

ESTIMATING AND TESTING THE INNOVATION VARIANCE OF ON-OFF EEG TIME SERIES

W. Härdle

Institut für Angewandte Mathematik
Im Neuenheimer Feld 294
D-6900 Heidelberg 1

Th. Gasser

Zentralinstitut für Seelische Gesundheit
J 5
D-6800 Mannheim 1

Summary.

Three estimation and testing procedures of the innovation variance of an ON-OFF electroencephalogram time series are considered. A Monte Carlo study of these procedures was carried out to compare the empirical power of the tests. The test based on a jackknife estimate seems to be appropriate to the human EEG situation since the other procedures, the F-Test and a test based on an estimate of Hannan and Nicholls are sensible to outliers.

I. INTRODUCTION

The electroencephalogram (EEG) is a graphical representation of electrical potential differences in the human brain. The changing dipole distribution in the neuro-electrical apparatus of the cerebral cortex can be considered as the source of the EEG. The normal potential changes of the human EEG are in frequency between .5 and 30 Hz and in amplitude between 10 and 500 μ V. The exact origin of the EEG is not localizable, since impulses transmitted along a single primary nerve fibre will act at varying degrees of intensity on many other secondary nerve cells. However the EEG is established since decades as a helpful diagnostic tool in neuropathology, (Dumermuth (1976), p. 2-5).

The properties of the EEG depend on exogene factors such as light stimulation or noise and on endogene factors such as age, drowsiness or mental retardation. With closed eyes the (occipital) main frequency of an adult EEG is about 10 Hz. This frequency is called the (dominant) alpha rhythm. If a subject opens its eyes

the alpha rhythm disappears after a short time. The resulting process is called an ON-OFF EEG time series which is clearly unstationary since the frequencies change under the different conditions "eyes closed" and "eyes opened".

Much research was done since the discovery of the EEG to derive diagnostically relevant parameters. One promising approach towards an "EEG-model" is the fit by an autoregressive scheme which was suggested by several authors (Ahlborn and Zetterberg (1976); see Zetterberg (1978) for further references). For an ON-OFF time series this method makes only sense if we fit each regime, the one with closed the other with opened eyes, separately by an autoregressive model.

This paper is concerned with estimating and testing differences in the innovation variance of ON-OFF time series. Detection of shifts in the statistical behavior of the EEG is useful in neurophysiology and psychology where the reaction of certain EEG properties after instructions or stimulations serve as a measure for the mental retardation or brain dysfunctions. Fuller (1977) studied the attenuation of the alpha rhythm during problem solving and Baumeister (1963,1967) considered the alpha responsiveness to photic stimulations in mental defectives. The amount of alpha power in the EEG reflects also the age and functional status of the brain. With maturation the dominant frequency becomes more rapid, and brain damage, dysfunctions or deterioration causes frequency slowing in the brain regions involved (John, E.R. et al.(1981), Ahn, H. et al.(1981)).

The observations are divided in two regimes, separated by the moment when the subject closes its eyes.

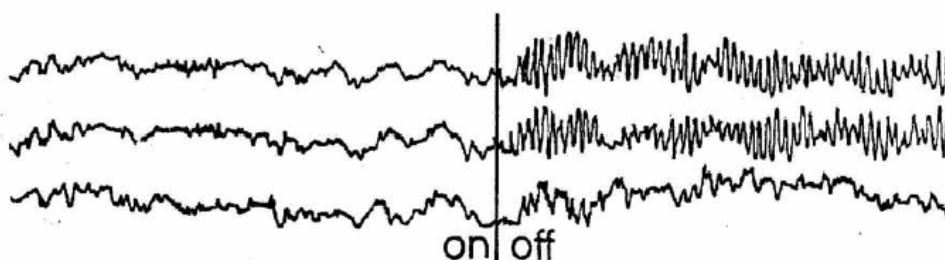


Figure 1

If the observations are assumed to be mutually independent and the data are normally distributed the right test for differences in variance would be the F-test. The assumption of independence is certainly not true for EEG time series and the normality of the data is also not evident. Box(1953) pointed out that the performance of the F-test heavily depends on the assumption of normality, so we studied similar to Davis(1978, 1979) tests which are robust with respect to deviations of normality. We also analyzed an estimate from Hannan and Nicholls (1977) although the assumptions for the asymptotic properties of this estimate include the normality of the data. This estimate is based on the periodogram and is thus quickly computed by means of the Fast Fourier Transform (FFT). Since most of the EEG analysts extract information out of the spectrum and so compute it in any case, this estimate can be get as an appendix to the FFT.

In the independent identically distributed situation there are a lot of papers proposing and comparing procedures for testing equality of variance (Shorack (1969); Miller (1968); Layard (1973)). Shorack for example compared several robust criteria on the basis of Pitman efficiency and Monte Carlo studies of power functions. In the non i.i.d. situation Davis (1978, 1979) evaluated some robust procedures such as the Box-Anderson data splitting technique and the Jackknife for an AR(p) model in both regimes. One of his basic assumptions is that the innovations in both data regimes have the same distribution after normalization by their variance. This cannot be said of an ON-OFF EEG time series since the generating innovations in the brain can be entirely different in both regimes(see figure 1).

In the next section we propose the model for an ON-OFF EEG time series and formulate the test problem. In section 3 we derive the jackknife estimate and the asymptotic properties of the test statistics involved. In section 4 we state the likelihood ratio procedure and discuss a test which is based on Hannan-Nicholls' estimate. The results of a small Monte Carlo study are presented in section 5 and in the last section we apply the techniques to the EEG.

II. AUTOREGRESSIVE MODEL OF AN ON-OFF EEG TIME SERIES

Let us denote with $Y_1(t)$ the EEG under opened eyes and with $Y_2(t)$ the EEG under closed eyes. We assume that the EEG is generated in both regimes by an autoregressive scheme of known order p . In reality the order is not known and may be possibly different in the two regimes. By reasons of simplicity we did not

consider these effects but by reading through the proofs it is clear that the statements also hold for different order p_1 and p_2 . We propose the following model

$$(2.1) \quad \begin{aligned} \sum_{s=0}^p \alpha_s^{(1)} Y_1(t-s) &= \varepsilon_t^{(1)} & t=1,2,\dots, T_1 \\ \sum_{s=0}^p \alpha_s^{(2)} Y_2(t-s) &= \varepsilon_t^{(2)} & t=T_1+1,\dots, T \end{aligned}$$

where the $Y_1(1-p), \dots, Y_1(0)$ are assumed to be fixed. The $\alpha_s^{(j)}$; $s=0, \dots, p$; $j=1,2$ are the autoregressive coefficients. $\varepsilon_t^{(j)}$; $j=1,2$ are the innovations which are an independent sample from the distributions $G_1(x/\sigma_1)$ for $t \leq T_1$ and $G_2(x/\sigma_2)$ for $t > T_1$. σ_j^2 ; $j=1,2$ denotes the innovation variance and the distributions satisfy the following set of constraints:

$$(2.2) \quad \int x dG_j(x) = 0, \int x^2 dG_j(x) = 1, \int x^k dG_j(x) = \gamma_j + 3 < \infty.$$

We rewrite the model (2.1) in terms of the back-shift operator B (i.e. $B Y_t = Y_{t-1}$):

$$(2.3) \quad \begin{aligned} P_1(B) Y_1(t) &= \varepsilon_t^{(1)} & t \leq T_1 \\ P_2(B) Y_2(t) &= \varepsilon_t^{(2)} & t > T_1, \end{aligned}$$

where $P_j(B) = \sum_{s=0}^p \alpha_s^{(j)} B^s$; $B^0 = I$. We will assume that the autoregressive parameters $\alpha_s^{(j)}$ correspond to stationary processes in the sense that all roots of the characteristic polynomials $P_j(B)$; $j=1,2$ be outside the unit circle in the complex plane.

We want to test the equality of the innovation variances σ_1^2 and σ_2^2 . As the alternative of $H_0: \sigma_1^2 = \sigma_2^2$ we take $H_1: \sigma_2^2 > \sigma_1^2$. The alternative $H_1: \sigma_1^2 \neq \sigma_2^2$ makes no sense for the ON-OFF EEG time series if we accept Berger's generator model for the alpha rhythm of the EEG (Berger (1930); Goldstein (1975)).

The test we are interested in is:

$$(2.4) \quad H_0: \Delta^2 = 1 \quad \text{vs.} \quad H_1: \Delta^2 > 1$$

where $\Delta^2 = \sigma_2^2 / \sigma_1^2$.

III. THE JACKKNIFE PROCEDURE

This method was used by Shorack (1969) for the i.i.d. case and by Davis (1979) for autoregressive time series with $G_1 = G_2$. For simplicity we present her only the "one-leave-out" jackknife. We want to estimate $\chi_j^2 = \log(\sigma_j^2)$. In this case the procedure works as follows. Let us define the pseudovalues

$$\begin{aligned} \eta_c^{(1)} &= T_1 \cdot \log(\hat{\sigma}_1^2) - (T_1 - 1) \cdot \log(v_c^{(1)}) \quad t=1, \dots, T \\ \eta_c^{(2)} &= T_2 \cdot \log(\hat{\sigma}_2^2) - (T_2 - 1) \cdot \log(v_c^{(2)}) \quad t=T_1+1, \dots, T \end{aligned} \quad (3.1)$$

where $T=T_1+T_2$ and

$$v_c^{(j)} = \sum_{s \neq t} \hat{\epsilon}_s^{(j)2} / (T_j - 1); j=1, 2 \quad \hat{\sigma}_j^2 = \sum_s \hat{\epsilon}_s^{(j)2} / T_j; \quad j=1, 2.$$

The jackknife estimate of $\log \sigma_j^2$ is now

$$\eta_j = \sum_t \eta_c^{(j)} / T_j; \quad j=1, 2.$$

The residual sum of squares is

$$W_j = \sum_t (\eta_c^{(j)} - \eta_j)^2.$$

The jackknife theorem is stated below.

(3.5) Theorem.

If $\min(T_1, T_2) \rightarrow \infty$

$$\frac{(\eta_2 - \eta_1 - \log \Delta^2)}{\left(\sum_{j=1}^2 W_j / T_j (T_j - 1) \right)^{1/2}} \xrightarrow{d} N(0, 1)$$

Proof: Since the proof is basically the same as Davis one it is only sketched in the appendix.

The asymptotic relation (3.5) allows us to test H_0 vs. H_1 on the basis of the estimates (3.3). If we leave more than one point out, say m , and fix $k_j = T_j/m$, letting m tend to infinity, the resulting distribution is a $t_{k_1+k_2-2}$ distribution. For long time series it may be convenient to pick $m > 1$ and to use the Student-t approximation.

IV. LIKELIHOOD RATIO AND HANNAN'S ESTIMATE

Let us rewrite the model (2.3) as a regression equation. The observations are $Y=(Y_1, \dots, Y_T)$, the first regime $Y^{(1)}=(Y_1, \dots, Y_{T_1})$ and the second $Y^{(2)}=(Y_{T_1+1}, \dots, Y_T)$. Define $x_i=(y_i y_{i-1} y_{i-2} \dots y_{i-p+1})'$, $\epsilon_{(1)}=(\epsilon_1, \dots, \epsilon_{T_1})'$, $\epsilon_{(2)}=(\epsilon_{T_1+1}, \dots, \epsilon_T)'$ and the design matrices are $X_1=(x_1 x_2 \dots x_{T_1})'$, $X_2=(x_{T_1+1} \dots x_T)'$. The model (2.1) now reads:

$$Y^{(j)} = X_j \theta_j + \epsilon_{(j)} \quad j=1, 2.$$

If Q denotes the likelihood ratio statistic for H and $\epsilon_{(j)}$ are assumed to be normal it is easy to show that:

$$-2 \log Q = \sum_{j=1}^2 T_j \log(\hat{\sigma}_j^2 / \sigma_j^2),$$

$$\text{where } \hat{\sigma}_j^2 = \|Y^{(j)} - X_j \theta_j\|^2 / T_j$$

$$\hat{\theta}_j = (X_j' X_j)^{-1} X_j' Y^{(j)} \quad \sigma_j^2 = \sum_{j=1}^2 \|Y^{(j)} - X_j \theta_j\|^2 / T.$$

One concludes that the likelihood ratio procedure depends only on $F = \sigma_2^2 / \sigma_1^2$. The effect of non-normality on the F -statistic is asymptotically the same as in the i.i.d. case as the following theorem shows.

(4.3) Theorem

$$\sqrt{\frac{T_1 T_2}{T_1 + T_2}} (F - \Delta^2) / \Delta^2 \xrightarrow{d} N(0, 2 + (\gamma_1 + \lambda \gamma_2) / (1 + \lambda)),$$

where $\lambda = \lim T_2 / T_1$.

Theorem (4.3) is also proven in the appendix. This statement indicates that if a consistent estimate of $\gamma_j; j=1, 2$ can be obtained the test would be asymptotically robust against non-normality. So one can expect that the significance level of the test tends to the nominal level as the sample size tends to infinity.

Another estimate of the innovation variance is proposed by Hannan and Nicholls (1977) and is based on the formula

$$\sigma_j^2 = \exp\left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \log 2\pi f(\omega) d\omega \right\},$$

where $f(\omega)$ is the spectral density of the process. For the following theorem we assume f to belong to the class $A_\alpha, 0 < \alpha < 1$ i.e. $\sup |f(\omega + \sigma) - f(\omega)| \leq A \cdot |\sigma|^\alpha$ (Zygmund (1968), p. 42). The estimate of σ_j^2 depends on a parameter m which can be interpreted as a smoothing parameter of the periodogram.

$$\sigma_j^2(m_j) = \frac{1}{m_j} \exp\left\{ M_j^{-1} \sum_{t=0}^{M_j-1} \log\left\{ M_j^{-1} \sum_{s=1}^{m_j} |W_j(\omega_{tm+s})|^2 - \psi(m_j) \right\} \right\}$$

where $M_j = \lceil (T_j - 1) / (2m_j) \rceil$ the largest integer greater than $(T_j - 1) / (2m_j)$, and

$$W_j(\omega_k) = T_j^{-1/2} \sum_t Y_j(t) e^{it\omega_k}, \quad \omega_k = 2\pi k / T_j, \quad 0 < k < \frac{1}{2} T_j$$

with k an integer and $\psi(x) = d \log \Gamma(x) / dx$ the Digamma function, tables of which are given by Abramowitz and Stegun (1966), p. 267-273. The subtraction of the Digamma function $\psi(m)$ is bias correcting. It may be shown that if $Y(t)$ is $N(\mu, \sigma^2)$ i.i.d., i.e. $f(\omega) = \sigma^2 / 2\pi$, $T_j^{1/2} (\sigma_j^2(m) - \sigma^2)$ is asymptotically normal (Davis and

Jones (1968)). Hannan-Nicholls' first theorem establishes both consistency and normality for the slightly more general class of spectra $f \in A_\alpha$. The asymptotic variance of $\hat{\sigma}^2(m)$ depends on $2m\psi'(m)$ which is tabulated below.

(4.6)

m	1	2	3	4	5	6	7	8
$2m\psi'(m)$	3.29	2.58	2.37	2.271	2.213	2.176	2.15	2.13

Table 1

This quantity decreases from 3.29 at $m=1$ to 2 at $m=\infty$, but changes only slowly after $m=3$. By the interpretation of m as a smoothing parameter it can be expected that for fixed T the bias will be smallest when m is one and will increase, in general, as m increases. The next theorem states the asymptotic behavior of $D_{m_1, m_2} = \hat{\sigma}_2^2(m_2) / \hat{\sigma}_1^2(m_1)$.

(4.7) Theorem

Let $f_j(\omega)$ the spectral density $Y_j(t)$ ($j=1,2$) be positive for $\omega \in [-\pi, \pi]$ and $f_j(\omega) \in A_\alpha$, $\alpha > \frac{1}{2}$ and let $\epsilon^{(j)}(t)$ be normal $N(0, \sigma_j^2)$.

$$\sqrt{\frac{T_1 T_2}{T_1 + T_2}} (D_{m_1, m_2} - \Delta^2) / \Delta^2 \xrightarrow{K} N(0, \sigma_D^2), \text{ where}$$

$$\sigma_D^2 = \frac{1}{(1+\lambda)} m_1 \psi'(m_1) + \frac{\lambda}{(1+\lambda)} m_2 \psi'(m_2).$$

Since $m\psi'(m)$ converges to 1 as m tends to infinity it follows that the Pitman a.r.e. of the test based on D_{m_1, m_2} and (4.7) with respect to the F-test tends to one. Hannan and Nicholls report that there is no doubt that their theorem holds for non-normal data too.

The term for $k = \frac{1}{2}T$, T even, was omitted for simplicity. Had it been included the exponent of the expression (4.5) would be changed but the asymptotic results still hold in that case. Since $\sigma_j^2(m_j)$ is computed out of periodogram values it is numerically more stable than the two other procedures which are coming out of a recursion due to Durbin (1960) or an inversion of a Toeplitz matrix. For a discussion of numerical aspects of the estimation of autoregressive coefficients we refer to Cybenko (1981).

V. MONTE CARLO STUDY

A small Monte Carlo computer simulation was conducted to study the power of the three proposed test procedures. Time series of length 400 were generated from an autoregressive model with order $p=1$. A possible

shift in innovation variance or autoregressive parameters occurred after 200 observations making the two regimes of length 200. The direct comparison of this Monte Carlo study with that of Miller (1969) or Davis (1979) is not possible since we are interested in one sided alternatives. We fixed Δ^2 at the following levels $\Delta^2 = 1; 1.2; 1.4; 1.6; 1.8$ whereas the papers of Miller and Davis deal with $\Delta^2 = 1; 2; 4; 6$ since they are testing $H_0: \sigma_1^2 = \sigma_2^2$ vs. $H_1: \sigma_1^2 \neq \sigma_2^2$.

The distributions of $\epsilon(t)$ run through the following set of combinations: normal/normal; normal/uniform; normal/double exponential; uniform/double exponential. Pseudo random numbers were generated using the algorithm RANORM (see Andrews, Bickel et al. (1972)) for the normal distribution. For the double exponential and the uniform random numbers we used the routines GGEXN and GGUPS respectively from the IMSL-library. Each series was generated 2000 times to estimate the probability of detecting a shift in innovation variance using significance level $p=.05$. The results of the simulation are presented in Table 2. The autoregressive coefficients were $\alpha_1^{(1)} = -.9$, $\alpha_1^{(2)} = -.8$.

Of course it is not fair to compare power functions of tests that do not maintain their nominal significance level. So one may argue that the power of the F-test in the heavy tailed situations (normal/double exponential) is high because it has higher significance level. In any event the F-test power was far away from the nominal level $p=.05$ and is thus a very suspect candidate for tests in real situations. We also used the Hannan-Nicholls estimate in the non-normal situation to receive a little bit more insight in its behavior. Since its power resembles that of the F-test in the heavy tailed situations it can be conceived that it is also sensitive to deviations of normality. We remember the fact that it is constructed only for the normal case and we conjecture that in the non-normal case some correcting term to the asymptotic variance should be added. It shows for $m=4$ the best power compared with the others ($m=1,2,3$). This seems to be due to the relatively sharp peak in the spectrum, generated by $\alpha_1^{(1)}, \alpha_2^{(2)}$ which are close to the boundary of the unit circle in the complex plane. The effect of increasing m is to flatten the spectrum and to approximate it to a white noise.

The jackknife does the best job of maintaining the nominal significance level. In the normal/normal case it clearly lacks behind the F-test but in all other cases it comes closest to $p=.05$. So the jackknife comes out as a reliable estimate in our Monte Carlo study and

seems to be a trustworthy procedure.

VI. APPLICATION TO ON-OFF EEG TIME SERIES

We applied the three proposed tests to one channel of the EEG of a child which was about 10 years old (Figure 2).



