

Heiler, Siegfried; Feng, Yuanhua

**Working Paper**

## Data-driven optimal decomposition of time series

Diskussionsbeiträge - Serie II, No. 287

**Provided in Cooperation with:**

Department of Economics, University of Konstanz

*Suggested Citation:* Heiler, Siegfried; Feng, Yuanhua (1995) : Data-driven optimal decomposition of time series, Diskussionsbeiträge - Serie II, No. 287, Universität Konstanz, Sonderforschungsbereich 178 - Internationalisierung der Wirtschaft, Konstanz

This Version is available at:

<https://hdl.handle.net/10419/101578>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

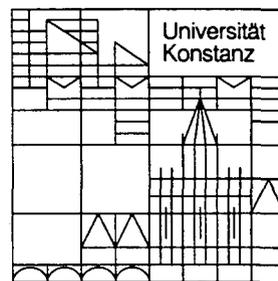
*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

**Sonderforschungsbereich 178**  
**„Internationalisierung der Wirtschaft“**

Diskussionsbeiträge



Juristische  
Fakultät

Fakultät für Wirtschafts-  
wissenschaften und Statistik

---

Siegfried Heiler  
Yuanhua Feng

**Data-Driven Optimal Decomposition  
of Time Series**

16. JAN 1996 Weltwirtschaft

W 113 (287) ~~mi~~ br sig gla

# **Data-Driven Optimal Decomposition of Time Series**

**Siegfried Heiler**

**Yuanhua Feng**

656685

**W 113 (287)**



Serie II - Nr. 287

Dezember 1995

# Data-Driven Optimal Decomposition of Time Series

Siegfried Heiler

Yuanhua Feng

University of Konstanz

## Abstract

A data-driven optimal decomposition of time series with trend-cyclical and seasonal components as well as the estimation of derivatives of the trend-cyclical is considered. The time series is smoothed by locally weighted regression with polynomials and trigonometric functions as local regressors. Two variates for the selection of the optimal bandwidths and the order of the polynomials are proposed with a particular approach for the estimation in the boundary areas of the time series. The second of these procedures can also be used for the selection of optimal bandwidths if only one component is considered. The smoothing of a time series without seasonal variations is just a special case for these procedures. The rate of convergence in the second procedure for this special case is discussed. A by-product of this work is the development of a *seasonal-difference-based* method to estimate the variance in a seasonal time series.

Keywords: Time Series Decomposition, Bandwidth Selection, Locally Weighted Regression.

## 1 Introduction

Decomposing economic time series into unobservable trend-cyclical and seasonal components has a long tradition. There exists a large number of different methodical approaches and also ready-made software systems to perform this. The method used by the German Federal Statistical Office, the Berlin-Method (BV: Berliner Verfahren) is one of the well known ones. In this approach a traditional nonparametric method, local regression, is used. Including a local fit of polynomials and trigonometric functions makes it possible to estimate the trend-cyclical and the seasonal components in a unified approach (Heiler, 1966). Recently Heiler and Michels (1994) (see also Heiler 1994, 1995) proposed a weighted version of BV based on the idea of locally weighted regression (LWR) introduced by Stone (1977) and Cleveland (1979). The main result of the present paper is a new data-driven version of this approach. A different procedure for time series decomposition with LWR is proposed by Cleveland et al. (1990), where seasonal fluctuations are treated in a different way and robustness weights are used.

Nadaraya and Watson introduced in 1964 the nonparametric kernel regression. Compared with kernel regression local regression has a very long tradition (see Cleveland and Loader, 1995 for a historical review and more references). LWR is also a kernel based method and can be interpreted as another type of kernel regression. Lejeune (1985) and Müller (1987) discussed the asymptotic equivalence between LWR and kernel regression. Based on these results it is possible to adapt some results for kernel regression to LWR. For finite samples LWR works better than kernel regression with its intuitional definition and model unbiasedness (see the next section). For details of the advantages of LWR compared with kernel regression see Hastie and Loader (1993) and Cleveland and Loader (1995). Recent researches show that LWR is a very attractive smoothing technique. We think that it is

also a powerful method for nonparametric estimation of many other econometric functionals discussed in Ullah (1988, 1989).

Some parameters, such as the bandwidths ( $h$ ), orders of the polynomials ( $p$ ) and the selected kernel function, which control the smoothness of the estimate have to be selected for decomposing time series with LWR. A data-driven method for selecting these parameters, especially for selecting the bandwidth and the order of the polynomials is very important in order to get an objective *best* estimation. Quite a few data-driven selection procedures for the bandwidth are proposed for kernel regression (see for instance Rice, 1983, 1984, Gasser et al., 1991, Härdle et al., 1992). Cleveland and Devlin (1988) and Cleveland et al. (1988) proposed a  $\tilde{M}$ -plot procedure for selecting bandwidth and the order of the polynomials for LWR. Fan and Gijbels (1992) and Ruppert et al. (1994) proposed some plug-in procedures for bandwidth selection for LWR based on asymptotic considerations. All of these proposals are for LWR with polynomials only. In this paper the bandwidth selection for time series decomposition with LWR is discussed. The R-statistic of Rice (1983, 1984) and the Double-Smoothing procedure of Härdle et al. (1992) (but without subjective pilot bandwidth) are adapted to this problem.

A very important problem is the estimation of the course of the components at the actual boundary (right end points) of the time series. We treat this problem in a natural way by selecting a left bandwidth  $h_l$  and a right bandwidth  $h_r$  pointwise for a moving average with given symmetrical ( $h_l = h_r$ ) or asymmetrical ( $h_l > h_r$ ) weights for each point in the right boundary area which are then used symmetrically for the left boundary area. The optimal order of polynomial is also selected pointwise by comparing the criteria for different  $p$ . In this paper we do not discuss the selection of the weight function (kernel). See Fedorov, Hackl and Müller (1993a, b) for researches in this direction.

In the next section we introduce the method of time series decomposition with LWR and discuss the finite sample properties of the estimates. Section 3 discusses the criteria for bandwidth selection and the estimation of variance. The data-driven procedures are described in section 4 while some real examples decomposed with these procedures are presented in section 5. Section 6 contains some concluding remarks.

## 2 Time Series Decomposition with LWR

We consider an equidistant time series  $\mathbf{Y}_t$ ,  $t \in \mathbf{I}$ , where  $\mathbf{I} = \{1, 2, \dots, T\}$  is the time index set. It is assumed that  $\mathbf{Y}_t$  follows an additive components model

$$Y_t = G(t) + S(t) + \epsilon_t, \quad t = 1, 2, \dots, T, \quad (1)$$

where  $\epsilon_t$  are assumed to be i.i.d. random variables with  $E(\epsilon_t) = 0$  and  $var(\epsilon_t) = \sigma^2$ .  $G(t)$  is the trend-cyclical component,  $S(t)$  is the seasonal component and  $\mu(t) = G(t) + S(t)$  is the nonrandom mean function. The proposed weighted version of BV to estimate  $\mu(t)$ ,  $G(t)$  and  $S(t)$  is described here briefly, for more details see Heiler (1994, 1995) and Heiler and Michels (1994).

The trend-cyclical component is assumed to be a smooth function with existing derivative of order  $p + 1$ , so that around a point  $t_0$  it can be developed in a Taylor series, yielding a local polynomial representation of order  $p$ . In a similar way it is assumed that the seasonal component with seasonal period  $s$  can be locally modeled by a Fourier series, i.e.

$$G(t) \doteq \sum_{j=0}^p \beta_{1j}(t_0)(t - t_0)^j, \quad S(t) \doteq \sum_{j=1}^q [\beta_{2j}(t_0) \cos \lambda_j(t - t_0) + \beta_{3j}(t_0) \sin \lambda_j(t - t_0)], \quad (2)$$

where  $q \leq [s/2]$  ( $[x]$  denotes the greatest integer less than or equal to  $x$ ),  $\lambda_1 = 2\pi/s$  and  $\lambda_j = j\lambda_1$  for  $j = 2, \dots, q$ . When  $\lambda_q = \pi$ , the last sinusoidal has to be omitted. With the vectors

$$\beta_1(t_0) = (\beta_{10}(t_0), \dots, \beta_{1p}(t_0))', \quad \beta_2(t_0) = (\beta_{21}(t_0), \beta_{31}(t_0), \dots, \beta_{2q}(t_0), [\beta_{3q}(t_0)])',$$

$$\beta(t_0) = (\beta_1(t_0)', \beta_2(t_0)')', \quad \mathbf{x}_1(t) = (1, (t - t_0), \dots, (t - t_0)^p)',$$

$\mathbf{x}_2(t) = (\cos \lambda_1(t - t_0), \sin \lambda_1(t - t_0), \dots, \cos \lambda_q(t - t_0), [\sin \lambda_q(t - t_0)])'$ ,  $\mathbf{x}(t) = (\mathbf{x}_1(t)', \mathbf{x}_2(t)')'$  (for  $\lambda_q = \pi$  the  $[\dots]$ -terms have to be omitted) (2) can be written in the form

$$G(t) \doteq \mathbf{x}_1(t)' \beta_1(t_0), \quad S(t) \doteq \mathbf{x}_2(t)' \beta_2(t_0) \quad \text{and} \quad \mu(t) \doteq \mathbf{x}(t)' \beta(t_0).$$

