, 21, 81–88.

[83] YANG, L. and HÄRDLE, W. (1996). Nonparametric Autoregression with Multiplicative Volatility and Additive Mean. Submitted to *Journal of Time Series Analysis*.

[84] YANG, S. (1981). Linear Functions of Concomitants of Order Statistics with Application to Nonparametric Estimation of a Regression Function. *Journal of the American Statistical Association*, 76, 658–662.

[85] YU, K. and JONES, M.C. (1998). Local Linear Quantile Regression. *Journal. Amer. Statist. Assoc.*, 93, 228-237.