Figure 2

Since $T_1 = 752$ and $T_2 = 669$ the asymptotic results of the foregoing chapters can be applied. The output of the tests are presented in Table 3.

Table 3. Teststatistics for the three derived procedures

	σ_1^2	σ_2^2	stat.	P
F	135.	156.	1.52	<.07
Jackkn.	4.78	4.95	1.649	<.05
$D_{1,1}$	137.	157.	2.38	<.05
$D_{2,2}$	132.	157.	3.17	<.05
$D_{3,3}$	129.	161.	3.71	<.05
$D_{4,4}$	127.	159.	3.76	<.05

The statistics are all significantly high to reject the hypothesis $H_0: \Delta^2=1$. The different Hannan-Nicholls estimates ($m=1,2,3,4$) increase the statistics as m increases. The chosen order for the first regime (ON) was $p=5$, for the second regime (OFF) it was $p=8$. Both were recommended orders due to a criterion of Schwarz (1978).

Table 2

	1	1.2	1.4	1.6	1.8
G_1 =normal, G_2 =normal					
F	.059	.4175	.822	.973	1.
Jackknife	.045	.344	.759	.96	1.
Hannan-Nicholls $m=1$.057	.3245	.65	.871	.956
$m=2$.067	.346	.706	.971	.999
$m=3$.062	.364	.748	.935	.999
$m=4$.055	.315	.714	.922	.999
G_1 =uniform, G_2 =normal					
F	.0625	.52	.9115	.992	1.
Jackknife	.0506	.467	.888	.99	1.
Hannan-Nicholls $m=1$.051	.294	.672	.889	.986
$m=2$.034	.314	.747	.948	.999
$m=3$.042	.348	.798	.966	.999
$m=4$.032	.314	.754	.955	.999
G_1 =double exp, G_2 =normal					
F	.0975	.382	.6725	.872	.949
Jackknife	.0725	.2885	.575	.8	.918
Hannan-Nicholls $m=1$.107	.349	.647	.836	.931
$m=2$.117	.397	.689	.864	.955
$m=3$.137	.424	.724	.882	.968
$m=4$.114	.393	.691	.873	.962
G_1 =double exp, G_2 =uniform					
F	.1085	.404	.733	.9085	.972
Jackknife	.069	.3125	.6335	.851	.943
Hannan-Nicholls $m=1$.104	.333	.648	.84	.953
$m=2$.108	.375	.703	.893	.971
$m=3$.112	.399	.734	.92	.984
$m=4$.97	.365	.703	.901	.971

VII. APPENDIX.

We begin with the proof of Theorem (3.5) and decompose the relevant terms in the following way:

$$(7.1) \quad T_j^{1/2} (n_j - \log \sigma_j^2) = T_j^{1/2} (\log G_j^2 - \log(\sigma_j^2)) - (T_j - 1) \sum_t (\log(v_t^{(j)} - \log G_j^2)) / T_j^{1/2}.$$

As in Davis (1979) the second term tends in probability to zero and the first term is treated by the CLT:

$$(7.2) \quad T_2^{1/2} (\log G_2^2 - \log(\sigma_1^2), \log G_2^2 - \log(\sigma_2^2)) \rightarrow N(0, \Gamma_1)$$

where $\Gamma_1 = \begin{pmatrix} \lambda(2+\gamma_1) & 0 \\ 0 & (2+\gamma_2) \end{pmatrix}$ and $\lambda = \lim T_2/T_1$.

Another application of Davis (1979) Lemma 4 yields:

$$(7.3) \quad W_j/T_j \xrightarrow{P} (2+\gamma_j), \quad j=1,2$$

and hence Slutsky's theorem completes the proof. To establish the asymptotic distribution of the F-statistic we observe that

$$(7.4) \quad (F - \Delta^2) / \Delta^2 = \frac{\sigma_1^2}{\sigma_2^2} \left(\frac{\sigma_2^2 - \sigma_1^2}{\sigma_2^2} - \frac{\sigma_1^2 - \sigma_1^2}{\sigma_1^2} \right).$$

If we now apply Lemma 1 of Davis (1979) and observe that $\hat{\sigma}_1^2 \xrightarrow{P} \sigma_1^2$ the conclusion of Theorem (4.3) follows.

(7.5) Lemma (Hannan - Nicholls)

$$T_j^{1/2} (\sigma_j^2(m_j) - \sigma_j^2) \xrightarrow{L} N(0, \sigma_{m_j}^2).$$

where $\sigma_{m_j}^2 = 2\sigma_j^4 \cdot m_j \cdot \psi'(m_j)$.

Hence by the multivariate form of the central limit theorem we have:

$$(7.6) \quad T_2^{1/2} (\sigma_1^2(m_1) - \sigma_1^2, \sigma_2^2(m_2) - \sigma_2^2) \xrightarrow{L} N(0, \Gamma_2).$$

where $\Gamma_2 = \begin{pmatrix} \lambda 2\sigma_1^4 m_1 \psi'(m_1) & 0 \\ 0 & 2\sigma_2^4 m_2 \psi'(m_2) \end{pmatrix}$, $\lambda = \lim T_2/T_1$

if we decompose $(D_{m_1, m_2} - \Delta^2) / \Delta^2$ in a similar fashion as we did in (7.4) the theorem (4.7) follows if we apply the CLT to each summand separately and keep track on

the norming factor $\frac{T_1 T_2}{T_1 + T_2}$.

REFERENCES

Abramowitz, M., Stegun, I.A., Handbook of Mathematical Functions Applied Mathematical Series. (1966).

Ahlbom, G., Zetterberg, L.H., A Comparative Study of Five Methods for Analysis of EEG, Technical Report (1976).

Ahn, H., Prichep, L., John, E.R., Developmental Equations Reflect Brain Dysfunctions Science. (1980), Vol. 210, pp. 1259.

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., Tukey, J.W., Robust Estimation of Location, Princeton University Press, (1972).

Baumeister, A.A., Hawkins, W.F., Alpha Responsiveness to Photic Stimulation in Mental Defectives, A.J. of Mental Deficiency, (1967), vol. 71, pp. 83.

Baumeister, A.A., Spain, C.J., Ellis, N.R., A Note on Alpha Block Duration in Normals and Retardates, A.J. of Mental Deficiency, (1963), vol. 67, pp. 723.

Berger, H., Über das Elektroencephalogramm des Menschen II, J. Psychol. Neurol. (Leipzig), (1930), vol. 40, pp. 160.

Box, G.E.P., Non-Normality and Tests on Variances, Biometrika, (1953), vol. 40, pp. 318.

Cybenko, G., The Numerical Stability of the Levinson-Durbin Algorithm for Toeplitz System of Equations. SIAM J. Sci. Stat. Comput. (1980), vol. 1, pp. 303.

Davis, H.T., Jones, R.H., Estimation of the Innovation Variance of a Stationary Time-Series, JASA, (1968), vol. 63, pp. 141.

Davis, W.W., Robust Methods for Detection of Shifts of the Innovation-Variance of a Time-Series, Technometrics, (1979), vol. 21, pp. 313.

Davis, W.W., Robust Interval Estimation of the Innovation-Variance of an Arma-Model, Ann. Stat. (1977), vol. 5, pp. 700.

Dumermuth, G., Electroencephalographie im Kindesalter, G. Thieme Verlag, Stuttgart, (1976).

Durbin, J., Estimation of Parameters in Time-Series Regression Models, J. Royal Stat. Soc. Series B, (1960), vol. 22, pp. 139.

Fuller, P.W., Computer Estimated Alpha Attenuation During Problem Solving in Children with Learning Disabilities, EEG and Clinical Neurophysiology, (1977), vol. 42, pp. 149.

Goldstein, S., Phase Coherence of the Alpha Rhythm During Photic Blocking, EEG and Clinical Neurophysiology, (1970), vol. 29, pp. 127.

Hannan, E.J., Nicholls, D.F., The Estimation of the Prediction Error Variance, JASA, (1977), vol. 72, pp. 843.

John, E.R., Ahn, H., Prichep, L., Developmental Equations for the Electroencephalogram, Science, (1980), vol. 210, pp. 1255.

Layard, M.W.J., Robust Large-Sample Tests for Homogeneity of Variances, JASA, (1973), vol. 68, pp. 195.

Miller, R.G. Jr., Jackknifing Variances, Ann. Math. Stat. (1968), vol. 39, pp. 568.

Schwarz, G., Estimating the Dimension of a Model, Ann. Math. Stat., (1978), vol. 6, pp. 461.

Shorack, G.R., Testing and Estimating Ratios of Scale Parameters, JASA, (1969), vol. 64, pp. 999.

Zetterberg, L.H., Recent Advances in EEG Data Processing, EEG Supplement Nr. 34, (1978), vol. 34, pp. 19.

Zygmund, A., Trigonometric Series, Cambridge University Press, (1968).

This work has been supported by the Deutsche Forschungsgemeinschaft, in particular by the Sonderforschungsbereich 123 "Stochastische Mathematische Modelle".

DIE BEURTEILUNG VON VERGIFTUNGSFÄLLEN MITTELS

DISKRIMINANZANALYSE AM MODELLBEISPIEL DIGOXIN

R. Aderjan und W. Härdle

(Institut für Rechtsmedizin der Universität Heidelberg)

Über Digitalisvergiftungen wird relativ häufig berichtet, zuletzt in einer umfassenden Studie von Flasch und Flasch (1981), (1). Tödlich verlaufende Fälle finden sich in der Literatur nur selten und in weniger als 10 Arbeiten sind Blut- und Gewebekonzentrationsdaten genügend beschrieben. (1,2,4 - 6, 8, 9, 11, 12). Kommen unterschiedliche Meßmethodik, niedrige Dosierung und Körperkonzentrationen sowie geringe therapeutische Breite zusammen, wie bei Digoxin, so ist der forensische Toxikologe von jeher vor größte Probleme gestellt, wenn es einen Vergiftungsverdacht zu klären gilt.

Welcher Stellenwert ist Konzentrationsmessungen von Digoxin in Blut und Geweben zuzuweisen? Abb. 1 zeigt eine Zusammenfassung von klinischen Serumspiegelbestimmungen nach Rietbrock (1978, 10) aus der zu ersehen ist, daß sich die Digoxin-Konzentrationsbereiche, die bei therapeutischen- und toxischen Wirkungen festzustellen sind, weitgehend überschneiden. Bereits die 50-%ige Überschreitung des mehr willkürlich angenommenen Maximalwertes von 2 ng Digoxin pro Milliliter Serum, der die Obergrenze des therapeutischen Bereiches darstellen soll, führt zu einer Intoxikationshäufigkeit von 93 %. Bereits bei 2 ng pro Milliliter werden in 16 % der Fälle Intoxikationen beobachtet. Leider ist die zugrundeliegende Anzahl der Fälle nicht so deutlich ersichtlich, wie in der Studie von Storstein (1977, 13), die die Verteilung der Meßwerte und die anteilige prozentuale Häufigkeit intoxikierter Patienten in Abhängigkeit zur Serumkonzentration übersichtlich darstellt. (Abb. 2)

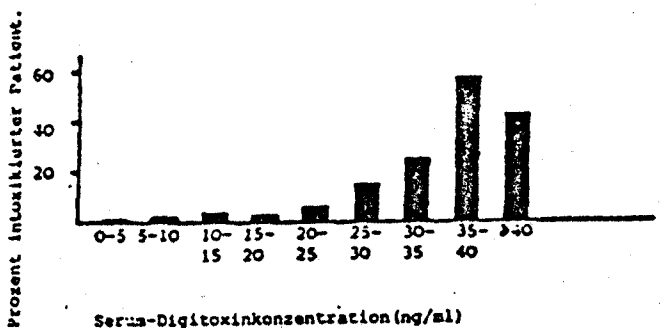
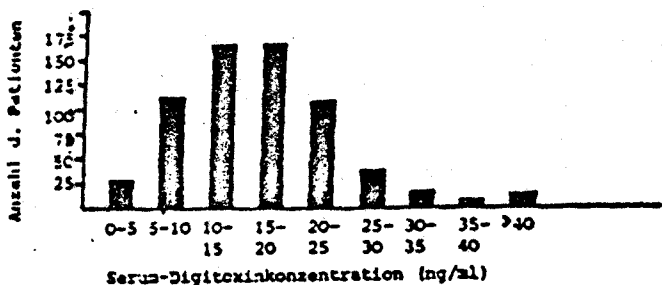
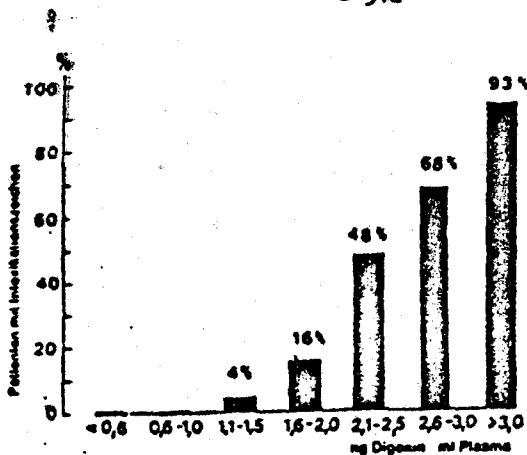
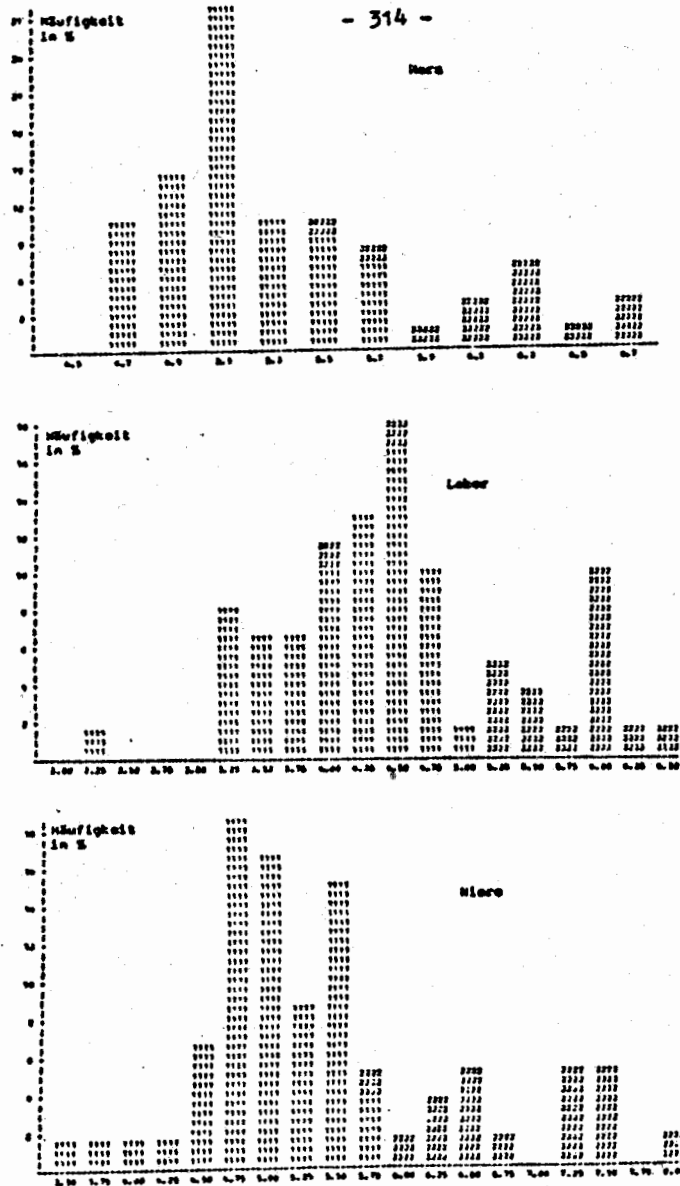


Abb. 4.2
 Häufigkeit der Digitalis-Intoxikation in Bezug auf die Serumkonzentration;
 obere Darstellung: Digoxin (n. Rietbrock 1978)
 mittlere u. untere Darstellung:
 Digoxin-Serumkonzentration von 649 unter therapeutischen Dosen befindlicher Patienten und Häufigkeit toxischer Glykosidwirkungen in den verschiedenen Konzentrationsbereichen (nach Storstein u. Mitarb. 1977).

Tabelle Nr. 1

Vergleich der Mittelwerte der Digoxin-Konzentrationen im Herz, Niere und Leber bei Patienten unter therapeutischen Dosen, bei Suiciden und bei den Vergiftungsfällen des Krankenhauses Rheinfelden mittels t-Test (bei der Ermittlung der Irrtumswahrscheinlichkeit im t-Test wurde berücksichtigt, daß die Varianzen der Stichproben nicht gleich sind) nach logarithmischer Transformation.

Gewebeart	Fallgruppe	Konzentration (ng/g, Mittelw. u. Standardabw.)	Irrtumswahrscheinlichkeit im t-Test <
Herz	Patienten	172 ± 1,3	0.0001
	Suicide	327 ± 2,4	
	Vergiftungen	439 ± 1,4	
Niere	Patienten	144 ± 1,6	0.0002
	Suicide	740 ± 1,6	
	Vergiftungen	830 ± 2,5	
Leber	Patienten	59 ± 1,7	0.0001
	Suicide	220 ± 1,8	
	Vergiftungen	319 ± 2,03	



Geht man von einer im Verteilungsgleichgewicht linearen Beziehung zwischen Blut- und Gewebekonzentrationen aus, so gilt für den Stellenwert gemessener Digoxin-Gewebekonzentrationen bei Vergiftungsverdacht ebenso, daß auf Grund von Bereichsüberschneidungen eine eindeutige Zuordnung eines Konzentrationswertes zu therapeutischen oder toxischen Glykosidwirkungen nur möglich ist, wenn toxische Wirkungen objektiviert sind.

Die Blut- und Gewebekonzentrationen von 45 therapeutisch mit Digoxin oder Beta-Methyldigoxin behandelten Patienten, die wir 1978 und 1979 untersuchten, (1) dienen als Vergleichskollektiv, wenn man sie den Daten von 13 Vergiftungsfällen des eigenen Untersuchungsgebietes gegenüberstellt (2). Bei 6 suicidalen- und 7 als homicidal zu betrachtenden Vergiftungen zeigt der Mittelwertvergleich der Konzentrationen in Herzmuskulatur, Leber und Nierengewebe nach logarithmischer Transformation (um symmetrische Verteilungen zu erzielen und nach Rücktransformation des log-Mittelwertes), daß sich die Kollektive hier signifikant unterscheiden. Für Skelettmuskulaturkonzentrationen und Gehirnkonzentrationen trifft dies nicht zu (Tab. 1) (1).

Wie klar im Grenzbereich der beiden Kollektive eine Zuordnung eines Organ-Meßwertes zu treffen ist, hängt davon ab, wie sehr sich die Konzentrationsbereiche überschneiden. Die Abb. 3 bis 5 zeigen die gemeinsame Verteilung der Digoxin-Konzentrationen der beiden Kollektive in Herz, Leber und Niere. In logarithmischem Maßstab aufgetragen, ist die zweigipflige Form zu erkennen.

Stellt man die Verteilungen entsprechend den Mittelwerten und der Standardabweichung normiert dar, so ergibt sich für die Herzmuskulatur, daß bei 257 ng pro Gramm im

Abb. 3-F
Darstellung der 2-gipfligen gemeinsamen Verteilungskurven nach logarithmischer Transformation der Konzentrationswerte von linksventrikulärer Herzmuskulatur (LHV), Lebergewebe (LE) und Nierenrinde (NR). 1 = Normalkollektiv, 2 = Vergiftungsfälle.

rechten Ventrikel sowie 289 ng pro Gramm im linken Ventrikel es mit 4,05 % bzw. 2,56 % gleich wahrscheinlich ist, daß ein beobachteter Meßwert einem der beiden Kollektive "toxisch" oder "therapeutisch" zuzuordnen ist. Bei der auf Grund unserer Beobachtungen (1) abgeleiteten Grenzkonzentrationen für den Beginn des toxischen Konzentrationsbereiches von 400 ng Digoxin pro Gramm Gewebe für den Herzmuskel ist die Wahrscheinlichkeit, daß nach therapeutischer Dosierung ein noch höherer Meßwert beobachtet wird nur noch 0,55 % bzw. 0,09 % (Abb. 6).