For  $t \in \mathbf{I}$ , the local regression parameters  $\beta_1(t_0)$  and  $\beta_2(t_0)$  are estimated by the locally weighted least squares criterion

$$Q(t_0) = \sum_{t=1}^T (Y_t - \mathbf{x}_1(t)' \beta_1(t_0) - \mathbf{x}_2(t)' \beta_2(t_0))^2 K\left(\frac{t - t_0}{h + 0.5}\right) \Rightarrow \min,$$

where  $K$  is a symmetrical density function (more precisely a kernel function of order 2) with compact support  $[-1, 1]$ . In order to cope with boundary problems we distinguish a left bandwidth  $h_l$  and a right bandwidth  $h_r$  and put  $h = \max(h_l, h_r)$ . With the  $T \times l$ -regressor matrix  $\mathbf{X} = (\mathbf{X}_1 : \mathbf{X}_2)$  with the rows  $x(t)'$ , where  $l = p + 2q + [1] < h_l + h_r + 1$ , as well as the weight matrix  $\mathbf{K}(t_0) = (k_{ij}(t_0)) = \text{diag}(k_{ii}(t_0))$  with

$$k_{ii}(t_0) = \begin{cases} K\left(\frac{i - t_0}{h + 0.5}\right) & , \quad i \in [t_0 - h_l, t_0 + h_r], \\ 0 & , \quad \text{otherwise,} \end{cases}$$

and the data vector  $\mathbf{y} = (y_1, \dots, y_T)'$  one immediately obtains the solution

$$\begin{aligned} \hat{\beta}(t_0) &= \begin{pmatrix} \hat{\beta}_1(t_0) \\ \hat{\beta}_2(t_0) \end{pmatrix} = (\mathbf{X}' \mathbf{K}(t_0) \mathbf{X})^{-1} \mathbf{X}' \mathbf{K}(t_0) \mathbf{y} \\ &= \begin{pmatrix} \mathbf{X}_1' \mathbf{K}(t_0) \mathbf{X}_1 & \mathbf{X}_1' \mathbf{K}(t_0) \mathbf{X}_2 \\ \mathbf{X}_2' \mathbf{K}(t_0) \mathbf{X}_1 & \mathbf{X}_2' \mathbf{K}(t_0) \mathbf{X}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{pmatrix} \mathbf{K}(t_0) \mathbf{y} \end{aligned} \quad (3)$$

and

$$\begin{aligned} \hat{\mu}(t) &= \mathbf{x}(t)' \hat{\beta}(t_0) = \mathbf{x}(t)' (\mathbf{X}' \mathbf{K}(t_0) \mathbf{X})^{-1} \mathbf{X}' \mathbf{K}(t_0) \mathbf{y} \\ &=: \mathbf{w}(t)' \mathbf{y} = \mathbf{x}_1(t)' \hat{\beta}_1(t_0) + \mathbf{x}_2(t)' \hat{\beta}_2(t_0) \\ &= (\mathbf{x}_1(t)', \mathbf{0}') (\mathbf{X}' \mathbf{K}(t_0) \mathbf{X})^{-1} \mathbf{X}' \mathbf{K}(t_0) \mathbf{y} + (\mathbf{0}', \mathbf{x}_2(t)') (\mathbf{X}' \mathbf{K}(t_0) \mathbf{X})^{-1} \mathbf{X}' \mathbf{K}(t_0) \mathbf{y} \\ &=: \mathbf{w}_1(t)' \mathbf{y} + \mathbf{w}_2(t)' \mathbf{y}, \end{aligned} \quad (4)$$

where  $\mathbf{0}$  denotes a vector of zeros of appropriate dimension. The final LWR estimations at the point  $t = t_0$  are

$$\hat{\mu}(t) = \mathbf{w}(t)' \mathbf{y}, \quad \hat{G}(t) = \mathbf{w}_1(t)' \mathbf{y} \quad \text{and} \quad \hat{S}(t) = \mathbf{w}_2(t)' \mathbf{y}, \quad (5)$$

where  $\mathbf{w}(t)$ ,  $\mathbf{w}_1(t)$  and  $\mathbf{w}_2(t)$  are defined in (4) with  $t = t_0$ .

For  $t \in [h_l + 1, T - h_r]$  all elements of  $\mathbf{w}(t)$ ,  $\mathbf{w}_1(t)$  and  $\mathbf{w}_2(t)$  with indices  $i \notin [t - h_l, t + h_r]$  are zeros. The non-zero part of  $\mathbf{K}(t)$ ,  $\tilde{\mathbf{K}} = (k_{ij}(t))$ ,  $i, j \in [t - h_l, t + h_r]$ , remains the same

and it can be shown that this is also true for the non-zero parts of the vectors  $\mathbf{w}(t)$ ,  $\mathbf{w}_1(t)$  and  $\mathbf{w}_2(t)$  which for  $t \rightarrow t + 1$  are shifted by one instant. Hence in the central part of the time series with  $h_l$  and  $h_r$  fixed the estimations work as moving averages with symmetrical weight systems (the non-zero parts of these vectors) for  $h_l = h_r$  or asymmetrical weight systems for  $h_l \neq h_r$ . To estimate  $\mu(t)$ ,  $G(t)$  and  $S(t)$  in the right boundary area one has to take  $h_r \leq T - t$ . In order to estimate  $\mu(t)$ ,  $G(t)$  and  $S(t)$  in the left boundary area one has to take  $h_l \leq t - 1$ .

With the notation  $\mathbf{x}_1^{(\nu)}(t) = (0, t^{(\nu)}, \dots, t^{p(\nu)})'$  ( $1 \leq \nu \leq p$ ) for the  $\nu$ -th derivative of  $\mathbf{x}_1(t)$  we obtain the estimate of the  $\nu$ -th derivative of the trend-cyclical component

$$\hat{G}^{(\nu)}(t) = \mathbf{w}^\nu(t)' \mathbf{y}$$

where

$$\mathbf{w}^\nu(t)' = (\mathbf{x}_1^{(\nu)}(t)', \mathbf{0}')(\mathbf{X}'\mathbf{K}(t)\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}(t). \quad (6)$$

$\hat{G}^{(\nu)}(t)$  works as a moving average, too.

In what follows we give some finite sample properties of the estimators discussed above. For the weight systems (the active kernel functions following the terminology of Hastie and Loader, 1993)  $\mathbf{w}(t)$ ,  $\mathbf{w}_1(t)$ ,  $\mathbf{w}_2(t)$  and  $\mathbf{w}^\nu(t)$  ( $1 \leq \nu \leq p$ ) with any  $p$ ,  $h_l$  and  $h_r$  and any kernel function it holds that

$$\sum_{i=1}^T w_i(t) = 1, \quad \sum_{i=1}^T w_{1i}(t) = 1, \quad \sum_{i=1}^T w_{2i}(t) = 0 \quad \text{and} \quad \sum_{i=1}^T w_i^\nu(t) = 0, \quad 1 \leq \nu \leq p.$$

This is one of the so called moment conditions. The above weight systems satisfy yet more general moment conditions.

Under the assumption of model (1) the variance factors (variance/ $\sigma^2$ ) of  $\hat{\mu}(t)$ ,  $\hat{G}(t)$ ,  $\hat{S}(t)$  and  $\hat{G}^{(\nu)}(t)$  are

$$\mathbf{w}(t)'\mathbf{w}(t), \quad \mathbf{w}_1(t)'\mathbf{w}_1(t), \quad \mathbf{w}_2(t)'\mathbf{w}_2(t) \quad \text{and} \quad \mathbf{w}^\nu(t)'\mathbf{w}^\nu(t),$$

respectively. If one uses a k-nearest neighbour method, i.e. the total bandwidth,  $h_T = h_l + h_r + 1$ , is fixed, the variance factor of  $\hat{G}(t)$  is strongly growing in the boundary area towards the ends of the time series. The variance factor of  $\hat{S}(t)$  remains almost the same in the boundary area and in the interior. In the interior it is about  $(s-1)$  times that of a kernel smoothing with  $K(u)$ . If  $s$  is large, for example  $s = 12$ , the variance factor of  $\hat{S}(t)$  is much larger than the variance factor of  $\hat{G}(t)$  with  $p \leq 3$ . In this case the variance of  $\hat{S}(t)$  is the dominant term of the variance of  $\hat{\mu}(t)$ .

The correlation coefficients between single estimations  $\hat{G}(t)$  and  $\hat{S}(t)$ ,  $\hat{G}(t)$  and  $\hat{G}^{(\nu)}(t)$  and  $\hat{S}(t)$  and  $\hat{G}^{(\nu)}(t)$  are

$$\rho_{gs}(t) = \frac{\mathbf{w}_1(t)'\mathbf{w}_2(t)}{\sqrt{(\mathbf{w}_1(t)'\mathbf{w}_1(t))(\mathbf{w}_2(t)'\mathbf{w}_2(t))}}, \quad \rho_{gg^\nu}(t) = \frac{\mathbf{w}_1(t)'\mathbf{w}^\nu(t)}{\sqrt{(\mathbf{w}_1(t)'\mathbf{w}_1(t))(\mathbf{w}^\nu(t)'\mathbf{w}^\nu(t))}}$$

and

$$\rho_{sg^\nu}(t) = \frac{\mathbf{w}_2(t)'\mathbf{w}^\nu(t)}{\sqrt{(\mathbf{w}_2(t)'\mathbf{w}_2(t))(\mathbf{w}^\nu(t)'\mathbf{w}^\nu(t))}},$$

respectively. These correlation coefficients are about zero if  $h_l$  or  $h_r$  is large, especially when  $h_l = h_r$ . It can be shown that, under the assumptions of model (1),  $\rho_{gs} \rightarrow 0$  as  $h_l \rightarrow \infty$  or  $h_r \rightarrow \infty$ . At an endpoint,  $t = 1$  (or  $t = T$ ),  $\rho_{gg^\nu}(t)$  and  $\rho_{gs}(t)$  are always negative for finite  $h_r$  (or  $h_l$ ).

LWR estimations are **model unbiased**, i.e.  $\hat{G}(t)$  is unbiased for a polynomial trend-cyclical component of order not larger than  $p$ , and  $\hat{G}^{(\nu)}(t)$  is unbiased for its  $\nu$ -th ( $\nu \leq p$ ) derivative.  $\hat{S}(t)$  is unbiased for a constant seasonal component. If  $h_l = h_r$ ,  $\hat{G}(t)$  (or  $\hat{S}(t)$ ) with a polynomial of even order  $p$  and of order  $p+1$  are the same. If the seasonal component is constant,  $\hat{S}(t)$  is unbiased while the bias for  $\hat{\mu}(t)$  and  $\hat{G}(t)$  is the same. In this case one can obtain similar asymptotic results as for LWR with polynomials only, but this will not be discussed here.