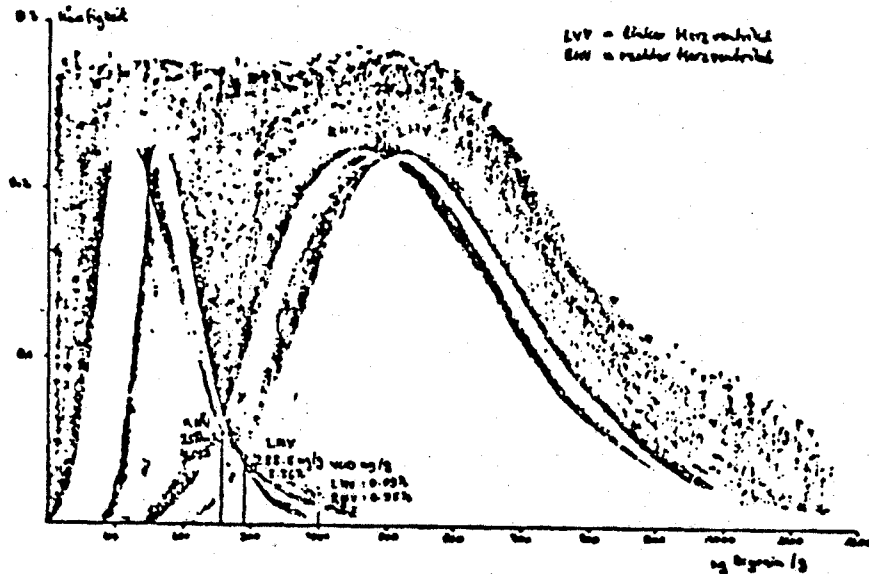


Abb. 6
Darstellung der standardisierten Verteilung der Digoxin-Konzentration in links-ventrikulärer (LHV)-rechtsventrikulärer (RHV) Herzmuskulatur nach therapeutischer Dosierung und nach toxischen Dosen berechnet aus Mittelwert und Standardabweichung nach logarithmischer Transformation. Die Schnittpunkte der Verteilungskurve zeigen die Konzentration an, die mit gleicher Wahrscheinlichkeit einem der beiden Kollektive zugeordnet werden kann. Die höhere Standardabweichung, die der toxischen Verteilungskurve zugrunde liegt, kann auch auf die stark unterschiedlichen toxischen, z.T. unbekanntem Dosen zurückzuführen sein.

	Mittelwert ± Standardabweichung der Organ-Konzentration in ng/g		Abgrenzungsschranke für den toxischen Konzentrationsbereich		Grenzwert, mg/g	Sicherheitszone
	nach therap. Dosierung n=45	nach toxischer Dosis n=6-12	ng/g	5 Wahrsch. Wahrsch. Schranke		
Vollblut	5,08 ± 2,47	40,7 ± 32,5	11	3,32	20	
Herzmuskel li. Ventrikel	178 ± 50,6	541 ± 155	209	2,56	400	
Herzmuskel re. Ventrikel	128 ± 58	485 ± 187	257	4,05	400	
Leber	67,7 ± 32,8	337 ± 191	115	11,3	250	
Nierenrinde	161 ± 72,1	1700 ± 766,9	318	5,7	500	
Nierenmark	162 ± 86,6	601 ± 199	329	6,7	500	
Skelettmuskel	30,8 ± 21,8	70,1 ± 31,6	43,4	18,7		- entfällt wegen zu starkem Streuungsbereich
Gehirn	26,8 ± 12,1	39,1 ± 29,1	26,4	57,9		- Überschneidung der Meßbereiche

Tabelle 2

Mittelwerte und Standardabweichungen der Digoxin-Konzentration in Körperflüssigkeiten und Geweben bei Patienten unter therapeutischen Dosen von D-Methylidigoxin und Digoxin (n = 45) sowie bei letalen Digoxinvergiftungen. Die nach logarithmischer Transformation der Einzelwerte abzufolgenden Verteilungen (s. Beispiel Herzmuskulatur Abb. 6) führen zu:

- einer Abgrenzungsschranke, dafür, daß es gleich wahrscheinlich ist, daß nach therapeutischer oder (letal-)toxischer Dosierung ein Meßwert zu beobachten ist. Die Wahrscheinlichkeit dafür, daß bei therapeutischer Dosierung ein Meßwert über dieser Schranke und dafür, daß nach toxischer Dosis ein Meßwert unter dieser Schranke zu beobachten ist, nimmt jeweils ab.
- der Wahrscheinlichkeit, mit der am Beginn des toxischen Konzentrationsbereiches (aufgrund der beobachteten Vergiftungsfälle und unter Berücksichtigung einer Sicherheitszone von 100 ng/g ab dem höchsten therapeutischen Meßwert) ein Meßwert als Folge einer therapeutischen Dosierung zu beobachten ist.

Die entsprechenden Daten für die Leber- und die Nierenkonzentrationen ergeben sich aus der Tabelle 2. Bei der Diskriminanzanalyse (7), einem statistischen Zuordnungsverfahren, werden die Kollektive mittels mehrerer Organparameter, mindestens 2, miteinander verglichen. Nimmt man beispielsweise Blut- und Nierengewebe oder Blut- und Lebergewebe, so entstehen durch zwei Parameter gebildete Gruppen (Abb. 7 und 8), die sich als elliptische Gebilde in der Fläche darstellen lassen. Deutlich erkennbar ist die verbesserte Unterscheidbarkeit der beiden Kollektive, die sich durch eine hyperbelähnliche Kurve voneinander abgrenzen lassen. Die Zuordnungswahrscheinlichkeit eines Falles ergibt sich aus dem Abstand zu dieser (nicht einge-

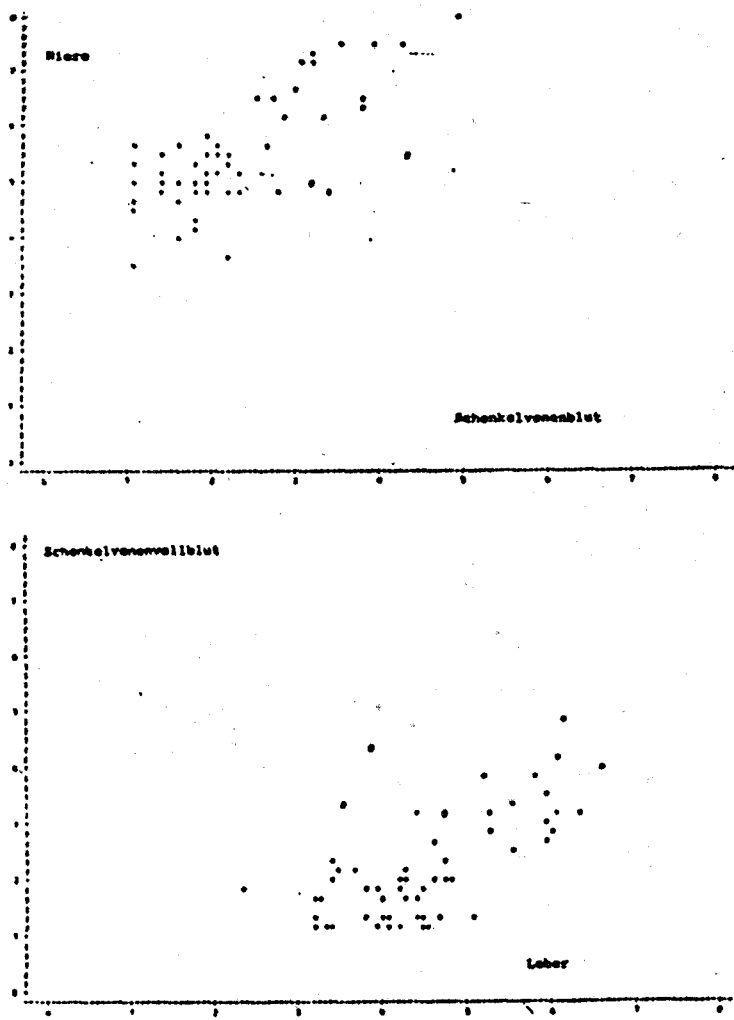


Abb. 3 und 4

Darstellung des Normalkollektivs und der Digoxinvergiftungen in den von den Parametern Lebergewebe und Schenkelvenenvenenblut sowie Nierengewebe und Schenkelvenenvenenblut auf gespanntem Raum.

- o) • Vergiftungen
- • Normalkollektiv
- o • Testfälle nach Literaturangaben

zeichneten) Grenzlinie. Bereits die 2-dimensionale Parameterkombination Leber und Schenkelvenenvenenblut (als Beispiel) bringt, neben der richtigen Einordnung jedes einzelnen Vergiftungsfalles für sich als neu genommen, eine 100 %ige Klassifikation literaturbekannter Konzentrationsdaten. (Obwohl deren unterschiedliche Untersuchungsmethodik einen direkten Vergleich der Meßwerte sicher nicht erlaubt).

Erhöht man die Parameterzahl auf 4, z. B. linker Herzventrikel, Leber, Niere und Schenkelvenenblut, so ist die Diskrimination nicht mehr graphisch darstellbar, da sie über das 3-dimensionale hinausgeht. Die Unterscheidungskraft wird jedoch so hoch, daß auch der Fall einer nur um knapp 1 Stunde überlebten Beta-Methyldigoxinvergiftung (Rietbrock 1978, 9) richtig zugeordnet wird. Ohne Berücksichtigung der Blutkonzentration von 75 ng/ml würde er fehlerhaft klassifiziert, da die Organkonzentrationen auf Grund der geringen Verteilungszeit noch nicht in einen einwandfrei als "toxisch" erkennbaren Konzentrationszustand gekommen waren.

Ich fasse zusammen:

Die Anwendung statistischer Methoden, insbesondere der Diskriminanzanalyse erlaubt eine Zuordnung von Digoxin-Blut- und Gewebekonzentrationen zu einem therapeutischen Vergleichskollektiv oder zu einem Kollektiv von Vergiftungsfällen. Die Unterscheidungskraft wird um so höher, je mehr Organparameter in das Verfahren einbezogen werden. Die Diskriminanzanalyse stellt den toxikologischen Beurteilungskriterien eine statistische Sicherung gegenüber, die um so besser ist, je größer die beobachteten Kollektive sind. Sollen die Ergebnisse verschiedener Laboratorien verwendet werden, so sind nur qualitätskontrollierte Meßwerte zweckdienlich.

ZUSAMMENFASSUNG

Der forensische Toxikologe steht bei seltenen Vergiftungen vor dem Problem, die quantitativen Analysebefunde seiner postmortalen Untersuchung an nur wenigen Vergiftungsfällen messen zu können. Die Grundlage für eine Beurteilung muß ohne entsprechende statistische Absicherung bleiben, solange nicht genügend Datenmaterial gesammelt werden konnte.

Bei Vergiftungen durch Arzneimittel kann ein geeigneter Weg dadurch eingeschlagen werden, daß die Blut- und Gewebekonzentrationen nach therapeutischer Dosierung denen nachgewiesener Vergiftungsfälle gegenübergestellt werden um so jeden neu vorkommenden Fall einer der beiden Einschätzungen "toxisch" oder "nicht toxisch" mit der entsprechenden Wahrscheinlichkeit zuordnen zu können.

Im Falle des klinisch häufig verordneten Herzglykosids Digoxin und seiner Derivate werden derartige Klassifikationen vorgenommen, indem von einem Kollektiv von 45 Patienten unter therapeutischen Dosen sowie von 13 Digoxin-Todesfällen die Konzentrationen in sektionstechnisch regelmäßig verfügbaren Körperflüssigkeiten und Organen unter Anwendung statistischer Methoden, insbesondere der Diskriminanzanalyse, miteinander verglichen werden.

LITERATURVERZEICHNIS

1. Aderjan, R.
Habilitationsschrift Heidelberg 1981
2. Arnold W., Püschel, K. (1979)
Toxikologische und morphologische Befunde bei Digoxinvergiftung in forensischer Sicht.
Z.Rechtsmed. 83, 265
3. Flasch, H. und Flasch, C. I. (1981)
Nebenwirkungshäufigkeit bei digitalisierten Patienten - Dokumentation und Analyse einer Literaturrecherche über Intoxikationsquoten.
Ärztl. Forsch. 29, 31
4. Iisalo, E., Nuutila, M. (1973)
Myocardial digoxin concentrations in fatal intoxications.
Lancet, 3 Febr., p. 257
5. Jelliffe, R. W. (1967)
Autopsy verification of suicide by digitalis. Report of a case with successful chemical identification of digitalis glycosides in gastric contents.
Am. J. Clin. Path. 47, 180
6. Larbig, D., Haasis, R., Kochsiek, K. (1978)
Die Glykosidkonzentration und ihre klinische Bedeutung.
Forum cardiologium 15, Boehringer Mannheim
7. Press, S.J. (1974)
Applied multivariate analysis.
Holt, Reinhart and Winston inc.
8. Reissell, P., Alha, A., Karjalainen, J., Nieminen, R., Ojala, K. (1975)
Digoxinintoxication determined post mortem.
Abstr. of the VIth Int. Congr. on Pharmacol. Helsinki, 386
9. Rietbrock, N., Wojahn, H. Weinmann, J. Hasford, J. Kuhlmann, J. (1978)
Tödlich verlaufene β -Methyldigoxin-Intoxikation in suicidal Absicht
Dtsch.Med.Wschr. 103, 1841
10. Rietbrock, N., Oeff, F., Martin, K., Kuhlmann, J.
Glykosidkonzentrationen im Plasma und Intoxikationshäufigkeit nach β -Methyldigoxin und β -Acetyldigoxin unter standardisierten Bedingungen.
Herz/Kreisl. 10. 267

11. Selesky, M., Spiehler, V., Cravey, R.H., Elliot, H.W. (1976).
Digoxin concentrations in fatal cases.
J. Forens. Sci. 22, 409
12. Steentoft, A. (1973)
Fatal digitalis poisoning.
Acta Pharmacol. et Toxicol. 32, 353
13. Storstein, O., Hansteen, V., Hatle, L., Hillestad, L.,
Storstein, L. (1977)
Studies on digitalis XIII: A prospective study of
649 patients on maintenance treatment with digitoxin
Am. Heart. J. 93, 434

I. INTRODUCTION .

The DACAPO plot package is designed to provide a rapid method for producing graphs, histograms and other graphical output. Usually "plotting" means the call of subroutines from the level of the data - generating program. The plot-procedure DACAPO is performed in a separate step, requiring no compiling or linkage time. This is always an advantage in large computer centers. The user can check the output data and misspecified values and enter corrections before he funnels it through a plot program.

Even "big" plots, i.e. with many movements of the pen, are produced in less than 10 seconds. A consequence of this is that the user runs quickly through the chain of incoming jobs, which pass the internal reader.

Any sequential data set in fixed length may serve as input to DACAPO. Even temporary data sets are allowed, if DACAPO is called in the data computing run.

DACAPO is available also in an interactive version, which permits the user to control his plot activities directly from the terminal. This feature is called by the command "DACAPO" from the TSO level.

To call DACAPO from batch level, one uses the following EXEC card

```
//PLOT EXEC DACAPO  
//S1.SYSIN DD *
```

```
-----  
control input to DACAPO  
-----
```

One may also give

```
//S1.SYSIN DD DSN=USERDSN,DISP=...
```

N is the number of the plot data set. DACAPO allocates SYSOUT(6). To display the plot on the TEKTRONIX 4014 one gives the command

"TEKT your job".

DACAPO is controlled by variables, which are described by the following symbols and conventions in this manual.

␣	Signifies BLANK-character
&	Signifies parameters of your choice
< >	Encloses default-parameters
	Separates alternative options, any one of which may be specified.
UPPERCASE	Information given in uppercase must be typed exactly as shown, although it may be entered with upper or lower case.
lowercase	Information given in lower case describes a parameter where one may type in characters or digits as desired.
	Meaning of characters:
a,b,c	a single character, or a string
n,n	integers
x,y,z	reals.

The arrangement of the plots on the plotter paper is handled automatically, insuring that no overruns occur. A frame is drawn surrounding each plot. The frame can also be defined directly by appropriate variables or by specifying a DIN format.

II. THE CONTROL VARIABLES

In this chapter the input to DACAPO is explained. Input takes place several steps of descending order. First the user defines general characteristics of the plot run. Then he descends one level by defining the type of the picture (scatter-plot for instance). Next, after specifying the scale, size, colour etc., he arrives at a "non-data" defining step, where he enters text strings.

STEP 1: Setup of general variables

=====

The first card in the input stream to DACAPO must be either a blank card or a control card which fixes general characteristics of all plots in the run.

GTIT|_____

If GTIT is given, DACAPO assumes that the user wants to have a title for all following plots.

&xcor,&ycor|<right uppermost edge of the picture>

These parameters center the title in the coordinates given in cm.

&scalti|<1.>

Scale-factor for the (title-)string to be drawn. The default size of the letters is an intermediate readable scale.

&scasym|<1.>

This parameter defines a scale-factor for the symbols to be plotted. It is relevant only for the supervisors HBOO, XYPL, because only this routines are able to draw symbols.

&scanu|<1.>

This parameter defines a scale-factor for the numbers to be plotted along the axis.

&text

In column 41-80, one enters the string &text for a title.

The first card has the following input format:

A4,6X,5F5.0,5x,20A2

!

Example:

GTIT 1. 9. MY TITLE
-----1-----;-----2-----;-----3-----;-----4-----;-----5-----;-----6-----;

Explanation:

A title is centered at the absolute coordinates (1.,9.) (in cm), with the default scale factors.

STEP 2 :Definition of supervisor and scaling

DACAPO has three supervisor routines.

HBOO

Plots histograms, graphs, step functions, and other graphs with constant increment on the x-axis.

XYPL

The most flexible routine, with the possibility of plotting confidence intervals and variable abscissas.

SCAT

Plots the data in a scatter diagram.

P3DI

Plots 3 dimensional objects

HBOO|XYPL|SCAT|P3DI|<____>

Specifies the routine; ____ (4 blanks) is
equivalent to HBOO

To scale by hand or automatically one uses:

AUTO|HAND|<____>

AUTO, which is equivalent to ____ , scales the
data automatically. HAND expects the min/max
values out of the control block \$CONTRO. AUTO and
SCAT together is not permitted.

Input format:

A4,6X,A4

Example:

HBOO HAND

-----1-----;-----2-----;-----3-----;-----4-----;-----5-----;-----6-----;

Explanation:

The routine HBOO, which is explained in chapter III controls the data. Scaling is done by hand.

STEP 3 : Definition of the characteristics of the plot
=====

In this step the user tells DACAPO in a control block, the physical characteristics of his plot, (e.g. size, scale and format options).

DDNAM=&n|<8> , &n =/ 5,6,7,98,99

DDNAM is the reference number to the ddname ftDDNAMft001 given in the JCL embedding of the DACAPO call.

DIN=&n|<5> , 2 <= abs(&n) <= 9

This variable defines a plot in DIN format, i.e. the outer frame line has exactly DIN format. If this variable is omitted, the plot size has to be defined by the values of AXLEN, AYLEN, XMIN, XMAX, YMIN, YMAX. If NEWPLO is set to 0, this variable may be omitted.
If &n is positive, the long side of the rectangle is horizontal; if negative, it is vertical.

COLOUR=<1>|2|3

COLOUR defines the colour of the actual plot:
1 = black
2 = blue
3 = red

NEWPLO=<1>|0

This variable tells whether the user wants a new frame or new scale. (1=yes, 0=no)

FORMAT=1|<0>

Each supervisor routine which was chosen in step 2 has a standard input format. To change this format one gives the value 1 to FORMAT.

LOGFL=1|<0>

Logarithmic scale on y-axis. (1 = yes , 0 = no)

LTEXT=<0>|1|2|3

0 = no labelling of x- or y-axis.
1 = only below the x-axis.
2 = only beside the y-axis.
3 = on both axis.

LWTEXT=<0>|&n

Any further text? &n is the number of strings. The position is give later on in step 4.

LINTYP=<0>|1|2|3|4|5

Determines the type of lines.
0 = step function.
1 = histogram
2 = piece-wise linear connection
3 = piece-wise linear with dashed lines. The dash length is controlled by the variable DASUL, which is initialized at .3 cm.
4 = curve through data points. (cubic spline interpolation, the smoothing parameter is given in a later step)
5 = symbols only.

HATCH=&x|<0.0>

To hatch the area between the graph and the x-axis, give HATCH a value not equal to zero. The hatching lines are plotted then in a distance given by &x (cm). This option should not be used together with SCAT or LINTYP=5 , because the graph together with the frames should be a closed area.

ANGLE=&x|<0.0>

Defines the angle of the hatching lines.

NTIM=&n|<1>

Defines how many times DACAPO should plot all lines. Repeated plotting generates broader and more visible lines, (good for xerox-copies).

NBSP=&n|<0>

This command shifts the data . &n greater than 0 means a backspacing of &n records; &n less than 0 means a read-forward of &n records.

LXSYM=&n|<5>

LXSYM can take every integer value.

This parameter defines the partition of the x-axis by tick marks and decimals, which indicate the scale values of the axis.

&n gt 0 &n ticks and decimal values are drawn.

&n eq 0 no ticks and no scale values are plotted.

&n lt 0 &n ticks, but no scale values are drawn.

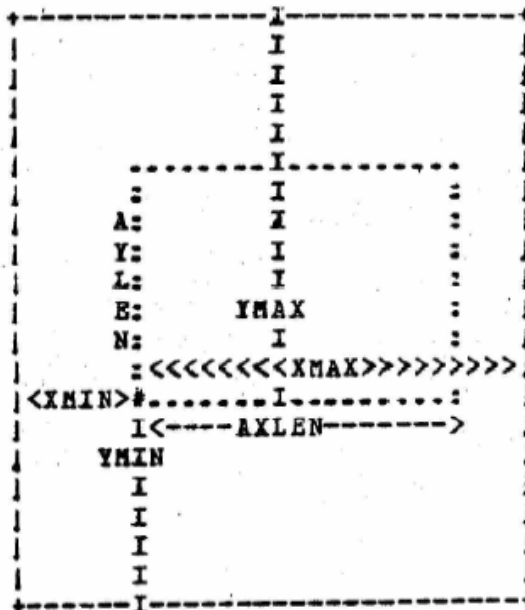
If &n is too great, the digits are drawn rectangular to the axis.

"tickmarks" : |----|----|----|----| (axis)
0. 2. 4. 6. 8. (datascale)

LYSYM=&n|<5>

LYSYM has the same meaning as LXSYM. It can take every integer value.

To understand the next variables look at the following picture.



----->>> # signifies the origin of the coordinate system <<---

AXLEN=&x|<DIN A5 value>

AYLEN=&x|<DIN A5 value>

XMIN =&x|<DIN A5 value>

XMAX =&x|<DIN A5 value>

YMIN =&x|<DIN A5 value>

YMAX =&x|<DIN A5 value>

AMINIX=&x|<0.0>|value from AUTO

IF scaling by hand was defined in step 2, &x must be specified; it is the minimum of the abscissas.

AMINIY=&y|<0.0>|value from AUTO

See AMINIX.

YSHIFT=&y|<0.0>

Shifts the data by &y in data coordinates in direction of y-axis.

XSHIFT=&x|<0.0>

Does the same as YSHIFT, but shifts in direction

of x-axis.

DASHL=&x|<0.3>

Defines the dashlength of the lines. This presupposes LINTYP=3. &x is the value in cm.

Example:

```
$CONTRO DIN=6, COLOUR=2, LTEXT=1,  
_FORMAT=1, LINTYP=3, $END
```

Explanation:

DACAPO generates a DIN A6 plot, which is drawn in blue. A title is put below the x-axis and the data comes in in a non-standard format (not the one of the supervisor). The data line is dashed with a length of 0.3 cm.

STEP 4 Text, format specifications and input for the supervisors
=====

If FORMAT is set to 1, one gives first the non-standard-format, which masks the data. The brackets are necessary.

Example:
(2F15.8)

Explanation:

DACAPO reads with format (2F15.8) and overrides the standard format of the previously defined supervisor.

If there were title flags in step 3, one gives title specifications in the following format.

```
&tit      &x  &y  &sc &ang      &abc  
-----1-----;-----2-----;-----3-----;-----4-----;-----5-----;-----6-----;  
(A4,6X,4F5.0,10X,20A2)
```

&tit=XTIT|YTIT

The flag shows whether the string is for the x-axis or the y-axis.

(&x,&y) |<centered below (beside) the axis>

The coordinates of the text

&sc|<1.>

The scale of the text

&ang|<.0>

The angle relative to the axis, specified by &tit.

&abc

The string

Example:

```

XTIT                                MYTEXT
-----1-----2-----3-----4-----5-----6-----;

```

Explanation:

The string MYTEXT is centered below the x-axis.

Example:

```

YTIT                                .5 MYTEXT ON Y-AXIS
-----1-----2-----3-----4-----5-----6-----;

```

Explanation:

Draws MYTEXT ON Y-AXIS centered beneath the y-axis, reduces size of the symbols to half of the standard height (computed from the length of the string and the axis).

For LWTEXT not equal to 0 the following card is necessary.

```

XYTI      &x      &y      &sc      &ang      &abc
-----1-----2-----3-----4-----5-----6-----;
(A4, 6X, 4F5.0, 10X, 20A2)

```

Example:

```

-----1-----2-----3-----4-----5-----6-----;
(4F20.7)
XYTI                                MY DATA
YTIT                                ARE PRETTY
XYTI      6.      -1.              DACAPO
XYTI      6.      -2.              PLOTS
XYTI      6.      -3.              MANY THINGS

```


Explanation:

Read with input format (4F20.7). The text "MY DATA" is centered below the x-axis, are "ARE PRETTY" beside the y-axis. Three strings are drawn in absolute positions (in cm) relative to the origin. For the definition of the origin see page 8.

STEP 5 : Control input to the supervisors
=====

For the supervisor HBOO one has the following input.

```
  &npts   &xinc   &xafv   &nsymb   &smooth  
-----1-----2-----3-----4-----5-----6-----;  
(I10,2F10.5,I10,f10.5)
```

&npts

The number of points

&xinc

The increment on the x-axis

&xafv

The starting point on the x-axis

&nsymb

The integer-equivalent of the symbol to be used

&smooth

The smoothing parameter for LINTYP=4.

For XYPL one enters only

```
  &npts   &smooth  
-----1-----2-----3-----4-----5-----6-----;  
(I10,f10.5)
```

&npts

The number of points to be plotted.

&smooth

The smoothing parameter for LINTYP=4.

For SCAT one must write

(3I10,4F10.5)

&npts

The number of points

&nchx

Number of categories in x

&nchy

Number of categories in y

&dx

The category size in x

&dy

The category size in y

&x1

The lowest value in x

&y1

The lowest value in y

Each x-y category which is not empty is represented by a cross.

To plot 3 dimensional data P3DI is the right routine.

```
&nx&ny  &xmin  &xmax  &ymin  &ycl  
-----1-----2-----3-----4-----5-----6-----  
-----&ymin  &zmin  &zmax  &ic1  
-----1-----2-----3-----4-----5-----6-----  
-----&x1  &z1  &za  &za  
-----1-----2-----3-----4-----5-----6-----  
(2I3,3e20.8/3e20.8,2I3/4e20.8)
```

&nx,&ny

The number of points in x, y direction

&xmin,&xmax

The extreme values in x direction

&ymin,&ymin

The extreme values in y direction

&zmin,&zmax

The extreme values in z direction

&lclu,&lclo = 1|2|3

The colour of the upper and the lower side of the plot.

&xl,&zl = &x|<0.0>

The length of the x, z-axis

&xa,&za = &x|<0.0>

The distance between the x and z-axis. The change of this variables generates another look at the picture.

The standard values generate a plot with suitable size.

Now we ask, why the name DACAPO? The reason is that the STEP 6 is exactly the same as the STEP 1 : one gives a control word, which defines a supervisor to produce another curve. How does this procedure stop? Remember "dacapo al fine". "fine" is simply given by

FINE.

III. THE SUPERVISOR ROUTINES

HBOO

This routine is constructed to draw histograms. The abscissas are computed out of $\&npts$, $\&xinc$, $\&xafv$ given in STEP 5. HBOO works only with equidistant grid. Input is given either by a nonstandard FORMAT option or in standard (8F10.5).

XYPL

This is the most powerful tool of DACAPO, but it needs a very specific input format. It plots points in every input sequence, is able to connect these piece-wise or to draw symbols only together with confidence intervals both in x-or y-direction.

The mask is given by

```
  &x          &y          &errx          &erry          &nsy
  -----1-----;-----2-----;-----3-----;-----4-----;-----5-----;-----6-----
(4E15.5,I10,30X)
```

($\&x, \&y$)

The coordinates of the points

($\&errx, \&erry$)

The confidence intervals (error crosses). A simple cross is drawn centered at ($\&x, \&y$) with length ($\&errx, \&erry$).

$\&nsymb$

The number of symbols if any wanted.

To produce only lines with XYPL, one fills this field up with blanks.

SCAT

Plots the data as isolated points. The data must be organized pairwise; thus on standard input (8F10.5) one has 4 pairs of abscissas and ordinates on one input card.

P3DI

Plots the data in a 3 dimensional figure. Only the z-values are to be given on input (standard format 8F10.5). The x, y values are computed out of the control variables for P3DI, which are supplied by the user in step 5.

Example of a full DACAPO job.

```
-----1-----;-----2-----;-----3-----;-----4-----;-----5-----;-----6-----;
GTIT                                     TITEL RECHTS OBEN
HBOO      AUTO
$CONTRO DDNAM=11,DIN=-4,COLOUR=1,NEWPLO=1,FORMAT=0,
LOGPL=0,LTEXT= 0,LWTEXT= 0,LINTYP= 2,YSHIFT=  0.0
DASHL=  0.30,NTIM= 1,LAXSYM= 0,HATCH=  0.0 ,ANGLE=  0.0 , $END
      13  1.00000  1.00000
FINE
-----1-----;-----2-----;-----3-----;-----4-----;-----5-----;-----6-----;
```

Explanation

A title is put on the right uppermost edge, the supervisor HBOO controls the input, which arrives in standard format (8F10.5) from unit 11. 13 points are connected by a simple polygonal curve, beginning with the data-point 1.0 on the x-axis, stepping forward with an increment of 1. The colour of the pen is black.

IV. INTERACTIVE VERSION

DACAPO performs the plot procedure in a separate step controlled by the variables explained in chapter II, III.

If the user does not want to program DACAPO by the variables, he

should use the interactive feature, to produce the plot directly from his terminal. By stepping through a prompter he generates the control input, which is stored in a data set INTER.

This data set should be pre-allocated by the user.

To call the prompter he enters simply the command DACAPO under TSO/mvs.

To plot from the terminal, the system needs allocations to the

data sets to be plotted. (In a batch call this corresponds to the allocations by DD-cards).

The file names are FT&iP001 , &i not equal to 5,6,7,98,99.

STEP I1 : Allocation of files to the user's session

=====

The second step is to define the general title if any wanted.

STEP I2 : Definition of a general title

=====

The third step tells DACAPO which one of the allocated data sets of STEP I1 it should use.

STEP I3 : Allocations to DACAPO
=====

Now the same things as above occur.
(see example below)

STEP I4 : Definition of the supervisor
=====

STEP I5 : Generation of the controlblock
=====

STEP I6 : "non data" input
=====

Example of a DACAPO session under TSO/avs. The dataset INTER, which contains the variables for DACAPO, was filled up in this session.

*** THIS IS THE TSOOUT STREAM ***

READY
dacapo
DATUM : 07.09
UHRZEIT : 17.02

BITTE GELEN SIE NUN IHRE ALLOKIERUNGEN UEBER FILES "FT@F001". SIE BRAUCHEN NUR DIE ZIFFER @@ ANGEBEN. DIE ZIFFERN @@ KOENNEN SIE BIS AUF DIE EINSCHRAENKUNG

| 1 <= @@ <= 97, @@ UNGLEICH 01,05,06,07 |

FREI WAEHLLEN. BITTE BEACHTEN SIE, DASS Z.B. 8 ALS 08 GESCHRIEBEN WERDEN MUSS.

-> ERSTE EINGABE: @@ ,WO STANDARD=08, DEFINIERT DEN DD-NAMEN FT@F001 (1<=@<=97,@ UNGLEICH 01,05,06,07)

-> ZWEITE EINGABE: IHRE DATEI(EN)

FILE-NUMMER ? :

YOUR DATA ? :
data(oostudy)
WOLLEN SIE NOCH MEHR DATEIEN ALLOKIEREN? (Y/N)

n
WOLLEN SIE ZUERST INTERAKTIV IHRE DACAPO-STEUERKARTEN ERSTELLEN ? (Y/N)
Y

***** D A C A P O *****
** VERSION 3.5 **
** APRIL 1980 **

-->> JETZT IST DIE NEUE OVERLAY VERSION DRIN !!!!
\$\$\$\$\$\$\$\$\$\$\$\$\$\$\$\$ HOT NEWS \$\$\$\$\$\$\$\$\$\$\$\$\$\$\$\$\$
| -- DIE NEUE OVERLAYSTRUCTUR ERFORDERT NUR NOCH DIE |
| UEBLICHEN 256 K BYTE , UND DAS BEI ERWEITERTEN |
| LEISTUNGSANGEBOT!!! (SIEHE LINTYP =4) |
| -- AB 9.6.80 STEHEN UEBER LINTYP=4 SPLINE-SMOOTHING |
| ROUTINEN ZUR VERFUEGUNG. (REINSCH 1967) |
| -->> SKRIPT DER NEUESTEN VERSION & NEWS DURCH |
| -->> EX 'U.H40.ALL.DACA.CLIST' |
\$

DER INTERACTIVE DIALOG, DEN SIE NUN MIT DACAPO FUEHREN, IST SO ANGELEGT, DASS SIE ZUGLEICH DIE VARIABLENNAMEN ERLERNEN, DIE DER DIALOG-JOB FUER SIE IN EINEN DATENSTAZ "INTER" SCHREIBT. DIESE SIND IN ECKIGEN KLAMMERN ANGEGBEN Z.B. <LINTYP>. DIES

ERLEICHTERT IHNEN DIE VERBESSERUNG VON "INTER", UM DACAPO
VERAENDERT ZU STARTEN.
STANDARD WERTE DIE DACAPO FUER SIE EINSETZT, WENN SIE DIE
BLANKTASTE DRUECKEN SIND DURCH STERNE GEKENNZEICHNET.
Z.B. * AUTOMATISCH *
BITTE SCHREIBEN SIE FEHLER IN DIE DATEL
U.H40.ALL.DACFEL(&UID)
-->> DURCH EINGABE VON "Q" VERLASSEN SIE DEN DIALOG UND KOMME!
SOGLEICH IN DEN PLOTSTEP. EINGABE VON " " = "N" = " NEIN".
... ENTERTASTE DRUECKEN

!
WOLLEN SIE EINEN TITEL FUER ALLE PLOTS? < GTIT > ##
DANN DEN TITEL BITTE EINGEBEN. (MAX. 40 ZEICHEN)

brooks2/preprint2

FILENUMMER DIESER DATENSATZES? ## * 8 * < DDNAM >

DEFINITION DES SUPERVISORS . * H * < ART >

* H * FUER HISTOGRAMME UND KURVEN MIT
KONSTANTEM INCREMENT AUF DER X-ACHSE
X FUER VORLIEGENDE X UND (!) Y-WERTE
VGL. PLOTMANUAL "DACAPO"
S FUER SCATTERPLOTS
D FUER PLOTS VON 3 DIMENSIONALEN
OBJEKTEN.

WELCHE ROUTINE (H=*/X/S/D?) SOLL DEN PLOT STEUERN?

SKALIERUNG ? < SKAL >

* AUTOMATISCH * ODER VON HAND? (A=*/H)

h

DIN FORMAT ? * 5 * < DIN >

-9,....,-2 FUER EINEN DIN PLOT IN HOCHFORMAT
0 FUER EIGENE BESTIMMUNG DER RAHMEN
2,..,*5*,..,9 FUER EINEN DIN PLOT IN QUERFORMAT

-4

FARBE DES PLOTS? ## * 1 * < COLOUR >

SCHWARZ= * 1 * , BLAU = 2 , ROT = 3

SIND DIE DATEN N I C H T IM STANDARDFORMAT? (Y/N=*)

STANDARDFORMATE DER SUPERVISOR:

HBOO = H (8F10.5)
XYPL = X (4E15.5,I10,10X) < FORMAT >
SCAT = S (4(2F10.5))
P3DI = D (8F10.5)

LOGARITHMISCHE SKALA? * 0 * < LOGFL >

1 LOG. SKALA NUR AUF DER X- ACHSE
2 LOG. SKALA NUR AUF DER Y- ACHSE
3 LOG. SKALA AUF BEIDEN ACHSEN
* 0 * LINEARE SKALA

BESCHRIFTUNG DER ACHSEN? * 0 * < LTEXT >

* 0 * KEINE BESCHRIFTUNG
1 < LTEXT> FUER X-ACHSEN BESCHRIFTUNG

```

2          FUER Y-ACHSEN BESCHRIFTUNG
3          FUER BESCHRIFTUNG DER X & Y-ACHSE
3
## WEITERE TEXTE ZEICHNEN ? * 0 * < LWTEXT > ##
* 0 *      KEINEN WEITEREN TEXT
&N        ANZAHL DER TEXTE -

## ZEICHENART ? * 2 * < LINTYP > ##
0          STUFENVERBUNDENE LINIE
1          HISTOGRAMM
* 2 *      POLYGONZUG
3          UNTERBROCHENE LINIE
4          GEGLAETTETE DATEN (CUB. SPLINES)
5          ES SOLLEN NUR SYMBOLE GEZEICHNET
          WERDEN, KEINE VERBINDUNG DER DATEN

4
## EIGENSKALIERUNG < AMINIX , AMAXIX > ##

GEBEN SIE NUN DIE MAXIMA UND MINIMA AUF DEN ACHSEN AN:
MINIMUM , MAXIMUM AUF X-ACHSE?
?
0. .2
## MIN, MAX AUF Y-ACHSE? < AMINIY , AMAXIY > ##
?
0. .5
## SCHRAFFIERUNG DES PLOTS ? < HATCH , ANGLE > ##
VORSICHT! DER PLOT MUSS EINE EINZIGE GESCHLOSSENE FLAECHE
BILDEN. (Y/N)

## VERSCHIEBUNG DER DATEN IN Y-RICHTUNG? (Y/N) ##

## DEFINIEREN SIE BITTE IHRE X-ACHSENUNTERTEILUNG ##
## WIEVIELE ZAHLEN SOLLEN AN DER ACHSE STEHEN? ##
* 5 * ODER &N , &N EINE INTEGERZAHL
          NEGATIVE WERTE:
          < LXSIM >   AUF DER GEGENUEBERLIEGENDEN SEITE
DER ACHSE
          WERDEN KEINE TICKMARKEN GEZEICHNET
0        OHNE TICKMARKEN UND OHNE SKALIERUNGSWERTE
--> TICKMARKEN: |.....|.....|.....|.....| (AXIS)
--> DEZIMALEN : .1   .25  .5   .75  1.

## DEFINIEREN SIE BITTE IHRE Y-ACHSENUNTERTEILUNG ##
## WIEVIELE ZAHLEN SOLLEN AN DER ACHSE STEHEN? ##
* 5 * ODER &N , &N EINE INTEGERZAHL
          NEGATIVE WERTE:
          < LYSIM >   AUF DER GEGENUEBERLIEGENDEN SEITE DER ACHSE
          WERDEN KEINE TICKMARKEN GEZEICHNET
0        OHNE TICKMARKEN UND OHNE SKALIERUNGSWERTE
--> TICKMARKEN: |.....|.....|.....|.....| (AXIS)
--> DEZIMALEN : .1   .25  .5   .75  1.

## MEHRERE MALE AUSZEICHNEN? * 1 * < NTIM > ##
* 1 *     EIN EINZIGES MAL