Further  $\hat{\mu}(t)$ ,  $\hat{G}(t)$ ,  $\hat{G}^{(\nu)}(t)$  and  $\hat{S}(t)$  are optimal in the sense of weighted variances, i.e. the sums

$$\begin{aligned} \sum_{i=t-h_l}^{t+h_r} w_i(t)^2/k_i(t), & \quad \sum_{i=t-h_l}^{t+h_r} w_{1i}(t)^2/k_i(t), \\ \sum_{i=t-h_l}^{t+h_r} w_{2i}(t)^2/k_i(t) & \quad \text{and} \quad \sum_{i=t-h_l}^{t+h_r} w_i^\nu(t)^2/k_i(t) \end{aligned}$$

are minimal in the class of all linear model unbiased local smoothers. Proofs of these properties are omitted. The proof of the properties of  $\mathbf{w}(t)$  for LWR without seasonal can be found in Müller (1987) and Lejeune (1985).

### 3 Estimations of the MASE and $\sigma^2$

As selection criterion for the parameters we use the mean averaged squared error (MASE) as a distance between  $\hat{\mu}(t)$  and  $\mu(t)$ ,

$$M = M(h_l, h_r) := T_{I^*}^{-1} \sum_{t \in I^*} E[\hat{\mu}(t) - \mu(t)]^2,$$

where  $I^*$  is a subset of the time index set  $I$  on which the smoothing with  $h_l$  and  $h_r$  is possible and  $T_{I^*}$  is the number of observations in  $I^*$ . Because  $h_l/T \rightarrow 0$  and  $h_r/T \rightarrow 0$  as  $T \rightarrow \infty$  it holds that  $T_{I^*}/T \rightarrow 1$  as  $T \rightarrow \infty$ . It is well known that the MASE splits up into a variance part and a bias part. The variance part of  $M(h_l, h_r)$  is given by

$$V = V(h_l, h_r) = T_{I^*}^{-1} \sum_{t \in I^*} \text{var}[\hat{\mu}(t)] = \sigma^2 T_{I^*}^{-1} \sum_{t \in I^*} \sum_{i=1}^T w_i(t)^2.$$

At each point the bias is defined by

$$b(t) = E[\hat{\mu}(t) - \mu(t)].$$

The systematic bias part is  $B = B(h_l, h_r) = T_{I^*}^{-1} \sum_{t \in I^*} b(t)^2$ . To estimate  $B(h_l, h_r)$  we define a general estimator of  $b(t)$  by

$$\hat{b}(t) = \mathbf{A}'(t)\mathbf{y}, \tag{7}$$

where  $\mathbf{A}(t) = \{a_1, a_2, \dots, a_T\}'$  is a weight vector with  $\sum_{i=1}^T a_i = 0$ . Then we estimate  $B(h_l, h_r)$  by

$$\hat{B} = \hat{B}(h_l, h_r) = T_{I^*}^{-1} \sum_{t \in I^*} \hat{b}(t)^2.$$

It is easy to show that there is a variance term in this estimate,  $T_{I^*}^{-1} \sigma^2 \sum_{t \in I^*} \sum_{i=1}^T a_i(t)^2$ , that should be allowed for. Therefore by subtracting this variance term, we estimate  $B$  by

$$\hat{B}_1 = \hat{B} - T_{I^*}^{-1} \hat{\sigma}^2 \sum_{t \in I^*} \sum_{i=1}^T a_i(t)^2,$$

where  $\hat{\sigma}^2$  is an estimator of  $\sigma^2$ . The variance part  $V$  is estimated by

$$\hat{V} = \hat{\sigma}^2 T_{I^*}^{-1} \sum_{t \in I^*} \sum_{i=1}^T w_i(t)^2.$$

Hence the final estimate of  $M$  is

$$\hat{M}(h_l, h_r) = \hat{V} + \hat{B} - T_{I^*}^{-1} \hat{\sigma}^2 \sum_{t \in I^*} \sum_{i=1}^T a_i(t)^2 = \hat{V} + \hat{B}_1. \quad (8)$$

In practice it can occur that  $\hat{B}_1$  becomes negative. Hence it seems more reasonable if one uses a slightly different estimator  $\tilde{M}(h_l, h_r)$  instead of the estimator  $\hat{M}(h_l, h_r)$ , where  $\tilde{M}(h_l, h_r) = \hat{M}(h_l, h_r)$  if  $\hat{B}_1 \geq 0$  and  $\tilde{M}(h_l, h_r) = \hat{V}$  if  $\hat{B}_1 < 0$ .

Two special cases of (8) are the R-statistic procedure proposed by Rice (1983, 1984) and the Double-Smoothing procedure proposed by Härdle et al. (1992) for the kernel regression.

In the R-statistic the bias  $b(t)$  at each point  $t$  is estimated by

$$\hat{b}_R(t) = \hat{\mu}(t) - y_t = \sum_{i=1}^T a_{iR}(t) y_i,$$

where

$$a_{iR}(t) = \begin{cases} w_i(t), & \text{for } i \neq t, \\ w_i(t) - 1, & \text{for } i = t, \end{cases}$$

and

$$\hat{B}_R = T_{I^*}^{-1} \sum_{t \in I^*} (\hat{\mu}(t) - y_t)^2.$$

Note that for fixed  $h_l$  and  $h_r$  the estimate  $\hat{\mu}(t)$  is a moving average. We obtain the R-statistic from (8):

$$\hat{R}(h_l, h_r) = T_{I^*}^{-1} \sum_{t \in I^*} (\hat{\mu}(t) - y_t)^2 + (2w_{t^*}(t^*) - 1) \hat{\sigma}^2, \quad (9)$$

where  $t^*$  can be any observation time point in  $I^*$ .  $\hat{R}(h_l, h_r)$  is an unbiased estimator of  $M(h_l, h_r)$ . The optimal bandwidths following this criterion are defined as the pair  $\hat{H}_R = \{\hat{h}_l, \hat{h}_r\}$  which minimizes (9). Discussions about the asymptotic properties of such an estimator for kernel regression can be found in Rice (1984) and Härdle et al. (1988).

Following the Double-Smoothing idea of Härdle et al. (1992) the bias  $b(t)$  at the point  $t \in \mathbf{I}^*$  is estimated by means of a pilot smoothing  $\hat{\mu}_g(t)$  with given pilot bandwidths  $g = \{g_l, g_r\}$  and weights  $w_{ig}(t)$  by

$$\hat{b}_D(t) = \sum_{j=1}^T w_j(t) \hat{\mu}_g(j) - \hat{\mu}_g(t) = \sum_{j=1}^T a_{jR} \hat{\mu}_g(j) = \sum_{i=1}^T a_{iD} y_i,$$

where the  $w_j(t)$  are the weights for a smoothing with bandwidths  $h_l$  and  $h_r$  and

$$a_{iD}(t) = \sum_{j=1}^T w_j(t) w_{ig}(j) - w_{ig}(t).$$

The systematic bias term is estimated by

$$\hat{B}_D = T_{I^*}^{-1} \sum_{t \in I^*} \hat{b}_D(t)^2.$$

The term  $T_{I^*}^{-1} \hat{\sigma}^2 \sum_{t \in I^*} \sum_{i=1}^T a_{iD}(t)^2$  causes an additional bias of the estimation of the bandwidths. It is more difficult to calculate than  $\hat{B}_D$  itself. The calculation of each  $a_{iD}(t)$  involves the convolution of two weight systems,  $w_j(t)$  and  $w_{ig}(j)$ . But in order to obtain  $\hat{B}_D$  we do not actually need to calculate  $a_{iD}(t)$  (see the formula of  $\hat{b}_D(t)$ ). The larger the pilot bandwidths compared to the selected bandwidths the smaller is the influence caused by this term. This term plays a similar role as the so called *nonstochastic term* in Sheather and Jones (1991) and Jones et al. (1991). One can use the idea proposed there, i.e. leave this term in  $\hat{M}(h_l, h_r)$ . We write these two cases (with or without this term) in one formula by introducing an indicator variable  $\Delta$

$$\hat{M}_D(h_l, h_r) = \hat{V} + \hat{B}_D - (1 - \Delta) T_{I^*}^{-1} \hat{\sigma}^2 \sum_{t \in I^*} \sum_{i=1}^T a_{iD}(t)^2, \quad (10)$$

where  $\Delta = 0$  or 1. If  $\Delta = 0$  (10) is the same as in Härdle et al. (1992), if  $\Delta = 1$  (10) is similar to the proposal of Müller (1985). The computational advantage with  $\Delta = 1$  is clear while the calculation of (10) for  $\Delta = 1$  is as easy as the calculation of (9) if the  $\hat{\mu}_g(i)$  are known. The rate of convergence of the bandwidth selector with  $\Delta = 1$  is sometimes slightly slower than that with  $\Delta = 0$ . But with a properly selected pilot bandwidth one can get a faster rate of convergence with  $\Delta = 1$ . See Heiler and Feng (1995) for more general cases.

For small pilot bandwidths one should take  $\Delta = 0$ . In the special case  $g_l = g_r = 0$ , i.e. no smoothing in the pilot step, we have  $\hat{M}_D(h_l, h_r) = \hat{R}(h_l, h_r)$ . If  $g_l > 0$  or  $g_r > 0$   $\hat{M}_D(h_l, h_r)$  is a biased estimator of  $M$  with a smaller variance than that in  $\hat{R}(h_l, h_r)$ . The optimal bandwidths  $\hat{H}_D = \{\hat{h}_l, \hat{h}_r\}$  following this criterion are the minimizers of (10). The rate of convergence of the Double-Smoothing bandwidth selector depends on the choice of the pilot bandwidth and the orders of the so called equivalent kernels in the pilot smoothing and the main smoothing (see section 4.2, Härdle et al., 1992 and Heiler and Feng, 1995). Sometimes it can even achieve the best convergence rate,  $T^{-\frac{1}{2}}$ . The convergence rate of  $\hat{h}_R$  is as slow as  $T^{-1/10}$ . A disadvantage of the Double-Smoothing method is that it requires a pilot smoothing or pilot bandwidths  $g_l$  and  $g_r$ . In the next section a Double-Smoothing

procedure with good asymptotic properties is proposed which does not require a subjective pilot smoothing. The pilot smoothing is obtained by using a R-statistic procedure.