```

```
&N      MEHRERE MALE. ES ENTSTEHEN BESSER KOPIERBARE LINIEN

## WIE LAUTET DER TITEL FUER DIE X-ACHSE? ## < XTIT >
x titel
## DER TITEL FUER DIE Y-ACHSE? ## < YTIT >
y titel
## SUPERVISOR H B O O ##
WOLLEN SIE DIE HBOO OPTIONEN VOM VORANGEGANGENEN
PLOT UEBERNEHMEN?(Y/N)

!
ANZAHL DER DATENPUNKTE, INCREMENT AUF X-ACHSE
UND DEN ANFANGSWERT EINGEBEN:< &NPTS,&XINC,&XAPW>
?
21 .01 0.
BITTE DEN SMOOTHING PARAMETER FUER DIE SPLINE
SMOOTHING ROUTINE EINGEBEN < &SMOOTH >
?
0.0006
## SOLL NOCH EINMAL GEPLOTTET WERDEN? < NEWPLO > ##
      A   JA UND ZWAR SOLL DIE NEUE KURVE IN DEN EBEN
          CREATIERTEN PLOT EINGEZEICHNET WERDEN.
      Y   JA, ES SOLL EIN NEUER PLOT, D.H. NEUER RAHMEN
          SKALIERUNG ETC. GEZEICHNET WERDEN.
      * N * NEIN

.... JETZT WIRD GEPLOTTET.
DER STANDARDNAME DER PLOTDATEIEN IST: PLOT.N@@,
MIT EINER ZWEISTELLIGEN GANZEN ZAHL @@.
WELCHE NR."@" SOLL DIE PLOTDATEI HABEN ? :
45
WOLLEN SIE EIN PROTOKOLL DES PLOTS SEHEN ?
(DRUCKER=A, TERMINAL=LEEREINGABE, DUMMY=D)

*****
*          DACAPO          *
*      VERSION  3.5      *
*      APRIL   1980     *
*****

=====
ALLE PLOTS HABEN DEN GENERALTITEL :
BROOKS2/PREPRINT2
=====

-- DER SKALIERUNGSFAKTOR FUER
   DIE SYMBOLE IST  1.00

-- SIE HABEN FOLGENDE PLOTSTEUERKARTEN EINGEGEBEN

XMIN=   .0              XMAX=   .0
YMIN=   .0              YMAX=   .0
AXLEN=  .0              AYLEN=  .0
XSHIFT= .0              YSHIFT= .0
```

```

AMINIX= .0          AMAXIX= .20000
AMINIY= .0          AMAXIY= .50000
ANGLE= .0           HATCH= .0
DASHL= .30000      DDNAM= 8
NBSP= 0             DIN= -4
NTIM= 1             COLOUR= 1
LTEXT= 3            LWTEXT= 0
LINTYP= 4           FORMAT= 0
LKSYM= 5            LYSYM= 5
LOGFL= 0            NEWPLO= 1
    
```

----- 1. PLOT, 1. KURVE-----

-- DIESER PLOT WIRD VON "HBOO" ERZEUGT

-- DIE AchSENEINTEILUNG WIRD VON "HAND" KONTROLLIERT

```

XTIT      0.0  0.0  0.0  0.0      X TITEL
YTIT      0.0  0.0  0.0  0.0      Y TITEL
    
```

*** PLOT JES2.TSU07132. LAENGE IN X : 22.00 CM ***
 *** AEV : 42 AEV-COSTS : 0.42 DM ***

IKJ562111 JOB F95 (TSU07132) EXECUTING

WENN SIE JETZT AN EINEM INTERAKTIVEM GRAPHISCHEN BILDSCHIRM
 SITZEN, KOENNEN SIE DURCH EINGABE VON

```

      TEK      TEKTRONIX 4014
      IBM      TEKTRONIX 618 + IBM 3277
    
```

SOGLEICH IHREN ERZEUGTEN PLOT ANSCHAUEN.

ibm

\$CONTRO

```

controlblock . . . . . 5
AMINIX . . . . . 9
AMINIY . . . . . 9
ANGLE
hatching angle . . . . . 7
AUTO
automatical scaling . . . . . 5
AXLEN . . . . . 9
AYLEN . . . . . 9
COLOUR . . . . . 6
DACAPO
command uader TSO . . . . . 19
from batch-level . . . . . 1
from TSO-level . . . . . 1
PROCLIB-procedure to call DACAPO . . . . . 1
DASHL
dash length . . . . . 10
DD-cards
corresponding allocations . . . . . 19
DDNAM
    
```

files FT&NF001	5
reference to dnames	5
the unitnumber of the data file	16
DIN	
DIN-format	6
equidistant grid	16
FORMAT	
standard input format	7
graphic output	1
GTIT	
general title	3
HAND	
scaling by hand	5
HATCH	
hatching	7
HBOO	13, 16
histograms	4
histograms	16
INTER	
data set for control variables	21
prompting the variables	19
storing control variables	19
INTERACTIVE VERSION	19
LINTYP	
type of lines	7
LOGFL	
logarithmic scale	7
LTEXT	
Text on axis	7
LWTEXT	
positioning strings	7
LXSYM	
defining the partition of the x-axis	8
defining the x-axis scale values	8
LYSYM	
defining the partition of the y-axis	8
defining the y-axis scale values	8
NBSP	
backspacing	8
read forward	8
NEWPLO	
new frame	6
new scale	7
nonstandard FORMAT option	16
NTIM	
broader lines	7
output SYSOUT(G)	1
P3DI	4, 14, 16
Scale factors	
numbers along the axis	3
symbols	3
title string	3
SCAT	4, 13, 16
STEP I1	
Allocation of files to the user's session	19
STEP I2	
Definition of a general title	19
STEP I3	
Allocations to DACAPO	20

STEP 14	Definition of the supervisor	20
STEP 15	Generation of the controlblock	20
STEP 16	"non data" input	20
STEP 1	Setup of general variables	3
STEP 2	Definition of supervisor and scaling	4
STEP 3	Definition of the characteristics of the plot	6
STEP 4	Text, format options and input for the supervisors	10
STEP 5	Control input to the supervisors	13
	supervisor routines	4
	symbols and conventions	2
	TEKTRONIX 4014	2
	THE CONTROL VARIABLES	3
	THE SUPERVISOR ROUTINES	
	how they work	16
	IMAX	9
	IMIN	9
	XSHIPT	9
	ITIT	
	x-titles	11
	XYPL	13, 16
	confidence regions	4
	supervisor for x,y-values	4
	IYTI	
	additional text	11
	YMAX	9
	YMIN	9
	YSHIPT	9
	YTIT	
	y-titles	11

Experimentelle Untersuchungen zum Verlauf der Alkoholkurve in der späten Eliminationsphase

R.Mattern, J.Bösche, K.Birk und W.Härdle

Zusammenfassung

Auf der Grundlage von 24 kontrollierten Trinkversuchen wurde die späte Phase der Alkoholelimination im Bereich unter 0,3‰ untersucht. Probanden im Alter von 22-33 Jahren tranken in 30 min 0,77 g Alkohol/kg Körpergewicht in Form von Cognac. Die Gipfelkonzentrationen lagen 2 h nach Trinkbeginn im Mittel bei 0,71‰, es erfolgten 9 Blutentnahmen, davon 5 unterhalb von 0,3‰. Oberhalb einer Blutalkoholkonzentration von 0,1‰ waren im Mittel Stundenabbauwerte von 0,137‰/h zu beobachten, die Streubereiche reichten von 0,103‰-0,207‰, höhere β_0 -Werte als 0,15‰ kamen nur in 2 Fällen als "Ausreißerwerte" vor. Die Meßdaten ließen sich bis zu einer Blutalkoholkonzentration von 0,1‰ mathematisch am besten durch eine lineare Funktion, unterhalb dieses Wertes durch eine exponentielle Funktion beschreiben. Der Umschlagpunkt zwischen beiden Bereichen lag im Mittel bei 0,052‰ mit 95% Konfidenzintervallen von 0,033 bis 0,070‰.

Summary

On the basis of 24 controlled drinking tests, the late phase of alcohol elimination in the range below 0,3‰ was investigated. Test persons of 22-33 years of age drank 0.77 g alcohol (cognac)/kg body weight in 30 min. Two hours after the start of drinking the mean peak concentrations amounted to 0.71‰; nine blood tests were then carried out; five of them gave values below 0.3‰. With a blood alcohol concentration above 0.1‰, mean reduction values of 0.137‰/h were observed; the range was 0.103‰-0.207‰; higher β_0 -values than 0.15‰ only occurred in two cases as "run-away values." The measurements up to a blood alcohol concentration of 0.1‰ were described mathematically best by a linear function, and below this value by an exponential function. The mean turning point between both ranges was 0.052‰ with 95% confidence intervals of 0.033‰-0.070‰.

Einleitung

Bei der Begutachtung von Trunkenheitsdelikten im Straßenverkehr spielt in der forensischen Praxis der Endbereich der Alkoholausscheidung eine verhältnismäßig geringe Rolle. Auffällig gewordene Kraftfahrer mit Blutalkoholkonzentrationen unter 0,5‰ bieten dem Polizeibeamten selten ein alkoholverdächtig erscheinendes Bild; die negativ verlaufende Atemalkoholprüfung begründet in solchen Fällen oft den polizeilichen Verzicht auf eine Blutentnahme. Dies mag bei Eintreffen der Polizei relativ kurz nach einem Vorfall keine wesentliche Beeinträchtigung der Beweislage darstellen - der Nachweis einer allein durch Alkohol verursachten Fahruntüchtigkeit ist bei derart niedrigen Konzentrationen ohnehin bekanntermaßen problematisch. Wenn der in Verdacht geratene Kraftfahrer jedoch erst viele Stunden nach dem Vorfall kontrolliert werden kann, sollten trotz negativem Erscheinungsbild und fehlendem Atemalkoholnachweis mindestens eine, besser zwei Blutproben gesichert werden. Wird nun bei derartigen Blutproben Alkohol im untersten Bereich, etwa nur um 0,1‰

nachgewiesen, so stellt sich für den Gutachter die Frage, ob, und wenn ja, wie eine Rückrechnung auf der Basis wissenschaftlicher Erkenntnisse durchgeführt werden kann.

Fallschilderung

Ein Fall aus der eigenen Begutachtungspraxis, zu dessen forensischer Abwicklung mehrere Sachverständige mit verschiedenen Auffassungen in 3 Instanzen und im Wiederaufnahmeverfahren tätig waren, gab uns Anlaß, die frühere Lehrmeinung zu überprüfen, wonach auf der Grundlage von Werten unterhalb 0,15 bis 0,2‰ nicht zurückgerechnet werden sollte.

Ein an Hyperurikämie und Leberparenchymschaden leidender Polizeibeamter mittleren Lebensalters verursachte mit seinem Dienstkraftfahrzeug einen Verkehrsunfall, beging Unfallflucht und wurde erst mehrere Stunden später gefunden. Unfallzeugen hatten deutliche Trunkenheitssymptome beobachtet. Es wurde eine Doppelblutentnahme durchgeführt, nachdem ein nach Umfang, Getränkeart und Trinkzeit spezifizierter Nachtrunk angegeben worden war.

Die Beweislage erlaubte keinen Ausschluß des Nachtrunkes; unter Zugrundelegung eines Abbauwertes von 0,1‰/h war zu berechnen, daß die als Nachtrunk aufgenommene Alkoholmenge im Zeitpunkt der ersten Blutentnahme bereits eliminiert war, so daß der noch nachgewiesene Blutalkohol als Beweis für eine vor dem Unfall vorhandene Alkoholbeeinflussung angesehen werden mußte.

Strittig war nun die Frage, ob einerseits für den Abbau des Nachtrunkes die Annahme eines β_{60} -Wertes von 0,1‰ berechtigt war, zum anderen, ob die gemessenen Blutalkoholkonzentrationen als Grundlage einer Rückrechnung dienen durften.

Die Blutalkoholkonzentrationen der im zeitlichen Abstand von 45 min vorgenommenen Blutentnahmen betragen gaschromatographisch in der ersten Probe 0,08‰, in der zweiten 0,03‰ im Mittel. Die entsprechenden fermentchemischen Werte lagen bei 0,10‰ und 0,04‰.

Aus dieser Konstellation der Blutalkoholwerte war zu schließen, daß hier offenbar gerade die Endphase der Alkoholausscheidung erfaßt worden war.

Eine Rückrechnung schien uns geboten, zumal an der Präzision, insbesondere der gaschromatographischen Messung und der beobachteten Differenz von 0,06‰ in 45 min, kaum Zweifel bestanden. Die Deutung der gemessenen Blutalkoholkonzentrationen als "endogener Alkohol" schied aus, nachdem selbst bei stoffwechselkranken Probanden im Nüchternblut allenfalls Tausendstel-Promille endogenen Ethanols zu erwarten sind (Sprung et al. 1981).

Problemstellung

Bei der Frage, welcher Abbaufaktor zugrunde zu legen war, hätte man im vorliegenden Fall rein rechnerisch aus der Doppelblutentnahme einen stündlichen Abbauwert von 0,08‰ ableiten können. Dem stand entgegen, daß nach gängiger Auffassung aus einem einzigen gemessenen Abbauwert nicht auf die individuelle Abbauleistung geschlossen werden sollte.

Im Schrifttum findet man vergleichsweise wenige, meist ältere experimentelle Daten, die sich als Grundlage für solche Berechnungsprobleme

in der späten Eliminationsphase anbieten (z.B. Wille u. Steigleder 1966; Scheer 1967; Wolf u. Wiens 1982). Eine für forensische Zwecke vorgeschlagene Methode fanden wir in der Dissertation von Klepsch (1969), die auch Forster u. Joachim in ihrem Band "Blutalkohol und Straftat" (1975) zitieren. Klepsch wies darauf hin, daß die Größe von β_0 zum Teil als Funktion der Blutalkoholkonzentration beschrieben werden könne (Abb. 2). Er schlug daher eine "gestaffelte Rückrechnung" vor, wonach von Bereichen unter 0,5‰ an β_0 -Werte von weniger als 0,1‰ einzusetzen sind, falls Rückrechnungen auf eine Tatzeitmindestkonzentration gefordert werden. Für den Konzentrationsbereich 0,2‰-0,1‰ gab er sogar stündliche Eliminationsraten von nur 0,05‰ an, Werte, die uns sehr niedrig vorkamen.

Eigene Untersuchungen

In einer Versuchsreihe erhielten 24 klinisch gesunde Versuchspersonen im Alter von 22-33 Jahren, darunter 7 weibliche Probanden, in der frühen Nachmittagszeit nach vorangegangener 4stündiger Nahrungskarenz 0,77 g Alkohol/kg KG in Form von Cognac. Die Leberenzyme der Versuchspersonen lagen im Normbereich. Die Trinkzeit betrug max. 30 min. 90 min nach Trinkende erfolgte die erste Blutprobe. Zu diesem Zeitpunkt lagen die Blutalkoholwerte im Durchschnitt bei 0,71‰.

Die weiteren Blutentnahmezeiten wurden individuell über Kontrollen durch Atemalkoholbestimmungen mit dem Infrarotmeßgerät Atalmer so festgelegt, daß 5 der insgesamt 9 Blutproben in die Endphase unter 0,3‰ fielen. In unmittelbarem Anschluß an die Blutentnahme wurde nach Möglichkeit eine Urinprobe gesichert. Alle Blutentnahmen erfolgten, wie in der Praxis, mit Ventülen aus Cubitalvenen der rechten und linken Seite im Wechsel (Zink u. Blauth 1982). Die Alkoholkonzentrationen wurden gaschromatographisch durch Doppelbestimmungen im Serum mit dem Multifract F 40 der Firma Perkin-Elmer gemessen.

Ergebnisse und Diskussion

Die graphische Auswertung der Versuchsergebnisse (Abb. 1) bestätigte, im Gegensatz etwa zu Rietbrock u. Abshagen (1971), die im neueren Schrifttum hinreichend bekannte Tatsache, daß der Konzentrationszeitverlauf der Alkoholkurve in der späten Eliminationsphase erkennbar nicht linear verläuft. (z.B. Wagner et al. 1976; Koppun u. Propping 1977; Wilkinson et al. 1980) (Abb. 1). Die Stundenabbauwerte im sog. linearen Bereich lagen zwischen 0,103 und 0,207‰, wobei 0,15‰ nur in 2 Fällen überschritten wurde. Der Mittelwert der gesamten Gruppe war in diesem Bereich mit 0,137‰ bemerkenswert niedrig, erhärtet aber die Befunde von Zink (1982) und bestätigt die Ergebnisse früherer Trinkversuche (z.B. Widmark 1932; Kulpe u. Mallach 1960; Forster et al. 1961; Springer 1972). Aber auch mit der durchschnittlichen Umsatzkapazität des Körpers bei Alkoholinfusion unter Bedingungen des Fließgleichgewichts stimmen diese Stundenabbauwerte gut überein (Förster u. Hartmann 1980).

Das Hauptinteresse galt nun dem Konzentrationsbereich, in dem sich der Übergang der linearen in eine exponentielle Elimination so deutlich vollzog, daß eine Änderung des β_0 -Wertes abzuleiten war. Wir untersuchten deshalb die Abhängigkeit der β_0 -Werte von der Blutalkoholkonzentration (Abb. 2). Hierzu wurden aus den Meßdaten der Blutalkoholbestimmungen Konzentrationsgruppen von 0,1‰ Breite gebildet. Die in diesen Gruppen beobachteten β_0 -Werte wurden arithmetisch gemittelt und mit ihren Streubereichen aufgetragen. Die Darstellung zeigt, daß sogar bis in den Bereich von 0,1‰ hinab die stündlichen Abbauwerte

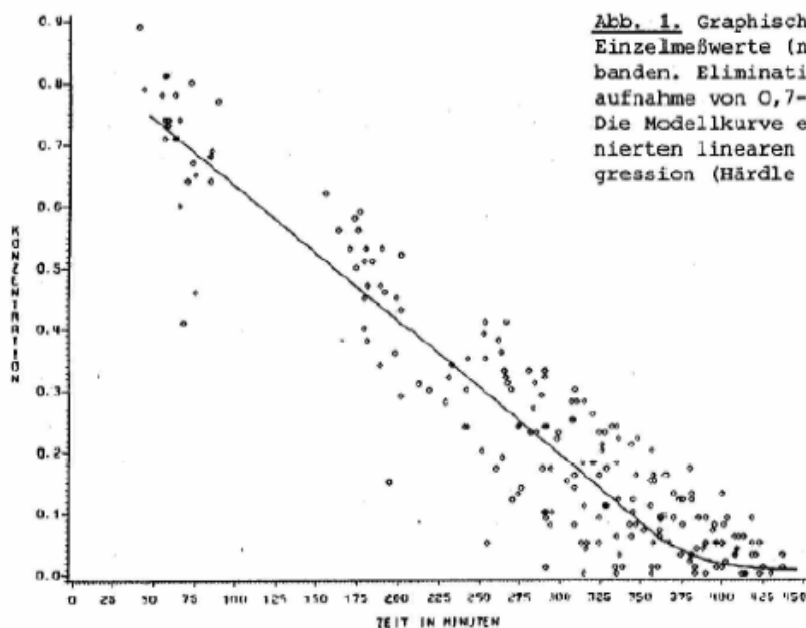


Abb. 1. Graphische Darstellung aller Einzelmeßwerte (n = 216) von 24 Probanden. Eliminationsphase nach Alkoholaufnahme von 0,7-0,8 g/kg Körpergewicht. Die Modellkurve entspricht einer kombinierten linearen und nichtlinearen Regression (Härdle u. Mattern 1983)

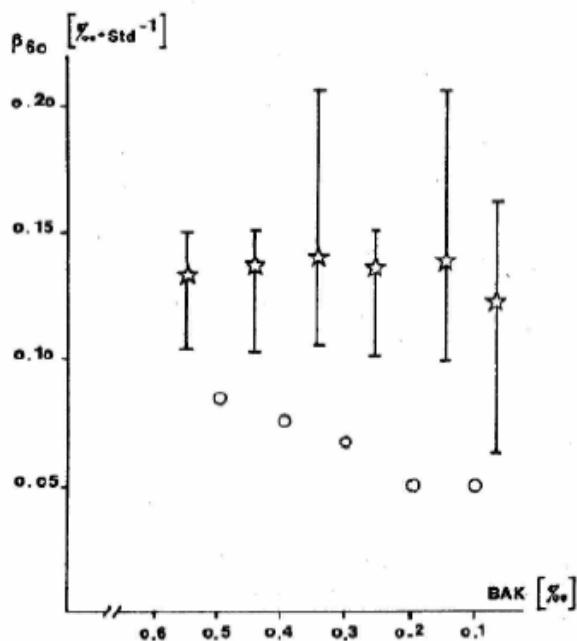


Abb. 2. β_{60} -Werte in der späten Eliminationsphase. ☆ Mittelwerte mit Streubereichen der eigenen Untersuchungen (24 Probanden). ○ Minimalwerte nach Klepsch (1969)

immer mindestens 0,1‰ betragen. Auch unterhalb von 0,1‰ - hier wurde der Bereich von 0,03 bis 0,09‰ als Gruppe zusammengefaßt - ergab sich mit 0,127‰ ein mittleres β_{60} über 0,1‰, die Streubereiche umfaßten hier Werte von 0,064-0,162‰.

Projiziert man in diese Darstellung die Ergebnisse von Klepsch (1969), so fällt auf, daß seine Untersuchungen deutlich niedrigere β_{60} -Werte lieferten, selbst wenn man unsere unteren Extremwerte berücksichtigt.

Ein wesentlicher Grund für diese Unterschiede liegt zweifellos im methodischen Ansatz von Klepsch (1969), der seinen Probanden mit 0,3-0,5 g/kg Körpergewicht recht kleine Alkoholmengen angeboten hatte. Außerdem bezieht er in seine Berechnungen Meßwerte von Blutproben ein, die bereits 45 min nach Trinkende entnommen worden waren. Unter diesen Voraussetzungen muß auch in der späten Eliminationsphase noch mit einer ausklingenden Resorption gerechnet werden.

In weiterführenden mathematisch-statistischen Analysen der experimentell gewonnenen Meßwerte wurde versucht, durch Entwicklung geeigneter mathematischer Modelle jenen Konzentrationsbereich zu bestimmen, von dem ab der Kurvenverlauf besser durch eine nichtlineare, als durch eine lineare Elimination beschrieben wird (Härdle u. Mattern 1983). Es konnte gezeigt werden, daß mit jenem Modell, das unter dem Kriterium der nichtlinearen Kleinste-Quadrate-Anpassung die Daten am besten beschreibt, der "Umschlagpunkt" für das untersuchte Kollektiv im Mittel bei 0,052‰ mit 95% Konfidenzbereichen von 0,033 bis 0,070‰ lag. Ähnliche Konzentrationen am "Umschlagpunkt" gaben Lester (1962), Larsen (1959), Scheer (1967) sowie kürzlich auch Wolf u. Wiens (1982) an.

Schlußfolgerungen

1. Gaschromatographisch bestimmte Blutalkoholkonzentrationen sind auch in der Größenordnung von 0,02-0,1‰ zuverlässige Meßwerte, die als Grundlage für Alkoholbegutachtungen geeignet sind.
2. In der Eliminationsphase kommen oberhalb einer Blutalkoholkonzentration von 0,1‰ Stundenabbauwerte von weniger als 0,1‰ nicht vor.
3. Die späte Eliminationsphase läßt sich mathematisch in einen linearen und nichtlinearen (exponentiellen) Teil gliedern. Die Grenzkonzentration zwischen diesen beiden Bereichen liegt in der Regel unter 0,1‰. In der exponentiellen Endphase der Elimination ist die Angabe eines β 60-Wertes nicht mehr sinnvoll.
4. Diese Feststellungen 1.-3. gelten für die untersuchten Probanden in der reinen Eliminationsphase. Die recht homogene Struktur und der Umfang des Kollektivs erlauben keine Verallgemeinerung; wir erwarten jedoch auch für ein anders zusammengesetztes Kollektiv keine grundsätzlich anderen Ergebnisse.

Literatur

- Forster B, Joachim H (1975) Blutalkohol und Straftat. Nachweis und Begutachtung für Ärzte und Juristen. Thieme, Stuttgart
- Forster B, Schulz G, Starck EJ (1961) Untersuchungen über den Blutalkoholabbau und seine forensische Bedeutung. Blutalkohol 1:2-7
- Förster H, Hartmann H (1980) Kann die alkoholbedingte Fahruntüchtigkeit beeinflußt werden? Dtsch Apoth Z 23:1045-1050
- Härdle W, Mattern R (1983) Mathematische Modellierung der Eliminationsphase des Ethanols. In: Barz J (Hrsg) Fortschritte der Rechtsmedizin. Springer, Berlin Heidelberg New York
- Klepsch D (1969) Die Alkoholelimination im Bereich niedriger Blutalkoholkonzentrationen. Inauguraldissertation, Universität Göttingen
- Koppun M, Propping P (1977) The kinetics of ethanol absorption and elimination in twins and supplementary repetitive experiments in singleton subjects. Eur J Clin Pharmacol 11/5:337-344

- Kulpe W, Mallach HJ (1960) Blutalkohol bei Leberkranken. Med Sachverst 56:270-274
- Larsen SA (1959) Determination of the hepatic blood flow by means of ethanol. Scand J Clin Lab Invest 11:340-347
- Lester D (1962) The concentration of apparent endogenous ethanol. Q J Stud Alcohol 23:17-23
- Rietbrock N, Abshagen U (1971) Pharmakokinetik und Stoffwechsel aliphatischer Alkohole. Arzneim Forsch 21:1309-1319
- Scheer H (1967) Über die Möglichkeiten einer Rückrechnung von niedrigen Blutalkoholwerten. Inauguraldissertation, Universität Frankfurt/Main
- Springer E (1972) Blutalkoholkurven nach Gabe von wässrigen Äthanollösungen verschiedener Konzentrationen. Blutalkohol 9:198-206
- Sprung R, Bonte W, Rüdell E, Domke M (1981) Zum Problem des endogenen Alkohols. Blutalkohol 18:65-70
- Wagner JG, Wilkinson PK, Sedmann PK, Kay DR, Weidler DJ (1976) Elimination of alcohol from human blood. J Pharm Sci 65:152-154
- Widmark EMP (1932) Die theoretischen Grundlagen und die praktische Verwendbarkeit der gerichtlich-medizinischen Alkoholbestimmung. Urban & Schwarzenberg, Berlin Wien
- Wilkinson PK, Reynolds G, Homes OD, Yang S, Wilkin LO (1980) Nonlinear pharmacokinetics of ethanol: The disproportionate AUC-dose relationship. Alcoholism (N4) 4: 384-389
- Wille R, Steigleder E (1966) Zur Frage der Rückrechnung von niedrigen Blutalkoholkonzentrationen. Blutalkohol 3:419-435
- Wolf M, Wiens N (1982) Zum Verlauf der Blutalkoholkurve im niedrigen Konzentrationsbereich. Beitr Gerichtl Med 40:63-67
- Zink P (1982) Über den Abfall der Blutalkoholkurve in Trinkversuchen und bei Doppelblutentnahmen. Blutalkohol 19:200-210
- Zink P, Blauth M (1982) Zur Frage der Beeinflussung der Blutalkoholkonzentration im Cubitalvenenblut durch die Blutentnahmetechnik. Blutalkohol 9:15-27

Klassifizierung von Digoxin-Blut- und Gewebekonzentrationen bei Vergiftungsverdacht* **

W. Härdle¹ und R. Aderjan²

¹Universität Heidelberg, Sonderforschungsbereich 123, Im Neuenheimer Feld 293, D-6900 Heidelberg 1, Bundesrepublik Deutschland

²Universität Heidelberg, Institut für Rechtsmedizin, D-6900 Heidelberg 1, Bundesrepublik Deutschland

Classification of the Digoxin Concentration in Blood and Tissues in Cases Under Suspicion of Poisoning

Summary. The clarification of a suspicion of poisoning at all times poses a problem to the forensic toxicologist, when a narrow margin of therapeutic safety and a low dosage coincide as in cases of digoxin poisoning. Statistical methods may serve as an aid. The post mortem digoxin concentration in the tissues of heart, kidney, liver and in blood of 45 patients who had received therapeutic daily doses and of 13 cases of fatal poisoning are compared.

After logarithmic transformation of the individual concentration values a two modal distribution is obtained. There is one concentration calculated with equal probability of being classified to "therapeutic or toxic", as well as the probability of observing the "critical" concentrations of 400 ng digoxin/g cardiac tissue, 500 ng/g kidney and 250 ng/g liver after therapeutic dosing.

Using the discriminant analysis each of the cases clearly falls into one of the two collectives "therapeutic" and "toxic", when taken as a separate observation. Concentration data of fatal poisonings taken from the literature are as successfully classified as the analytical results of some exhumed bodies under suspicion but not poisoned. As expected the power of discrimination increases with the number of parameters. Because of the relatively slow body distribution of digoxin the blood taken from peripheral vessels is of most important evidence.

Key words: Digoxin, tissue distribution - Poisoning, digoxin

* Diese Arbeit wurde z.T. durch Mittel der Deutschen Forschungsgemeinschaft, SFB 123 „Stochastische Mathematische Modelle“, unterstützt

** Frau Prof. Dr. Dr. M. Geldmacher-v. Mallinckrodt in Verehrung und Dankbarkeit gewidmet

Sonderdruckanfragen an: Dr. R. Aderjan (Adresse siehe oben)

Zusammenfassung. Das Zusammentreffen von geringer therapeutischer Breite und niedriger Dosierung stellte den forensischen Toxikologen beim Abklären eines Vergiftungsverdachts von jeher vor größte Probleme. Hierbei können statistische Methoden helfen: Die postmortalen Digoxinkonzentrationen in Blut und Geweben von 45 Patienten, die unter therapeutischen Dosen standen, werden denen gegenübergestellt, die nach 13 tödlichen Vergiftungsfällen festzustellen waren (Schenkelvenenblut, Herz, Leber, Nieren).

Nach logarithmischer Transformation der Konzentrationswerte läßt sich je Organ eine zweigipflige Verteilung erkennen, die aus den beiden Gruppen gebildet wird. Es läßt sich jeweils die Konzentration angeben, bei der es gleich wahrscheinlich ist, daß ein beobachteter Meßwert „therapeutisch oder toxisch“ ist, ferner die Wahrscheinlichkeit, mit der nach therapeutischer Dosierung oberhalb der „kritischen“ Konzentration von 400 ng Digoxin/g Herz, 500 ng/g Niere und 250 ng/g Leber noch Meßwerte zu beobachten sind.

Unter Anwendung einer Diskriminanzanalyse mit zunächst zwei Variablen, läßt sich jeder Fall für sich als neue Beobachtung genommen, eindeutig einem von zwei Kollektiven („therapeutisch oder toxisch“) zuordnen. Die Klassifizierung von Konzentrationsdaten literaturbekannter Vergiftungen gelingt ebenso wie die von verdächtigen exhumierten Leichen, wenn Vergiftungsverdacht gegeben war, der widerlegt wurde.

Die Unterscheidungsmöglichkeit nimmt erwartungsgemäß zu, wenn die Anzahl der betrachteten Organparameter erhöht wird. Wegen der langsamen Körperverteilung von Digoxin gehört das Blut zum wichtigsten Untersuchungsgut.

Schlüsselwörter: Digoxin, Gewebeverteilung – Vergiftung, Digoxin

Einleitung

Ist bei Vergiftungsverdacht der postmortale Nachweis des fraglichen Stoffes erbracht, so steht der forensische Toxikologe häufig vor dem Problem, daß die Kriterien für die Einordnung der quantitativen Befunde aus diesen (wenigen) zu beurteilenden Ereignissen heraus zu entwickeln sind. Der Schluß – der Mensch ist tot, also müssen diese Konzentrationen tödlich gewirkt haben – muß zumeist ohne entsprechende statistische als Grundlage dafür dienen, den Umkehrschluß zu ziehen.

Bei den klinisch häufig verordneten Digitalis-Präparaten – hier Digoxin und Derivate – kann ein geeigneter Weg zur Gewinnung von Kriterien beschritten werden: Es sind die Blut- und Gewebekonzentrationen nach therapeutischer Dosierung denen nachgewiesener Vergiftungsfälle gegenüberzustellen, um so jeden neu vorkommenden Fall einer der beiden Einschätzungen „toxisch“ oder „nicht toxisch“ mit einer entsprechenden Wahrscheinlichkeit einordnen zu können. Dies kann jedoch nur unter zwei Voraussetzungen geschehen: Erstens müssen die Konzentrationswerte gültig sein – Fehler können in der Art, dem Alter und dem Erhaltungszustand des zu untersuchenden Materiales oder in

methodischen Gründen liegen [1] – und zweitens darf bei der Glykosidvergiftung eine klinische Diagnose bzw. die Vorgeschichte der Einordnung nicht völlig widersprechen.

Im folgenden soll aufgezeigt werden, welche Zuordnungsmöglichkeiten bestehen, welche Klassifizierungen vorgenommen werden können, wenn von einem Kollektiv von 45 Patienten unter therapeutischen Dosen von Digoxin bzw. β -Methyldigoxin sowie von 13 Digoxin-Todesfällen die Konzentrationen in sektionstechnisch regelmäßig verfügbaren Körperflüssigkeiten und Organen unter Anwendung statistischer Methoden, insbesondere der Diskriminanzanalyse, miteinander verglichen werden. Die Einordnung von literaturbekannten Daten sowie einer Gruppe von exhumierten Verstorbenen, die unter Vergiftungsverdacht standen, wird unter Berücksichtigung klinischer Befunde diskutiert.

Material und Methoden

Normalkollektiv

Die Glykosidkonzentrationen wurden bei 45 therapeutisch digitalisierten Patienten in folgenden Körperflüssigkeiten und Organen radioimmunologisch bestimmt [3, 7], Tabelle 3, NGR=1 u. 2).

Schenkelvenenblut	(=VBL)
Herzmuskel (li. Ventr.)	(=HLV)
Herzmuskel (re. Ventr.)	(=HRV)
Leber	(=LEG)
Nierenrinde	(=NRR)
Nierenmark	(=NRM)
Skelettmuskel	
(M. pectoralis major)	(=SKM)
Gehirn	(=GEH)

Vergiftungen

Bei Vergiftungen mit β -Methyldigoxin oder Digoxin [1] wurden in 13 Fällen Konzentrationsdaten möglichst der gleichen Organe unter Beibehaltung der Analysenmethodik erhoben (Tabelle 3, NGR=3, 4 und 5).

Literaturbekannte Vergiftungen

Die Konzentrationsdaten von Vergiftungsfällen, die in der Literatur beschrieben sind [2, 4-6, 9-12], wurden zunächst ohne Berücksichtigung von Unterschieden der Analysenmethodik zusammengestellt (Tabelle 3, NGR=6).

Exhumierungen

Von einer Gruppe exhumierter Leichen wurden die Digoxin-Konzentrationen in noch verfügbaren Organen mit gleicher Analysenmethodik [1] untersucht. Häufig standen nur Vollblut unbekannter Herkunft und Skelettmuskelproben und Lebergewebe zur Verfügung (Tabelle 3, NGR=7 u. 8).

Statistische Methoden

Jedem Fall wird das Profil seiner Konzentrationen in den einzelnen Kompartimenten als multivariable Größe X zugeordnet. Beispielsweise ist $X=(VBL, LEG)=(135, 450)$ für den Fall

V2 (Tabelle 3). Das Ziel der statistischen Analyse ist es, das Normalkollektiv und die Vergiftungsfälle auf der Basis der gemessenen Konzentrationen X geeignet voneinander zu trennen. Dazu verwendeten wir ein diskriminatorisches Verfahren, das auf folgender Distanzbildung basiert:

$$D_j^2(X) = (X - \bar{X}_j) \text{cov}^{-1} (X - \bar{X}_j) - 2 \log p_j,$$

wobei der Index j die Gruppenzugehörigkeit angibt, d.h. $j=1$ = normal, $j=2$ = vergiftet. X bezeichnet den Mittelwert in der j -ten Gruppe und cov schließlich die Kovarianzmatrix [8]. Die Zahlen p sind gewisse a priori Wahrscheinlichkeiten, die die subjektiven Erfahrungen widerspiegeln. Sie wurden von uns auf $p=1/2$ gesetzt, d.h. wir nahmen von einem zu klassifizierenden Fall an, daß er zu gleicher Wahrscheinlichkeit der Vergiftetengruppe oder dem Normalkollektiv zuzuordnen war. Klassifiziert wird nun mit folgender Größe:

$$p(j/X) = \exp(0.5 D_j^2(X)) / \sum_{j=1}^2 \exp(0.5 D_j^2(X)) \quad j = \text{Gruppenindex}$$

Das Diskriminationsverfahren geschieht nun folgendermaßen:

- 1) Ordne Patient m. Vektor X in Gruppe 1 falls $p(1/X) > p(2/X)$;
- 2) Ordne Patient m. Vektor X in Gruppe 2 falls $p(1/X) < p(2/X)$;
- 3) Ordne Patient m. Vektor X in Gruppe 1, 2 falls $p(1/X) = p(2/X)$.

Dieses Verfahren kann in jeder beliebigen Parameterdimension d vorgenommen werden. Ist $X = (\text{LHV}, \text{LEG})$, wie oben, hat d den Wert 2. Die von uns ausgeführten Klassifizierungen wurden aufgrund der beschränkten Datenmenge nur für $d=2, 3, 4$ und verschiedene Konzentrationsvariable durchgeführt.

Im eindimensionalen Fall ($d=1$) läßt sich das Verfahren an Abb. 4 veranschaulichen. Hier ist der indifferente Fall $p(1/X)$ durch einen trennenden Punkt (289 ng/g für $X = \text{LHV}$) gegeben. Jede Konzentration, die links von diesem Wert fällt, erfüllt $p(1/X) > p(2/X)$, ist also dem Normalkollektiv zuzuschlagen. Hingegen sind $p(2/X) > p(1/X)$ von den Werten rechts dieses trennenden Punktes erfüllt, also der Vergiftetengruppe zugeordnet. In höherer Dimension z.B. $d=2$, wird die Trennung durch eine Gerade gegeben. In Abb. 5 hat man sich diese Trennlinie ungefähr als Verbindung der Punkte (0.8) und (7.0) vorzustellen.