The selectors of optimal bandwidths discussed above are all defined for a general linear smoother and can be used for selecting bandwidths of time series decomposition with LWR (In fact it is easy to show that the M-plot procedure proposed by Cleveland and Devlin (1988) and Cleveland et al. (1988) for LWR is just the same as the R-statistic procedure.) If we have pilot estimations of  $G(t)$ ,  $S(t)$  and  $G^{(\nu)}(t)$  we can also select bandwidths for these components with a Double-Smoothing procedure, respectively. Though other CV-like bandwidth selection procedures described in Rice (1983, 1984) and Härdle et al. (1988) are shown to be asymptotically equivalent to the R-statistic procedure, most of these criteria should not be used for time series decomposition with LWR, because some of them achieve a trivial minimum at the "no smoothing" point, where  $\hat{\mu}(t) = y_t$ , while some others become negative for small bandwidths.

To estimate  $\hat{M}(h_l, h_r)$  in (9) or (10) one also needs a nonparametric estimator of the unknown variance  $\sigma^2$ . Rice (1984) proposed an estimator based on a local constancy assumption. This estimator is

$$\hat{\sigma}_1^2(I^*) = \frac{1}{2(T_{I^*} - 1)} \sum_{t \in I^*, t+1 \in I^*} (y_{t+1} - y_t)^2. \quad (11)$$

Further Gasser et al. (1986) proposed another estimator of the variance based on a local linearity assumption which is given by

$$\hat{\sigma}_2(I^*) = \frac{2}{3(T_{I^*} - 2)} \sum_{t, t+1, t+2 \in I^*} (y_{t+1} - \frac{1}{2}(y_t + y_{t+2}))^2. \quad (12)$$

This idea has been extended to a general definition of a so called *difference-based* estimation of variance in nonparametric regression with a given *difference-sequence* by Hall et al. (1990). A difference sequence of order  $m$  is a real sequence  $D_m = \{d_0, d_1, \dots, d_m\}$  such that

$$\sum_{j=0}^m d_j = 0 \quad \text{and} \quad \sum_{j=0}^m d_j^2 = 1, \quad m = 1, 2, \dots \quad (13)$$

The estimation of the variance based on  $D_m$  is

$$\hat{\sigma}_m^2 = (T_{I^*} - m)^{-1} \sum_{t, \dots, t+m \in I^*} \left( \sum_{j=0}^m d_j y_{j+t} \right)^2. \quad (14)$$

For any  $m$   $\hat{\sigma}_m^2(I^*)$  is always a root-T consistent estimator of  $\sigma^2$  (Hall et al., 1990). The estimators in (11) and (12) are two special cases of  $\hat{\sigma}_m^2(I^*)$  with  $m=1$  and  $m=2$ , respectively.

In order to estimate the variance of a time series with seasonal component the seasonal fluctuations must be allowed for as well as the trend-cyclical movements. Therefore we define a *seasonal-difference-sequence*  $D_{ms} = \{d_j\}$ ,  $j = 0, 1, \dots, m$  which in addition to (13) also satisfies the condition

$$S_i = \sum_{j=0}^m d_j \delta_{ij} = 0, \quad i = 0, 1, \dots, s-1,$$

where  $s$  is the seasonal period and

$$\delta_{ij} = \begin{cases} 1, & \text{if } (j-i)/s \text{ is an integer,} \\ 0, & \text{otherwise.} \end{cases}$$

For  $s = 12$  two simple examples of seasonal-difference-sequences are

$$D_{ms}^1 = \frac{1}{c_1} \left\{ \begin{array}{c} -1, 2, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, \\ 1, -2, 1 \end{array} \right\}$$

and

$$D_{ms}^2 = \frac{1}{c_2} \left\{ \begin{array}{c} -1, 2, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, \\ 2, -4, 2, 0, 0, 0, 0, 0, 0, 0, 0, \\ -1, 2, -1 \end{array} \right\},$$

where  $c_1$  and  $c_2$  are two constants defined by (13). With  $D_{ms}^1$  (or  $D_{ms}^2$ ) we can obtain a root- $T$  consistent estimation of  $\sigma^2$  under the assumptions of model (1) and the assumption that the seasonal is locally constant. If for finite samples  $S(t) \doteq \frac{1}{2}(S(t-s) + S(t+s))$  holds,  $t = s+1, \dots, T-s$ , then  $D_{ms}^2$  is better. Other proposals for nonparametric variance estimation in LWR based on residuals can be found in Ruppert et al. (1994) and Fan and Gijbels (1992, 1995).

## 4 The proposed Data-driven Procedures

In this section two data-driven procedures for selecting optimal bandwidths and order of polynomials are proposed. The first one is based on the R-statistic with a special solution of the boundary problem. The second one consists of two stages. At the first stage one gets pilot estimation of the optimal bandwidths based on the R-statistic with a polynomial of a fixed (higher) order and the pilot estimations  $\hat{\mu}_g(t)$ ,  $\hat{G}_g(t)$ ,  $\hat{S}_g(t)$  and  $\hat{G}_g^{(\nu)}(t)$ . At the second stage one uses the pilot estimations in the Double-Smoothing criterion (10) in order to select the optimal bandwidths and proper orders of the polynomials for the trend-cyclical component and the seasonal as well as a second estimation of bandwidths and orders of the polynomials for  $\hat{\mu}(t)$ . One can estimate the optimal bandwidths for  $\hat{G}^{(\nu)}(t)$ , too. The algorithm for the second procedure is the same as for the first one but with  $\hat{M}_D(h_l, h_r)$  instead of  $\hat{R}(h_l, h_r)$ . The Mean Averaged Squared Errors of  $\hat{G}(t)$ ,  $\hat{S}(t)$  and  $\hat{G}^{(\nu)}(t)$  are estimated analogously as in (10). Hence we describe only the algorithm for the first procedure in detail.

### 4.1 Data-driven procedure based on R-statistic

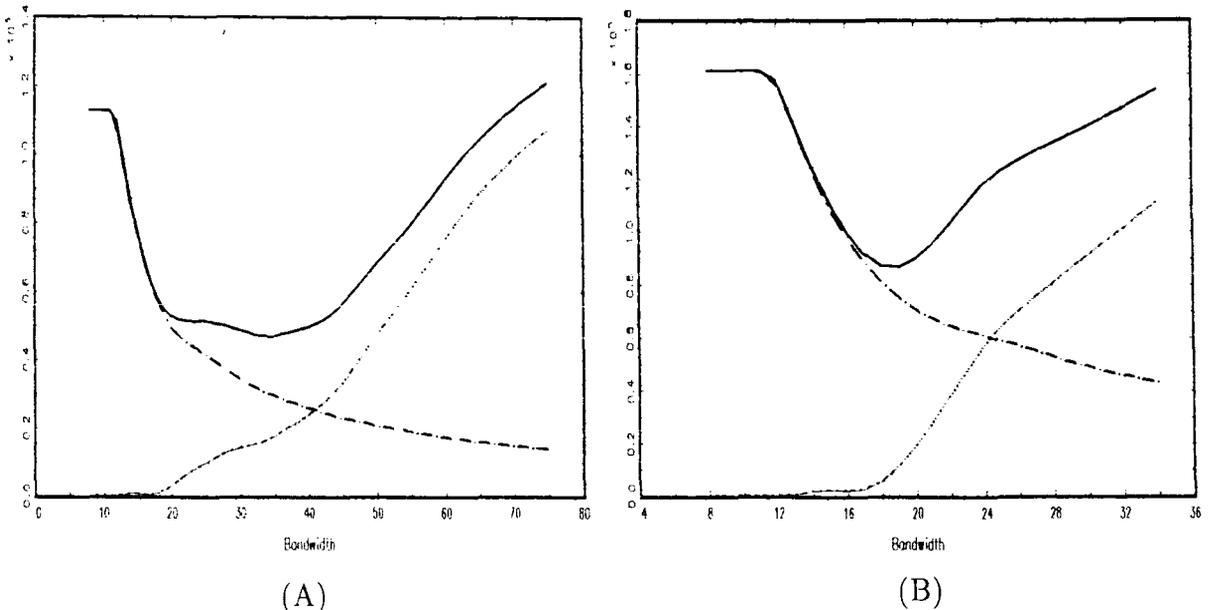
The main results of the proposed procedure are a bandwidth  $\hat{h}_m$  with  $\hat{p}$  for the estimations in the central part  $t \in [\hat{h}_m + 1, T - \hat{h}_m]$  (defined after the selection of  $\hat{h}_m$ ) with  $\hat{h}_l = \hat{h}_r = \hat{h}_m$  and bandwidths  $\{\hat{h}_l(t), \hat{h}_r(t)\}$  for the estimation at a point  $t$  in the right boundary area,  $t = T - \hat{h}_m, T - \hat{h}_m + 1, \dots, T$ , with proper order of the polynomial,  $\hat{p}(t)$ . The estimation at a point  $t$  in the left boundary area is effectuated symmetrically with the bandwidths  $\{\hat{h}_l(T - t + 1), \hat{h}_r(T - t + 1)\}$  and a polynomial of order  $\hat{p}(T - t + 1)$ . All bandwidths are selected following a similar iterative procedure.

At the beginning one has to put three parameters: the type of the difference sequence, the kernel function and the maximal order of the polynomials. As alternatives for the difference

sequences the procedures provide (11) and (12) for  $s=1$  (without seasonality) and  $D_{m_s}^1$  and  $D_{m_s}^2$  for  $s \geq 3$ . Which one of them should be used depends on the data on hand. For example, if it is assumed that the trend-cyclical is locally linear and the seasonal is locally constant, then one should choose  $D_{m_s}^1$  (Note that this does not mean that one has to take  $p = 1$ ). The choice of kernel function plays a more important role for time series decomposition with LWR than for kernel regression but we do not discuss this question here. In the procedures four kernel functions of the type  $C_\mu(1 - u^2)^\mu$  are provided as alternatives for  $\mu = 0, 1, 2, 3$ . They are the Rectangular Kernel, the Epanechnikov Kernel, the Bisquare Kernel and the Triweight Kernel, respectively. In the examples in this paper the difference sequence  $D_{m_s}^1$  and the Bisquare Kernel ( $\mu = 2$ ) are used. When the variance  $\sigma^2$  compared with the bias term is very small, a polynomial of high order ( $p > 3$ ) should be better following the criteria in section 3. Therefore a maximal order of polynomials,  $p_m$ , is given beforehand, for example  $p_m = 3$ , and only orders of the polynomials which are not larger than  $p_m$  are considered.