Tabelle 1. Logarithmische Mittelwerte und Abgrenzungsschranken für die Erfassung therapeutischer Konzentrationswerte. Das 99%ige Niveau entspricht den anhand der Suizide abgeschätzten Grenzkonzentrationen

Abgrenzungsschranken in ng/g zur Erfassung der therapeutischen Werte mit einer Wahrscheinlichkeit von:

Gewebeart	Mittelwert In-Skala	95%	97,5%	99%	99,9%	Beginn des toxischen Konzentrations- bereichs
Herz li. Ventr.	5,147 ± 0,27	267	291	324	386	400
Herz re. Ventr.	4,748 ± 0,492	256	300	363	500	400
Nierenmark	4,941 ± 0,574	358	430	539	782	500
Nierenrinde	4,973 ± 0,507	331	390	475	661	500
Leber	4,077 ± 0,569	156	189	239	352	250
Muskulatur	3,22 ± 0,627	70	86	110	165	—

Tabelle 2. Mittelwerte und Standardabweichungen der Digoxin-Konzentration in Körperflüssigkeiten und Geweben bei Patienten unter therapeutischen Dosen von β -Methyl Digoxin und Digoxin ($n=45$) sowie bei letalen Digoxinvergiftungen. Die nach logarithmischer Transformation der Einzelwerte abzuleitenden Verteilungen (s. Beispiel Herzmuskulatur Abb. 4) führen zu (a) einer Abgrenzungsschranke, dafür, daß es gleich wahrscheinlich ist, daß nach therapeutischer oder (letal-)toxischer Dosierung ein Meßwert zu beobachten ist. Die Wahrscheinlichkeit dafür, daß bei therapeutischer Dosierung ein Meßwert über dieser Schranke und dafür, daß nach toxischer Dosis ein Meßwert unter dieser Schranke zu beobachten ist, nimmt jeweils ab; (b) der Wahrscheinlichkeit, mit der am Beginn des toxischen Konzentrationsbereiches (aufgrund der beobachteten Vergiftungsfälle und unter Berücksichtigung einer Sicherheitszone von 100 ng/g ab dem höchsten therapeutischen Meßwert) ein Meßwert als Folge einer therapeutischen Dosierung zu beobachten ist.

	Mittelwert \pm Standardabweichung der Digoxin-Konzentration in ng/g		Abgrenzungsschranke gleicher Wahrscheinlichkeit (a)		Wahrscheinlichkeit, mit der die Grenze z. toxischen Konzentrationsbereich nach therap. Dosierung beobachtet wird (b)	
	Nach therap. Dosierung $n=45$	Nach toxischer Dosis $n=6-13$	ng/g	% Wahrscheinlichkeit	Grenzkonz. ng/g	% Wahrscheinlichkeit
Vollblut	5,8 \pm 2,4	40,7 \pm 32,5	11	3,92	20	0,06
Herzmuskel li. Ventrikel	178 \pm 50,6	541 \pm 155	289	2,56	400	0,09
Herzmuskel re. Ventrikel	128 \pm 58	485 \pm 167	257	4,05	400	0,55
Leber	67,7 \pm 32	333 \pm 191	115	11,3	250	0,48
Nierenrinde	161 \pm 72	1200 \pm 764	318	5,7	500	0,64
Nierenmark	162 \pm 86	601 \pm 199	329	6,7	500	1,26
Skelettmuskel	30,8 \pm 21	70 \pm 31	43	18,7		Entfällt wegen zu starker Überschneidung der Wertebereiche
Gehirn	26,8 \pm 12	39 \pm 29	26	57,9		

Tabelle 3. Zusammenfassung der gesamten Konzentrationsdaten in ng/g bzw. ng/ml, die für die Diskriminanzanalyse zur Verfügung standen

OBS	NPR	VBL	LHV	RHV	SKM	NRR	NRM	LEG	GEH	NGR
1	1	5	110	98	10	143	125	25	—	1
2	2	3	144	45	12	205	143	30	—	1
3	3	5	156	175	15	115	63	70	—	1
4	4	10	245	135	32	165	115	110	—	1
5	5	7	280	120	14	150	250	120	—	1
6	6	4	180	100	13	170	105	55	—	1
7	7	4	280	200	15	262	355	105	—	1
8	8	4	244	150	15	150	80	160	—	1
9	9	5	305	175	32	115	75	80	—	1
10	10	7	280	200	28	320	175	70	—	1
11	11	3	143	70	12	102	70	60	—	1
12	12	3	120	90	10	110	60	85	—	1
13	13	6	205	165	23	65	135	65	—	1
14	14	9	175	98	38	265	250	70	—	1
15	15	5	165	90	16	55	100	82	—	1
16	16	9	120	90	12	40	135	32	—	1
17	17	5	145	94	14	110	250	27	—	1
18	18	9	150	110	16	120	150	40	—	1
19	19	4	255	112	30	118	200	90	—	1
20	20	7	210	280	12	230	380	68	—	1
21	21	4	185	80	45	125	280	79	—	1
22	22	3	120	25	15	35	180	28	—	1
23	23	4	140	80	39	120	145	45	—	1
24	24	4	225	100	41	165	241	25	—	1
25	25	14	170	125	40	275	360	101	—	1
26	A	8	160	195	15	275	—	30	21	2
27	B	6	220	250	45	125	150	50	18	2
28	C	3	108	80	20	89	95	25	15	2
29	D	10	160	150	32	125	50	30	22	2
30	E	5	145	75	90	300	250	55	28	2
31	F	9	124	65	35	200	175	40	27	2
32	G	6	115	63	80	80	90	45	—	2
33	H	3	145	175	90	155	225	65	40	2
34	I	7	164	88	55	130	135	110	21	2
35	K	6	165	185	32	80	35	10	18	2
36	L	5	160	55	31	150	80	80	14	2
37	M	6	145	70	45	225	210	90	19	2
38	N	4	160	195	27	190	150	60	45	2
39	O	8	245	175	40	162	125	110	50	2
40	P	7	210	250	95	190	135	120	40	2

Tabelle 3. (Fortsetzung)

OBS	NPR	VBL	LHV	RHV	SKM	NRR	NRM	LEG	GEH	NGR
41	R	9	200	160	20	255	250	70	40	2
42	S	6	168	145	30	140	69	90	25	2
43	T	8	160	125	18	250	200	100	40	2
44	U	3	155	105	30	275	250	95	12	2
45	W	3	170	175	10	150	55	50	12	2
46	S1	22	—	—	50	1400	—	81	10	3
47	S2	28	—	—	110	500	—	250	45	3
48	S3	25	760	540	—	1320	—	200	90	3
49	S4	45	320	300	40	640	750	180	17	3
50	S6	15	480	420	80	620	680	390	32	3
51	S5	18	700	680	—	450	375	400	40	3
52	V1	52	510	—	—	1700	—	760	—	4
53	V2	135	510	—	—	2850	—	450	—	4
54	V3	24	490	—	—	1470	—	580	—	4
55	V4	34	560	—	—	1920	—	370	—	4
56	V5	70	750	—	—	1880	—	424	—	4
57	V6	17	—	—	—	—	—	200	—	5
58	V7	45	—	—	—	560	—	330	—	5
59	V8	13	560	—	—	680	—	250	—	7
60	V9	20	266	—	—	760	—	380	—	7
61	V0	—	333	—	—	290	—	58	—	5
62	A1	30	300	—	—	—	—	—	—	6
63	A2	12	282	—	—	—	—	—	—	6
64	A3	24	624	—	—	—	—	—	—	6
65	A4	30	—	—	—	130	—	35	—	6
66	A5	75	143	160	12	237	144	47	23	6
67	A6	25	—	—	—	143	—	110	—	6
68	E1	6	—	—	31	—	—	88	—	8
69	E2	6	—	—	8	—	—	24	—	8
70	E3	12	—	—	10	—	—	42	—	8
71	E4	—	—	—	15	—	—	—	—	8
72	E5	6	—	—	22	—	—	60	—	8

OBS=Beobachtungs-Nr.; NPR=Interne Bezeichnung; VBL=Schenkelvenenvollblut; LHV=linker Herzventrikel; RHV=rechter Herzventrikel; SKM=Skelettmuskulatur; NRR=Nierenrinde; NRM=Nierenmark; LEG=Lebergewebe; GEH=Gehirn; NGR=1, 2=Normalkollektiv; NGR 3, 4, 5=Vergiftungen; NGR=6=Literaturdaten; NGR=7, 8=Exhumierte Leichen

Ergebnisse und Diskussion

Begründung für die Anwendung einer Diskriminanzanalyse

Schon die einfache Gegenüberstellung der Digoxin-Konzentrationen in den für die forensisch-toxikologische Beurteilung wichtigen Körperorganen und Flüssigkeiten zeigt, daß ab 400 ng/g für das Herz, ab 250 ng/g für die Leber und ab 500 ng/g für die Nieren und ab 20 ng/g für das Schenkelvenen(voll)blut der Beginn des toxischen Konzentrationsbereiches anzunehmen ist. Die „therapeutischen“ Konzentrationen weisen in allen untersuchten Kompartments schiefe, eher logarithmische Verteilungen auf [1]. Die Wahrscheinlichkeit, daß ab einem Konzentrationswert therapeutische Konzentrationen unterhalb eines beobachteten Meßwertes liegen, ist in Tabelle 1 wiedergegeben. Danach entspricht etwa das 99%-Niveau den oben angeführten Grenzen. In Abb. 1-4 sind die Verteilungen der therapeutischen und der toxischen Konzentrationen dargestellt. Die Daten sind zum besseren Verständnis logarithmiert; dadurch werden Verteilungen mit mehr Symmetrie erzielt, die für viele Anwendungen statistischer Methoden geeigneter sind. Deutlich sind zwei Gipfel der gemeinsamen Verteilung und Bereichsüberschneidungen der Einzelverteilungen erkennbar. Unter der nicht völlig korrekten Annahme symmetrischer Verteilungen und gleicher Varianz ergibt sich, daß der Schnittpunkt der Verteilungskurven (Diskriminationsschranke, s. Abb. 4 für Herzmuskulatur) den in Tabelle 2 wiedergegebenen Konzentrationen und Wahrscheinlichkeiten für die Zuordnung zu einer der beiden Gruppen (s. „1“ oder „2“ in Abb. 1-3) entspricht. Für die oben angeführten Grenzwerte für den Beginn des toxischen Konzentrationsbereiches ergibt sich eine Wahrscheinlichkeit von $<1\%$ bis $<0,1\%$, daß in dieser Höhe eine Gewebekonzentration therapeutischer Dosierung zu beobachten ist (Tabelle 3).

Aus vielen Untersuchungen ist bekannt, daß eine Überschneidung des therapeutischen und toxischen Konzentrationsbereiches bei klinischen Digoxin-Serumspiegelbestimmungen eine Zuordnung und die Annahme einer Intoxikation nicht rechtfertigt, solange nicht klinische Zeichen dafür sprechen. Dies gilt auch für die Gewebekonzentrationen.

In einzelnen Geweben, besonders Skelettmuskulatur und Gehirn, können so starke Überschneidungen festgestellt werden, daß ein einzelner Meßwert eine Zuordnung im Zweifelsfall nicht gestattet. Daß in einzelnen Gruppen durchaus eine individuell höhere Anreicherung einer Glykosiddosis beobachtet werden kann, muß noch nicht für die Aufnahme einer toxischen Dosis sprechen. Wenn aber die Anzahl der zu vergleichenden Individualdaten erhöht wird, kann die Unterscheidung zunehmend besser getroffen werden. Der Erfahrene pflegt die Zuordnung „mit einem Blick“ durchzuführen. Demgegenüber ermöglicht die Diskriminanzanalyse mit einer zunehmenden Anzahl von Variablen, bei genügendem Umfang der Vergleichsgruppen ein zahlenmäßig faßbares Verständnis der zu treffenden Entscheidung. Als grundlegende Informationsträger sind

Abb. 1-3. Darstellung der 2-gipfligen gemeinsamen Verteilungskurven nach logarithmischer Transformation der Konzentrationswerte von linksventrikulärer Herzmuskulatur (LHV), Lebergewebe (LEG) und Nierenrinde (NRR). 1= Normalkollektiv, 2= Vergiftungsfälle

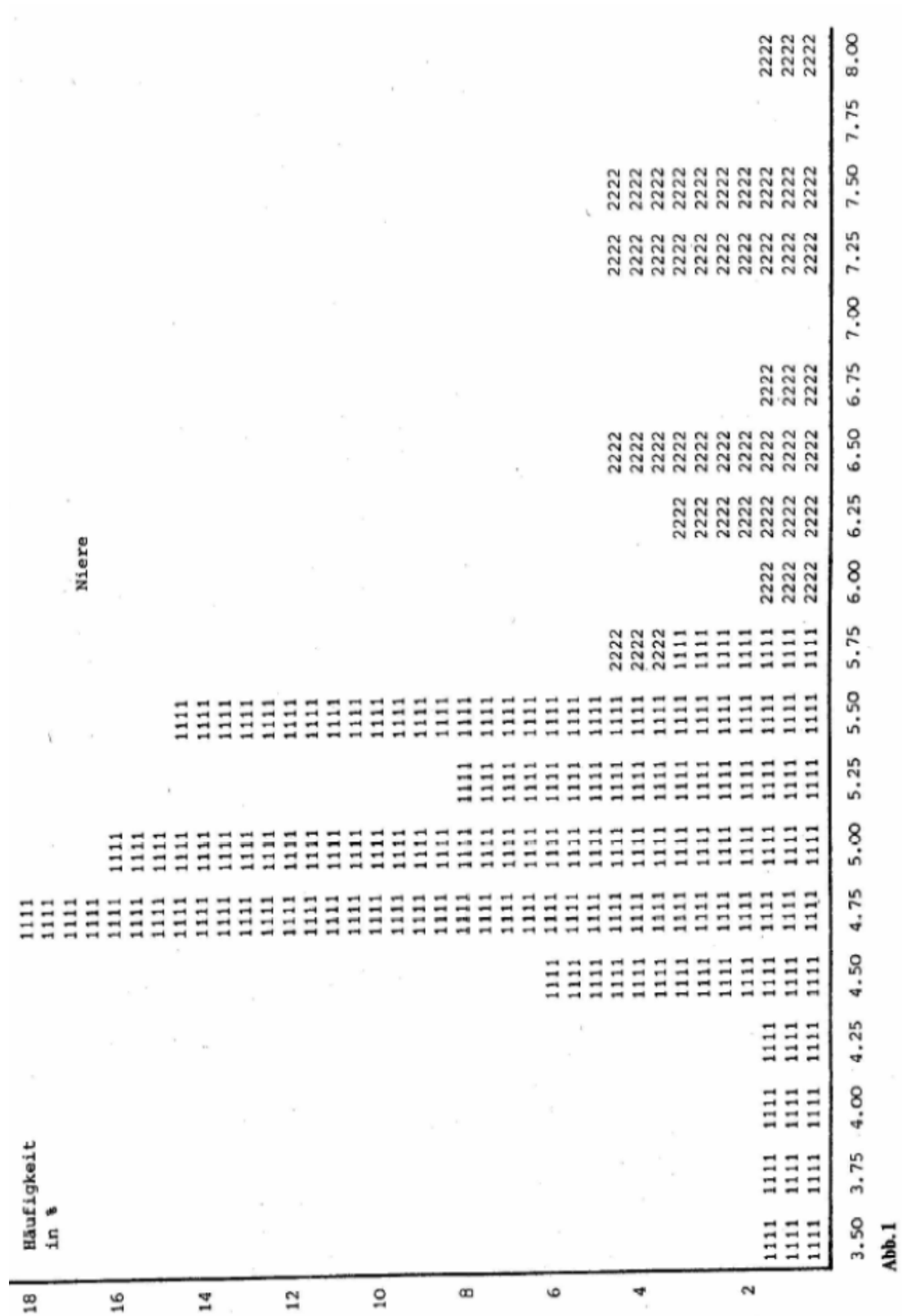


Abb. I

Härdle, W. and Aderjan, R. (1983) **Klassifikation von Digoxin- Blut und Gewebe-konzentrationen bei Vergiftungsverdacht.**

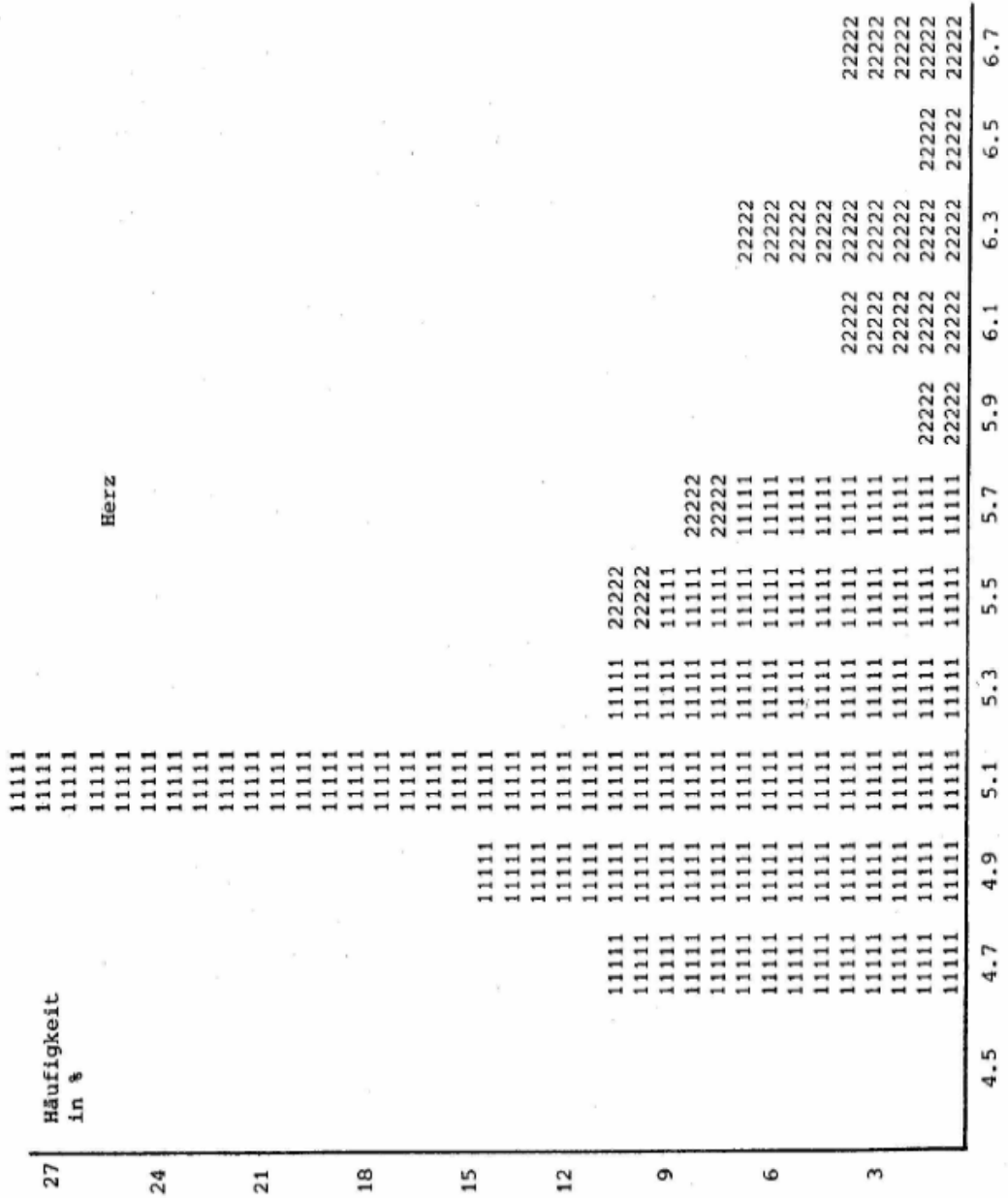


Abb.2

Härdle, W. and Aderjan, R. (1983) **Klassifikation von Digoxin- Blut und Gewebe-konzentrationen bei Vergiftungsverdacht.**