#### 4.1.1 Selection of $h_m$ and $p$ in the central part:

In this step  $h_l = h_r = h$  and the weight systems are symmetrical. The basic idea is to search an optimal bandwidth  $h_m(p)$  for a given  $p$  in an interval  $h \in [h_{min}, h_{max}]$  on a support  $t \in \mathbf{I}^*$ , where  $h_{min}$  is close to the possible smallest bandwidth. In the proposed procedures we use  $h_{min} = [\frac{1}{2}(s + p)] + 1$ , for instance for  $s = 12$  and  $p = 2$  we have  $h_{min} = 8$ .  $h_{max}$  is a given maximal bandwidth. In the proposed procedures  $h_{max} = [T/4]$  is used at the beginning with which the used observations for LWR is about one half of all observations. The support  $\mathbf{I}^* = [h_{max} + 1, T - h_{max}]$  is used to search an optimal bandwidth for given  $h_{max}$ . This is the largest support on which the smoothing with the bandwidths  $h_l = h_r = h_{max}$  is possible. It is clear that the selected optimal bandwidth depends on the support  $\mathbf{I}^*$ . This is shown in figure 1. Figure 1 exhibits  $\hat{R}(h)$  (solid line) as well as  $\hat{V}$  (dots and dashes) and  $\hat{B}_1$  (dots)



**Figure 1:** The estimated  $\hat{R}(t)$  (solid line),  $\hat{V}(t)$  (dots and dashes) and  $\hat{B}_1(t)$  (dots) for a data set STROM ( $T = 300$ ) with two different  $h_{max}$ : (a)  $h_{max} = 75$  and (b)  $h_{max} = 23$ .

estimated by the R-statistic for a data set STROM ( $T = 300$ , Schlittgen and Streitberg, 1991, page 65). Figure 1(a) shows these estimations with  $p = 2$ ,  $h_{min} = 8$ ,  $h_{max} = 75$

and the support  $\mathbf{I}^* = [76, 225]$ . There are two local optima in figure 1(a):  $h_{m1} = 23$  and  $h_{m2} = 34$ .  $h_{m2} = 34$  is the optimal bandwidth on  $\mathbf{I}^* = [76, 225]$ . In figure 1(b) another maximal bandwidth  $h_{max} = 34$  (the optimal one in (a)) is used, now only one optimum at  $h_m = 19$  occurs. We see that  $h = 34$  could not be a global optimal bandwidth.

We define  $h_m(p)$  for a given  $p$  as a *stable global optimal bandwidth in the central part*, i.e.  $h_m$  is optimal on some supports  $\mathbf{I}^{*'} = [h'_m + 1, T - h'_m]$  with  $h'_m > h_m$  as well as on the support  $\mathbf{I}^* = [h_m + 1, T - h_m]$  itself. For the STROM data with  $p = 2$  the bandwidth  $h_m = 19$  is such an optimal bandwidth. In fact  $h_m = 19$  is optimal for all supports  $\mathbf{I}^{*'} = [h'_{max} + 1, T - h'_{max}]$  with  $19 \leq h'_{max} \leq 53$ .

In the following the proposed automatic iterative procedure to select  $\hat{h}_m$  and a proper  $\hat{p}$  is described. At the beginning we put  $p = p_m$  if  $p_m$  is even or  $p = p_m - 1$  if  $p_m$  is odd, because  $\hat{G}(t)$ ,  $\hat{S}(t)$  and  $\hat{\mu}(t)$  are the same for  $p = p_m$  and  $p = p_m - 1$  if  $p_m$  is odd. The value  $h_{max0} = \lceil \frac{1}{4}T \rceil$  is used at this stage.

1. Select an optimal bandwidth  $h_{m0}$  with given  $p$  and  $h_{max0}$  on the support  $\mathbf{I}_0^* = [h_{max0} + 1, T - h_{max0}]$ ;
2. Put  $h_{max1} := h_{m0} + 1 + \lceil \frac{1}{3}(h_{max0} - h_{m0}) \rceil$ . Then select a new optimal bandwidth  $h_{m1}$  on the support  $\mathbf{I}_1^* = [h_{max1} + 1, T - h_{max1}]$ . If  $h_{m1}$  is stable, then goto step 3, otherwise put  $h_{max0} := h_{max1}$ ,  $h_{m0} = h_{m1}$  and repeat this step;
3. Put  $h_{max2} := h_{m1} + 1$  and select a new optimal bandwidth  $h_{m2}$  on the support  $\mathbf{I}_2^* = [h_{max2} + 1, T - h_{max2}]$ . If  $h_{m2}$  is also stable, then goto step 4, otherwise put  $h_{m1} = h_{m2}$  and repeat this step;
4. Put  $h_{max3} := h_{m2}$  and select a new optimal bandwidth on the support  $\mathbf{I}_3^* = [h_{max3} + 1, T - h_{max3}]$ . If  $h_{m3}$  is stable, too, put  $h_m(p) = h_{m3}$  and goto step 5. If  $h_{m3}$  is not stable, put  $h_{m1} = h_{m3}$  and go back to step 3;
5. If  $p - 2 < 0$ , then goto step 6, otherwise put  $h_{max0} := \lceil 1.2h_m(p) \rceil$ ,  $p := p - 2$  and go back to step 1;
6. Select the smallest bandwidth (called  $\hat{h}_m$ ) from all  $h_m(p)$ .  $\hat{h}_m$ , the respective  $\hat{p}$  and  $\hat{R}(\hat{h}_m, \hat{p})$  are the outcomes of this procedure in the central part.

The definition of  $h_{max}$  in step 2 and 3 ensure that a larger optimal bandwidth can also occur in the next step. One can put different values for  $h_{max}$  in step 2, 3 or step 5. It might cause a difference in the computing time, but it does not cause any change in the results. For a given  $p$  the minimal bandwidth  $h_{min}$  remains the same in all steps. In all steps it must be  $h_{max} \ll \lceil T/2 \rceil$  such that there are enough observations in the support  $\mathbf{I}^*$ . In this procedure we put  $h_{max} \leq \lceil T/4 \rceil$ . Since the assumption of constant variance often does not hold exactly for real data sets, the variance on a support  $\mathbf{I}^* = [a, b]$  is estimated on a related support  $\mathbf{I}_v^* = [a', b']$  where  $b' = \min(T, b + \lceil \frac{m}{2} \rceil)$  and  $a' = \max(1, b' - m - b + a)$ , where  $m$  is the order of the difference-sequence. The details of this iterative procedure can partly be seen in figure 1 and the corresponding explanations.

Sometimes the iterative optimal bandwidth for a given  $p$  does not converge to a  $h_m(p)$  but pendulates between a  $h_m^p$  and  $h_m^p + 1$ . In this case the proposed procedure is an endless loop and if this happens either  $h_m^p$  or  $h_m^p + 1$  is taken automatically as  $h_m(p)$ .

### 4.1.2 Selection procedure in the right boundary part:

After selection of  $\hat{h}_m$  the treatment boundary parts are clear. The right boundary part is then given by the interval  $[T - \hat{h}_m + 1, T]$ . The considered bandwidths at a point  $t$  in the right boundary part are limited to pairs  $\{h_l, h_r\}$  with either  $h_l = h_r$  or  $h_l > h_r$  and  $h_r = (T - t)$ . For  $t < T$  a pair of bandwidths  $\{h_l, h_r\}$  with  $h_l \neq h_r$  and  $h_r < (T - t)$  is also possible, but we do not consider this situation here. Under this limitations the selection of a pair  $\{\hat{h}_l, \hat{h}_r\}$  is equivalent to the selection of  $\hat{h}_l$ .

Now the minimal value of  $h_l(p)$  is given by  $h_{min} = \max(\lfloor \frac{1}{2}(s+p) \rfloor + 1, s+p+1 - (T-t))$ ,  $h_{max}$  is given as the maximal value of  $h_l$  used at step one. The comparative support determined by  $h_{max}$  is now  $\mathbf{I}^* = [h_{max} + 1, t]$  if  $t > T - h_{max}$  or the same as in the central part if  $t \leq T - h_{max}$ . The procedure to select  $\hat{h}_l$  is the same as for  $\hat{h}_m$ . But since  $h_l \neq h_r$  is now frequent, one should use  $p = p_m$  at the beginning and  $p - 1$  at step 5 for the next iteration, respectively. The selection of  $h_{max0}$  at each point is also changed.

The time series is then decomposed with the selected bandwidths and orders of the polynomials. If the time series does not have seasonal variation, the seasonal period  $s$  takes the value 1. In this case the procedure works automatically as a usual optimal smoothing procedure of a time series with LWR.

## 4.2 The Data-driven Double-Smoothing procedure

The selection of the pilot bandwidth for the Double-Smoothing procedure is an unresolved problem in Härdle et al. (1992). Our proposal is to select pilot bandwidths using the procedure proposed above with a fixed pilot order of the polynomial  $p_p$ . Then one uses these bandwidths (or multiplied by a factor  $\alpha_\Delta(T)$ ) to get the pilot smoothing  $\hat{\mu}_g(t)$ . The second stage of the Double-Smoothing procedure is exactly the same as the procedure based on the R-statistic. But now the criterion  $\hat{M}(h_l, h_r)$  is calculated from (10). The bias at a point  $t$  is estimated by  $\hat{b}_D(t) = \sum_{i=1}^T a_{iR} \hat{\mu}_g(i)$  with the pilot estimation  $\hat{\mu}_g(i)$ .

In the sequel we explain why the pilot bandwidths can be selected with the R-statistic and discuss the rate of convergence of the selected bandwidths with the proposed Double-Smoothing procedure for time series without seasonality shortly. If we talk about the order of a bandwidth it always means a relative one i.e. the absolute one divided by  $T$ . We assume that one uses a fixed order of polynomial  $p$  at the second stage, too (say  $p_p$  and  $p$  are both odd, to simplify). In this case the equivalent kernel function is of order  $s = p_p + 1$  at the first stage and of order  $r = p + 1$  at the second (main) stage. If  $p_p$  (or  $p$ ) is even the equivalent kernel functions are of order  $s = p_p + 2$  and  $r = p + 2$  in the central part and of order  $s = p_p + 1$  and  $r = p + 1$  at the end point  $t = T$ .

Selecting a pilot bandwidth for the Double-Smoothing procedure with the R-statistic is a good starting point. It is well known that the selected bandwidth  $\hat{g}$  based on the R-statistic is already of order  $T^{-\frac{1}{2s+1}}$  and the optimal bandwidth  $h_0$  is of order  $T^{-\frac{1}{2r+1}}$ . When  $\Delta = 0$  one can obtain the asymptotically optimal order of the pilot bandwidth by keeping the two terms containing  $g$  in theorem 1 of Härdle et al. (1992) in balance. These two terms are of order  $(T^{-2}g^{-(4r+1)})^{1/2}$  and  $g^s$ , respectively. The asymptotically best choice of the pilot bandwidth  $g_{opt}$  is to take it of order  $T^{-\frac{2}{4r+2s+1}}$ . It holds that

$$g_{opt} = O\left(T^{-\frac{2}{4r+2s+1}}\right) = \hat{g}\alpha_0(T),$$

where  $\alpha_0(T) = O(T^{\frac{4r-(2s+1)}{(4r+2s+1)(2s+1)})}$ . It is easy to show that  $h_0/g_{opt} \rightarrow 0$  as  $T \rightarrow \infty$ . The case  $\Delta = 1$  is a special case of Heiler and Feng (1995) with  $\Delta = 1$  and  $\delta = 0$ . Following theorem 1 in Heiler and Feng (1995) the asymptotically best choice of the pilot bandwidth is

$$g_{opt} = O\left(T^{-\frac{1}{2r+s+1}}\right) = \hat{g}\alpha_1(T),$$

where  $\alpha_1(T) = O(T^{\frac{2r-s}{(2r+s+1)(2s+1)})}$ .

Selecting the pilot bandwidth  $g_{opt} = \hat{g}\alpha_\Delta(T)$  is asymptotically optimal up to a very complicated constant term which does not affect the rate of convergence. In the central part with  $\Delta = 0$  the rate of convergence of the Double-Smoothing bandwidth selector with the pilot bandwidth  $g_{opt} = \hat{g}\alpha_0(T)$  is  $T^{-\frac{2s}{4r+2s+1}}$  if  $s \leq 2r$  or  $T^{-\frac{1}{2}}$  (a root-T estimator) if  $s \geq 2r + 1$ . For the latter a slightly weak condition is given in Härdle et al. (1992), remark 2. For example for  $p_p = 5$  and  $p = 3$ , i.e.  $s = 6$  and  $r = 4$ , the rate of convergence is  $T^{-\frac{12}{29}}$ . If  $\hat{g}$  is used simply as the pilot bandwidth we obtain the slower one of the two rates  $T^{-\frac{s}{2s+1}}$  and  $T^{-\frac{4(s-r)+1}{4s+2}}$ . This is slower than the rate of convergence by using  $g_{opt} = \hat{g}\alpha_0(T)$  and is always slower than  $T^{-\frac{1}{2}}$ . When  $s = 6$  and  $r = 4$  the rate of convergence is  $T^{-\frac{9}{26}}$ . If  $\Delta = 1$  the resulting rate of convergence with the pilot bandwidth  $g_{opt} = \hat{g}\alpha_1(T)$  is  $T^{-\frac{s}{2r+s+1}}$  which is a little slower than that with  $\Delta = 0$ . Again if  $\hat{g}$  is used simply as the pilot bandwidth we obtain the slower one of the two rates  $T^{-\frac{s}{2s+1}}$  and  $T^{-\frac{2(s-r)}{2s+1}}$ . For example the rate of convergence in the special case  $s = 6$  and  $r = 4$  with  $\hat{g}$  as the pilot bandwidth is  $T^{-\frac{4}{13}}$ . This rate of convergence can be achieved with  $\Delta = 0$  and  $r = s = 2$  (i.e. usual kernel regression) by taking the optimal pilot bandwidth  $g_{opt}$  of order  $T^{-\frac{2}{3}}$  (Härdle et al., 1992). For given  $r$  the larger  $s$  the higher is the rate of convergence. Furthermore a Double-Smoothing bandwidth selector by using a pilot bandwidth of the form  $g = CT^\nu h^\delta$  as in Jones et al. (1991) with  $\delta = -\frac{2r}{s}$  and correctly specified  $C$  and  $\nu$  yields a much higher rate of convergence. For example, if  $s > r$  the bandwidth selector is always root T consistent (Heiler and Feng, 1995).

If we use a pilot bandwidth  $g = g_l(T)$  with  $g_r = 0$  for the bandwidth selection at the right end  $t = T$  the theorem 1 in Härdle et al. (1992) and its extension in Heiler and Feng (1995) also hold and the rate of convergence of the bandwidth selector  $\hat{h}_l(T)$  remains the same as the one discussed above. For example if we choose  $p_p = 5$ ,  $p = 2$  and  $\Delta = 1$  and use  $\hat{g}_l(T)$  as the pilot bandwidth, here  $s = 6$  and  $r = 3$ , the rate of convergence of  $\hat{h}_l(T)$  is  $T^{-\frac{6}{13}}$ . This rate is faster than that in the central part. We see, with a polynomial of even order the regression function  $\mu(T)$  is harder to estimate at the end points but the optimal bandwidth  $h_0(T)$  is easier to select, the similar phenomenon has been touched upon in Härdle et al. (1988).

To estimate the optimal bandwidths of the trend-cyclical component or the seasonal component one should only use the pilot estimation and the pilot weight system (when  $\Delta = 1$  the pilot weight system is even not needed) for  $G(t)$  or  $S(t)$  in (10), respectively. The time series decomposition with these selected bandwidths is only optimal for the considered component, it is not optimal for the overall estimation  $\hat{\mu}(t)$  or for the other component.

## 5 Examples with real Data sets

In this section we give some examples of time series decomposed with the proposed procedures. All of the data sets mentioned here are provided by the **IFO-Institut** for economic

**Table 1: Selected Parameters for the Time Series of the Gross Domestic Product in Germany (1968-1994, Quarterly)**

Observation point ( $t$ )	Parameters selected by R-statistic			Parameters selected by Double-Smoothing		
	$\hat{h}_l(t)$	$\hat{h}_T(t)$	$\hat{p}(t)$	$\hat{h}_l(t)$	$\hat{h}_T(t)$	$\hat{p}(t)$
$T$	8	9	1	8	9	1
$T - 1$	13	15	3	14	16	2
$T - 2$	12	15	3	13	16	3
$T - 3$	12	16	3	13	17	3
$T - 4$	12	17	3	8	13	2
$T - 5$	12	18	3	13	19	3
$T - 6$	8	15	2	9	16	2
$T - 7$	9	17	2	10	18	3
$T - 8$	11	20	2	10	19	2
$T - 9$	9	19	2	9	19	2

Notes:  $\hat{h}_l(t)$ ,  $\hat{h}(t)$  and  $\hat{p}(t)$  are selected left bandwidth, total bandwidth and order of polynomial at a point  $t$ , respectively. The results are listed from the right endpoint ( $t = T$ ) to the central part ( $t = T - 9$ ).

research, München, Germany. The first example is the time series of the quarterly gross domestic product in Germany from 1968 to 1994 (billion DM, at current prices). As mentioned above the parameters are selected in the central part and at each point in the right boundary area. The bandwidths  $\hat{h}_l(t)$  and the orders of the polynomials  $\hat{p}(t)$  (here and in the following  $p_m = 2$  in the central part and  $p_m = 3$  in the boundary area) for  $\hat{\mu}(t)$  selected by the R-statistic and by the Double-Smoothing procedure with  $\Delta = 1$ , respectively, are given in table 1.  $p_p = 5$  is used in the pilot smoothing. The selected bandwidths  $\hat{g}$  are used as pilot bandwidths for the Double-Smoothing. The selected right bandwidth  $\hat{h}_r(t)$  at any point  $t$  in the boundary area is always  $T - t$  and is hence omitted. The total bandwidth,  $\hat{h}_T(t) = \hat{h}_l(t) + \hat{h}_r(t) + 1$ , is also given in table 1.

From table 1 we can see that for this example the bandwidths as well as the orders of polynomials in the central part selected by the R-statistic (denoted as  $\hat{h}_R$  and  $\hat{p}_R$ , respectively) and by the Double-Smoothing procedure (denoted as  $\hat{h}_D$  and  $\hat{p}_D$ , respectively) are the same. But the bandwidths as well as the orders of polynomials selected by these two procedures at a point in the boundary area are often different. The orders of polynomials at different points also differ from each other.

Table 2 exhibits the same results as in table 1 for the time series of quarterly private investment in equipment in Germany from 1968 to 1994 (billion DM, at the price in 1991). For this example the bandwidths in the central part selected by R-statistic and by Double-Smoothing respectively, i.e.  $\hat{h}_R$  and  $\hat{h}_D$ , are different.

It is clear that the selected bandwidth should be large if the variance in the time series is relatively large, and it should be small if the variance is relatively small. Further the bandwidth depends also on the form of the trend-cyclical component. As a descriptive

**Table 2: Selected Parameters for the Time Series of the Privat Investment in Equipment in Germany (1968-1994, Quarterly)**

Observation point ( $t$ )	Parameters selected by R-statistic			Parameters selected by Double-Smoothing		
	$\hat{h}_l(t)$	$\hat{h}_T(t)$	$\hat{p}(t)$	$\hat{h}_l(t)$	$\hat{h}_T(t)$	$\hat{p}(t)$
$T$	7	8	1	7	8	1
$T - 1$	12	14	2	13	15	3
$T - 2$	10	13	3	13	16	3
$T - 3$	10	14	3	13	17	3
$T - 4$	11	16	3	12	17	3
$T - 5$	10	16	3	13	19	3
$T - 6$	7	14	2	14	21	3
$T - 7$	7	15	2	8	16	3
$T - 8$	...	...	...	10	19	2
$T - 9$	...	...	...	9	19	2

*Notes:*  $\hat{h}_l(t)$ ,  $\hat{h}(t)$  and  $\hat{p}(t)$  are selected left bandwidth, total bandwidth and order of polynomial at a point  $t$ , respectively. The results are listed from the right endpoint ( $t = T$ ) to the central part ( $t = T - 7$  by R-statistic and  $t = T - 9$  by Double-Smoothing, respectively).

measure we use a noise to curvature ratio (NCR), defined as

$$NCR = \left[ \frac{\sum_{t=3}^T (\hat{r}_t - \bar{r})^2}{\sum_{t=3}^T (\Delta^2 \hat{G}(t))^2} \right]^{1/2},$$

where  $\hat{r}_t$  are the residuals and  $\Delta^2 \hat{G}(t) = \hat{G}(t) - 2\hat{G}(t-1) + \hat{G}(t-2)$  are the second differences. The estimations used in NCR are obtained with the bandwidths and orders of polynomials selected by the R-statistic. For example this ratio is equal to 2.16 for the time series of the monthly unemployment rate in Germany from January 1977 to April 1995 (see Figure 2(A)). Here we obtain  $\hat{h}_R = 17$ . But for the time series of indices of domestic orders received in Germany from January 1978 to December 1994 (1985 = 100) (see Figure 2(B)) NCR is 8.08, that is about four time as large as for the unemployment series. In this case  $\hat{h}_R = 32$  is selected.

The Double-Smoothing procedure is also used to select the bandwidths and orders of polynomials for a single component. The result shows that the optimal bandwidths for the  $\hat{S}(t)$  are larger than the ones for  $\hat{\mu}(t)$  and with lower orders of polynomials while the optimal bandwidths for the  $\hat{G}(t)$  are smaller than the ones for  $\hat{\mu}(t)$ .

As an example the results of the decomposition for the time series of quarterly privat investment in equipment with the bandwidths and the orders of the polynomials selected by the Double-Smoothing procedure are shown in figure 3. Figure 3(a) shows the estimations of the trend-cyclical with the original data, figure 3(b) shows the estimations of the seasonal, while  $\hat{G}'(t)$ , the first derivative, is shown in figure 3(c).

**Insert figure 2 and figure 3 near here**

## 6 Some Remarks

The procedures proposed here are based on the assumption that  $\epsilon_t$  are uncorrelated. If  $\epsilon_t$  are autocorrelated, both the estimation method of the variance and the criteria themselves must be adapted. If the autocorrelation functions are known it is not difficult to obtain a new estimator of  $\sigma^2$  and new criteria for the R-statistic procedure and the Double-Smoothing procedure. But if the autocorrelation functions are unknown one has to estimate them nonparametrically. A modified R-statistic is proposed by Hart and Wehrly (1986) by using repeated measurements data (see also Azzalini, 1995).

In the procedures all seasonal frequencies are used in the model. For a simple seasonal component one might only need to use a proper subset of all seasonal frequencies in the model. Similar to the order of the polynomial a smaller subset of the seasonal frequencies causes a smaller variance but a larger bias while a larger subset of the seasonal frequencies causes the opposite. The variance of the estimation caused by the  $\hat{S}(t)$  is about  $2q_1 - 1$  times, if  $\lambda_q = \pi$  and it is in the model, or  $2q_1$  times, otherwise, of the variance for a normal kernel regression, where  $q_1$  is the number of the seasonal frequencies used in the model. Similar to the choice of  $p$  one can choose a subset of the seasonal frequencies by comparing the minima of the criterion for all possible subsets. Our experiment shows that for most economic time series all seasonal frequencies should be used in the model.

For LWR without seasonality the optimal weight system is the Epanechnikov Kernel (Müller, 1987). But this is probably not true for the time series decomposition model because the correlation between  $\hat{G}(t)$  and  $\hat{S}(t)$  depends on the kernel function as well as on the order of the polynomial. In this paper the problem of forecasting is not considered. However forecasting is also possible by LWR or other nonparametric methods (see e.g. Cao et al., 1992, Fedorov, Hackl and Müller, 1993b and Heiler and Michels, 1994). As far as we know, the problem of selecting the bandwidths for a single component is for the first time considered here. This paper is only a first step in this direction.

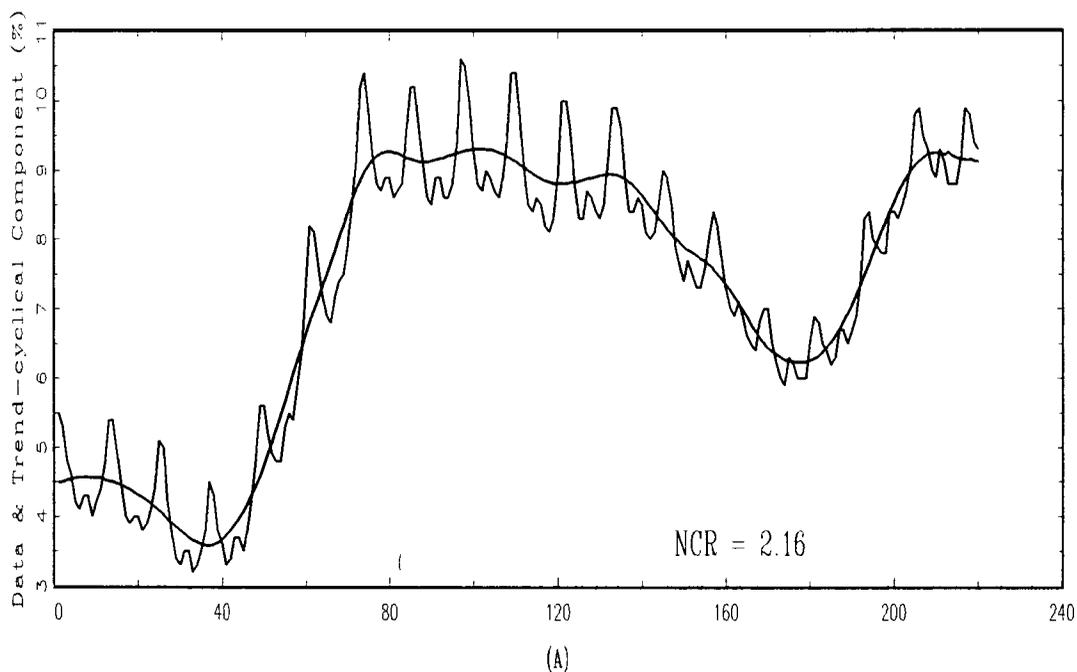
## References

- [1] AZZALINI, A. (1995): Topics in Repeated Measurements Analysis: Nonparametric Smoothing Methods. Work report, University of Padua.
- [2] CAO, R., GONZÁLEZ-MANTEIGA, W., PRADA-SÁNCHEZ, J.M., GARCÍA-JURADO, I. and FEBRERO-BANDE, M. (1992): Forecasting Using a Semiparametric Model. In DODGE, Y. and WHITTAKER, J. (Eds): *Computational Statistics, Volume 1*, 327-30. Physica, Heidelberg.
- [3] CLEVELAND, R.B., CLEVELAND, W.S., McRAE, I.E., TERPENNING, I. (1990): STL: A Seasonal-trend decomposition Procedure Based on LOWESS (with discussion). *Journal of Official Statistics*, vol. 6, No. 1, 3-73.
- [4] CLEVELAND, W.S. (1979): Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74, No. 36, 829-836.
- [5] CLEVELAND, W.S., DELVIN, S.J. (1988): Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *J. Amer. Statist. Assoc.*, 83, 596-610.

- [6] CLEVELAND, W.S., DELVIN, S.J., GROSSE, E. (1988): Regression by Local Fitting. Methods, Properties and Computational Algorithmus. *Journal of Econometrics*, 37, 87-114.
- [7] CLEVELAND, W.S. and LOADER, C. (1995): Smoothing by Local Regression: Principles and Methods. Work report, AT&T Bell Laboratories.
- [8] FAN, J. and GIJBELS, I. (1992): Variable Bandwidth and Local Linear Regression Smoothers. *Annals of Statistics*, 20, 2008-36.
- [9] 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.
- [10] FEDOROV, V.V., HACKL, P. and MÜLLER, W.G. (1993a): Moving Local Regression: the Weight Function. *J. Nonparametric Statist.*, 2, 355-68.
- [11] FEDOROV, V.V., HACKL, P. and MÜLLER, W.G. (1993b): Optimized Moving Local Regression: Another Approach to Forecasting. In MÜLLER, W.G., WYNN, H.P. and ZHIGLJAVSKY, A. (Eds): *Model-Oriented Data Analysis 3*, 137-44. Physica, Heidelberg.
- [12] GASSER, T., KNEIP, A. and KÖHLER, W. (1991): A Flexible and Fast Method for Automatic Smoothing. *J. Amer. Statist. Assoc.*, 86, 643-52.
- [13] GASSER, T., SROKA, L. and JENNEN-STEINMETZ, C. (1986): Residual Variance and Residual Pattern in Nonparametric Regression. *Biometrika*, 73, 625-33.
- [14] HÄRDLE, W., HALL, P. and MARRON, J.S. (1988): How Far Are Automatically Chosen Regression Smoothing Parameters from Their Optimum? (with discussion). *J. Amer. Statist. Assoc.*, 83, 86-99.
- [15] HÄRDLE, W., HALL, P. and MARRON, J.S. (1992): Regression Smoothing Parameters That Are not Far from Their Optimum. *J. Amer. Statist. Assoc.*, 87, 227-33.
- [16] HALL, P., KAY, J.W. and Titterington, D.M. (1990): Asymptotically Optimal Difference-based Estimation of Variance in Nonparametric Regression. *Biometrika*, 77, 521-8.
- [17] HART, J. and WEHRLY, T.E. (1986): Kernel Regression Estimation Using Repeated Measurements Data. *J. Amer. Statist. Assoc.*, 81, 1080-88.
- [18] HASTIE, T. and LOADER, C. (1993): Local Regression: Automatic Kernel Carpentry (with discussion). *Statistical Science*, 8, 120-43.
- [19] HEILER, S. (1966): *Analyse der Struktur Wirtschaftlicher Prozesse durch Zerlegung von Zeitreihen*. Dissertation, Tübingen.
- [20] HEILER, S. (1994): Kernel Method in Time Series Analysis. In A. ZELIAS (Ed): *Prestrzenno-Czasowe Modelowanie i Prognozowanie Zjawisk Gospodarczych* Akademia Ekonomiczna w Krakowie. Krakow 1994, pp. 94-103.
- [21] HEILER, S. (1995): Zur Glättung Saisonaler Zeitreihen. Forthcoming.

- [22] HEILER, S. and FENG, Y. (1995): A Simple Root  $n$  Bandwidth Selector for Nonparametric Regression. *Unpublished discussion paper, University of Konstanz.*
- [23] HEILER, S. and MICHELS, P. (1994): *Deskriptive und Explorative Datenanalyse*, R. Oldenbourg, München.
- [24] JONES, M.C., MARRON, J.S. and PARK, B.U. (1991): A Simple Root  $n$  Bandwidth Selector. *Ann. Statist.*, 19, 1919–32.
- [25] LEJEUNE, M. (1985): Estimation Non-paramétrique par Noyaux: Regression Polynomiale Mobile. *Revue de Statistiques Appliquées*, 33, 43–68.
- [26] MÜLLER, H.-G. (1985): Empirical Bandwidth Choice for Nonparametric Kernel Regression by Means of Pilot Estimators. *Statistics & Decisions*, Supplement Issue No. 2, 193–206.
- [27] MÜLLER, H.-G. (1987): Weighted Local Regression and Kernel Methods for Nonparametric Curve Fitting. *J. Amer. Statist. Assoc.* 82, 231–8.
- [28] RICE, J. (1983): Methods for Bandwidth Choice in Nonparametric Kernel Regression. *Computer Science and Statistics*. 1983, 186–90.
- [29] RICE, J. (1984): Bandwidth Choice for Nonparametric Regression. *Annals Statistics*, 12, 1215–30.
- [30] RUPPERT, D., SHEATHER, S.J. and WAND, M.P. (1994): An Effective Bandwidth Selector for Local Least Squares Regression, unpublished research report.
- [31] SHEATHER, S.J. and JONES, M.C. (1991): A Reliable Data-based Bandwidth Selection Method for Kernel Density. *J. R. Statist. Soc. B*, 53, 683–90.
- [32] SCHLITZGEN, R. and STREITBER, B. (1991): *Zeitreihenanalyse*. R. Oldenbourg, München.
- [33] STONE, C.J. (1977): Consistent Nonparametric Regression (with discussion). *Annals of Statistics*, 5, 595–620.
- [34] ULLAH, A. (1988): Nonparametric Estimation of Econometric Functionals. *Canadian J. Economics*, XXI, No.3, 625–58.
- [35] ULLAH, A. (1989): Nonparametric Estimation and Hypothesis Testing in Econometric Models. In ULLAH, A. (Ed): *Semiparametric and Nonparametric Econometrics*, 101–29. Physica, Heidelberg.

## Monthly Unemployment Rate in Germany from January 1977 to April 1995



## Indices of Domestic Orders Received in Germany from January 1978 to December 1994

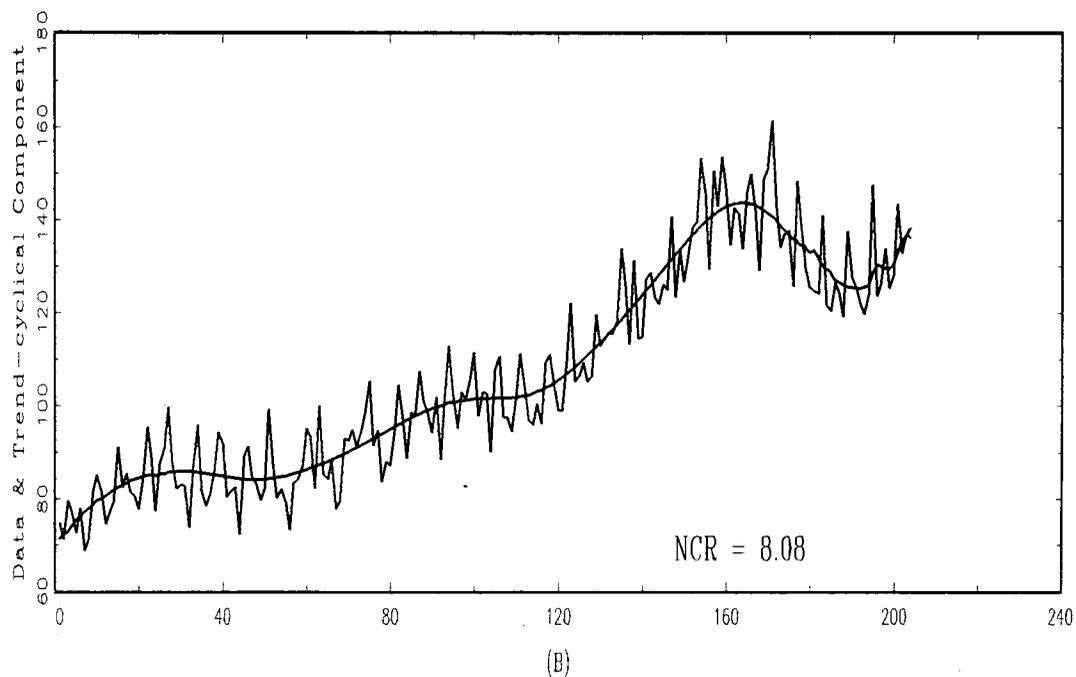
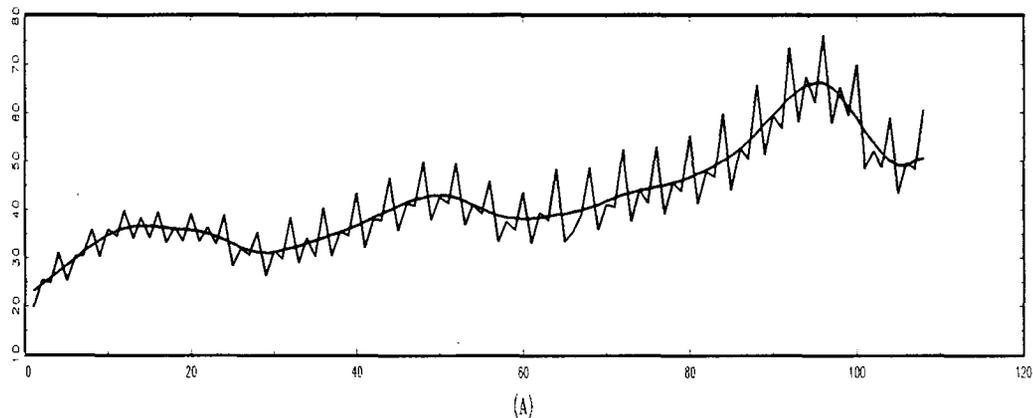
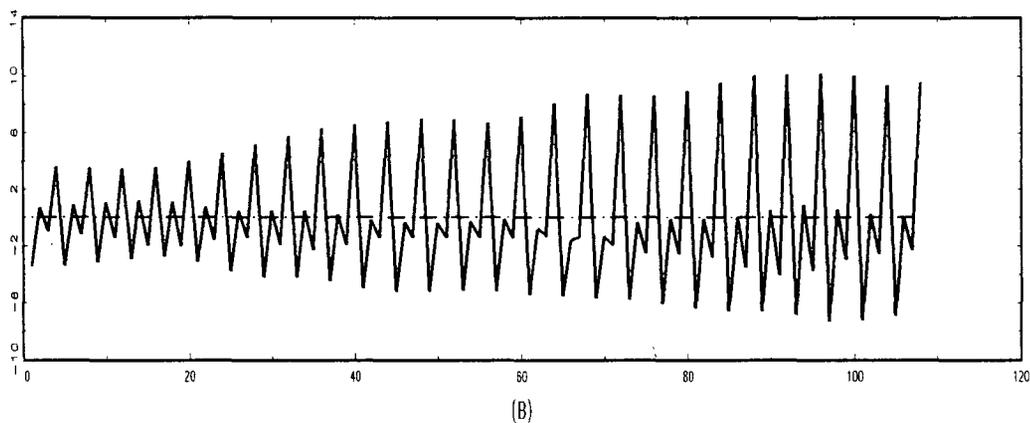


Figure 2. Two examples of seasonal time series: (A) time series with small noise to curvature ratio (NCR) and (B) time series with large NCR. The bandwidths selected by the R-statistic in the central part are 17 and 32, respectively.



The Seasonal Component



The Derivative of First Order of the Trend-cyclical Component

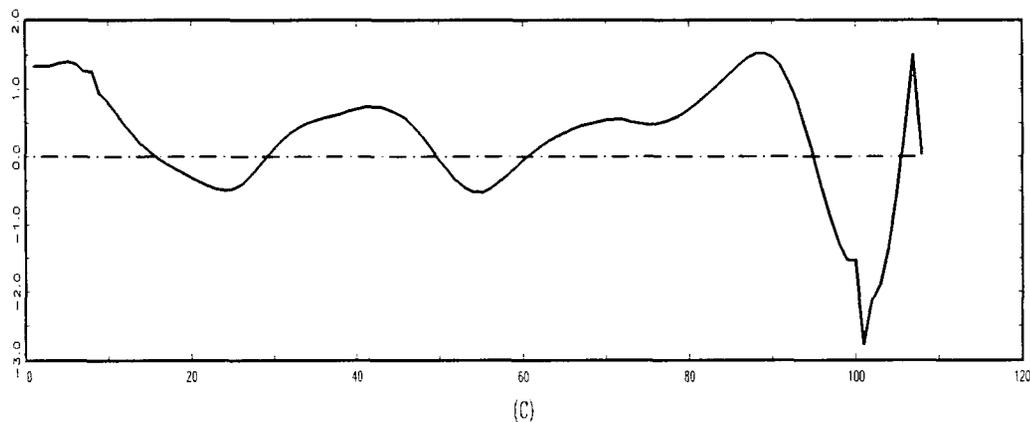


Figure 3. The decomposition results of the time series of quarterly private investment in equipment in Germany with the bandwidths and orders of polynomials selected by the Double-Smoothing procedure ( $P_p=5$ ,  $P_m=3$  and the Bisquare Kernel is used).