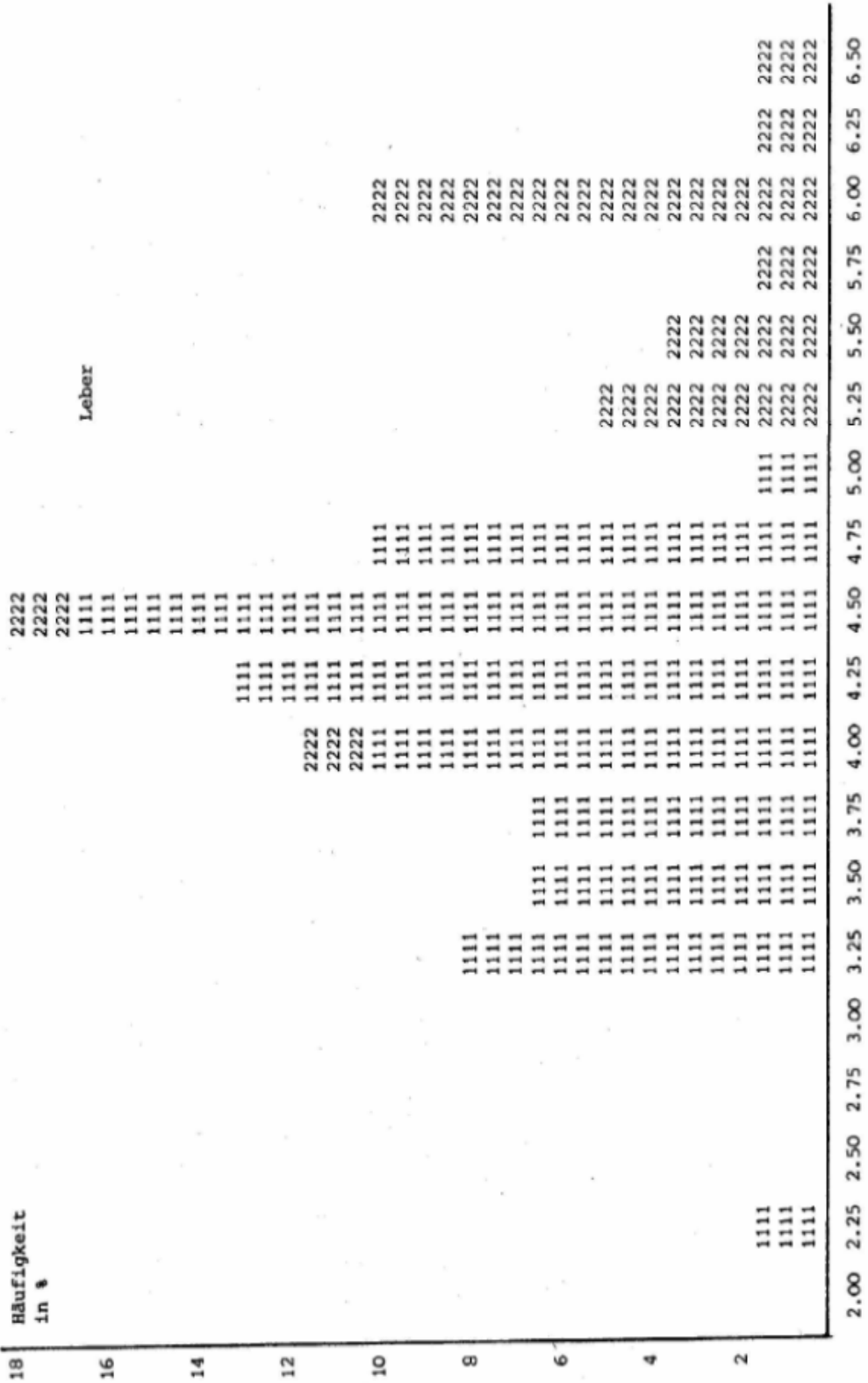


Abb. 3

Härdle, W. and Aderjan, R. (1983) **Klassifikation von Digoxin- Blut und Gewebe-konzentrationen bei Vergiftungsverdacht.**

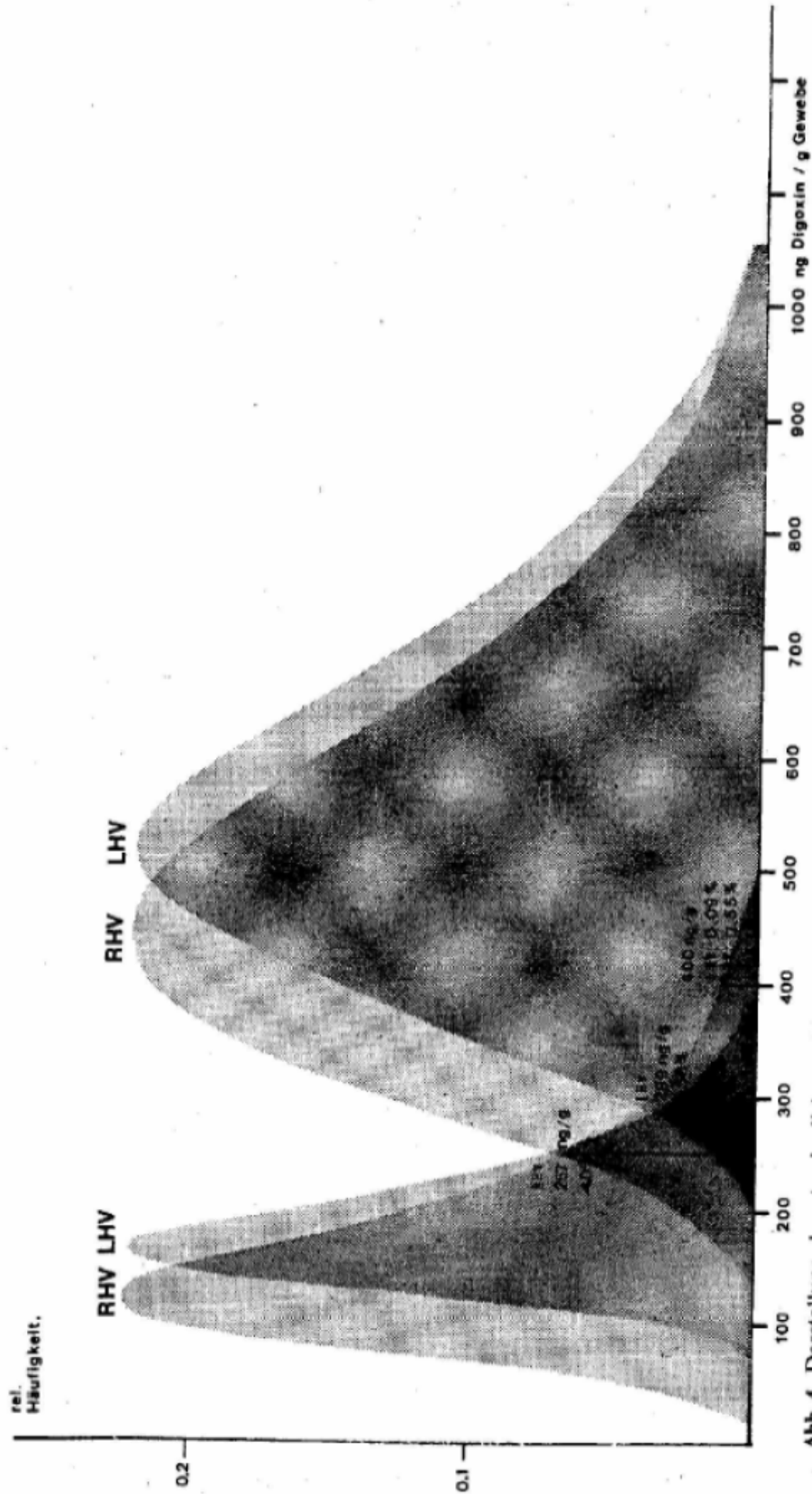


Abb. 4. Darstellung der standardisierten Verteilung der Digoxin-Konzentration in linksventrikulärer (LHV)-rechtsventrikulärer (RHV) Herzmuskulatur nach therapeutischer Dosierung und nach toxischen Dosen berechnet aus Mittelwert und Standardabweichung nach logarithmischer Transformation. Die Schnittpunkte der Verteilungskurve zeigen die Konzentration an, die mit gleicher Wahrscheinlichkeit einem der beiden Kollektive zugeordnet werden kann. Die höhere Standardabweichung, die der toxischen Verteilungskurve zugrunde liegt, kann auch auf die stark unterschiedlichen toxischen, z. T. unbekannteren Dosen zurückzuführen sein

Härdle, W. and Aderjan, R. (1983) **Klassifikation von Digoxin- Blut und Gewebe-konzentrationen bei Vergiftungsverdacht.**

nach den vorliegenden Ergebnissen die Digoxin-Konzentrationen im Schenkelvenenblut Herzmuskulatur, Leber und Nieren anzusehen. Konzentrationen in Skelettmuskulatur und Gehirn sind nur dann heranzuziehen, wenn andere Gewebe nicht verfügbar sind.

Durchführung der Diskriminanzanalyse

Das diskriminatorische Verfahren geschah wie folgt:

1) Festlegung der Dimension des Parametervektors (Anzahl der zu vergleichenden Variablen, s. II, 5).

2) Erstelle mit den pro Organ gemessenen Konzentrationsdaten die Kovarianzstruktur der beiden Kollektive – Vergiftete (NGR=3, 4, 5) und therapeutisch Digitalisierte (NGR=1, 2).

3) Prüfe die Trennbarkeit der gebildeten Gruppen mit dem ausgewählten Parametervektor (z.B. Lebergewebe, Vollblut), indem jeder Fall für sich auf seine Gruppenzugehörigkeit getestet wird.

4) Prüfe, in welche Gruppe die Daten der literaturbekannten Vergiftungen und die von Exhumierungen (NGR=6, 7, 8 s. Tab. 3) jeweils zugeordnet werden.

Die Lage der Daten zueinander in dem von den Parametern (Leber, Vollblut) bzw. (Vollblut, Nierenrinde) aufgespannten Raum ist in Abb. 5 bzw. Abb. 6 als Beispiel dargestellt. Deutlich erkennbar ist die verbesserte Unterscheidbarkeit der Kollektive im Vergleich zu der eindimensionalen Betrachtungsweise (Abb. 1-3). Welche Kombination an zweidimensionalen Parametern aus den Meßwerten der vier wichtigsten Organe auch immer gebildet werden, die Zuordnung nach (3) ist praktisch immer eindeutig. In der Gruppe der therapeutisch Digitalisierten wird höchstens ein Fall fehlklassifiziert, (aufgetreten bei Fall NPR=25, Leber, Vollblut, Tabelle 3). Die Vergiftungen werden stets in den unterschiedlichen Parametern richtig eingeordnet. Während dieses Ergebnis völlig den Erwartungen entspricht, wie man auch aus Abb. 5 und Abb. 6 unwillkürlich empfindet, ist die richtige Einordnung der literaturbekannten Vergiftungen (NGR=6) und der Exhumierungen (NGR=8) nicht ohne weiteres evident.

Die Parameterkombination (Leber, Vollblut) erbrachte eine 100%ige Klassifizierung der literaturbekannten Daten von Vergiftungen. Die 2 Exhumierungen der Gruppe NGR 7 wurden den Vergiftungen zugeordnet, obwohl nach klinischen Befunden eine tödlich verlaufende Herzglykosid-Vergiftung nicht vorgelegen hatte. Dies verdeutlicht, daß der Gültigkeit der Werte (exhumierte Leichen sind mit rasch nach dem Tod untersuchten nicht vergleichbar) besondere Beachtung zu schenken ist [1].

Um so mehr befriedigt, daß die Daten aus der Literatur bei dieser Kombination richtig zugeordnet werden (soweit die entsprechenden Daten erhoben wurden), obwohl mit unterschiedlichen Labormethoden untersucht wurde.

In derselben Weise wurden die Parameterkombinationen (linker Herzventrikel, Leber) und (linker Herzventrikel, Niere) gebildet. Hier allerdings wurde der einzige literaturbekannte Fall, bei dem diese Kombination gebildet werden konnte, in die Normalgruppe klassifiziert. Diese Vergiftung wurde nur 1 Stunde

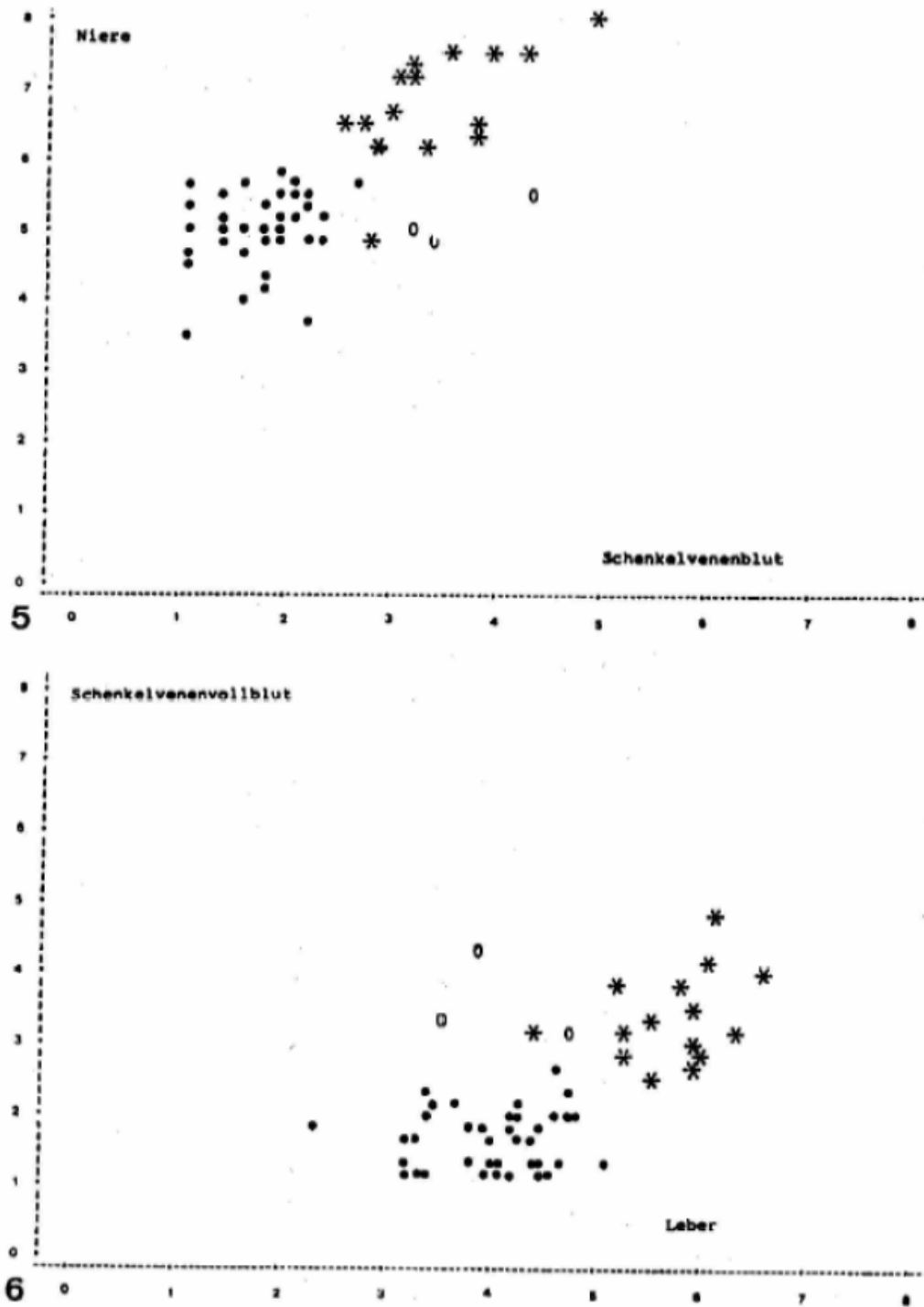


Abb. 5 und 6. Darstellung des Normalkollektivs und der Digoxinvergiftungen in dem von den Parametern Lebergewebe und Schenkelvenenvollblut sowie Nierengewebe und Schenkelvenenvollblut auf gespanntem Raum. * = Vergiftungen, ● = Normalkollektiv, 0 = Testfälle nach Literaturangaben

Härdle, W. and Aderjan, R. (1983) **Klassifikation von Digoxin- Blut und Gewebe-konzentrationen bei Vergiftungsverdacht.**

lang überlebt, weshalb die aufgenommene Dosis noch nicht vom Blut in die einzelnen Gewebekompartments verteilt war [10]. Von den Exhumierungen konnten nur die beiden Fälle von $NGR=7$ in das Verfahren aufgenommen werden, welche wiederum in die Vergiftetengruppe zugeordnet wurden. Eine Erhöhung der Parameterdimension auf vier Variablen, nämlich (linker Herzventrikel, Leber, Nierenrinde, Vollblut) erbrachte demgegenüber nicht nur eine 100%ige Klassifizierung im Schritt (3), sondern auch die Einordnung des einzig verwertbaren Literaturfalles [10] ($NPR=A\ 5$, $NGR\ 6$) in die Vergiftetengruppe. Als Konsequenz daraus ist abzuleiten, daß bei Vergiftungsverdacht mindestens die Organe Herz, Leber, Niere und Schenkelvenenblut zu asservieren wären, um die erhöhte Unterscheidungsfähigkeit auszunutzen. Dabei verlangt die Berücksichtigung der frühen Verteilungsphase nach der Aufnahme des Giftes, da die Konzentration des Schenkelvenenblutes als Variable beteiligt ist.

Literatur

1. Aderjan R (1981) Tödliche Vergiftungen mit Herzglykosiden-Nachweis und rechtsmedizinisch-toxikologische Befundbewertung. Habilitationsschrift, Heidelberg
2. Arnold W, Püschel K (1979) Toxikologische und morphologische Befunde bei Digoxinvergiftung in forensischer Sicht. Z Rechtsmed 83 : 265-271
3. Doster S (1980) Probleme des Nachweises von Digoxinintoxikationen aus Leichengewebe (β -Methyldigoxin). Dissertation, Heidelberg
4. Iisalo E, Nuutila M (1973) Myocardial digoxin concentrations in fatal intoxications. Lancet 257
5. Jelliffe RW (1967) Autopsy verification of suicide by digitalis. Report of a case with successful chemical identification of digitalis glycosides in gastric contents. Am J Clin Pathol 47 : 180-185
6. Larbig D, Haasis R, Kochsiek K (1978) Die Glykosidkonzentration und ihre klinische Bedeutung. Form cardiologium 15. Boehringer, Mannheim
7. Petri H (1980) Probleme des Nachweises von Digoxinintoxikationen aus Leichengewebe (Digoxin u. acetylierte Derivate). Dissertation, Heidelberg
8. Press SJ (1972) Applied multivariate analysis. Holt, Reinhart and Winston, New York
9. Reissell P, Alha A, Karjalainen J, Nieminen R, Ojala K (1975) Digoxinintoxication determined post mortem. Abstr VIth Int Congr Pharmacol 386
10. Rietbrock N, Wjahn H, Weinmann J, Hasford J, Kuhlmann J (1978) Tödlich verlaufene β -Methyldigoxin-Intoxikation in suizidaler Absicht. Dtsch Med Wochenschr 103 : 1841-1844
11. Selesky M, Spiehler V, Cravey RH, Elliot HW (1976) Digoxin concentrations in fatal cases. J Forensic Sci 22 : 409-417
12. Steentoft A (1973) Fatal digitalis poisoning. Acta Pharmacol Toxicol 32 : 353-357

Eingegangen am 21. Juni 1982

Mathematische Modellierung der Eliminationsphase des Äthanols

W. Härdle und R.Mattern

Zusammenfassung

Es werden verschiedene Methoden zur mathematischen Modellierung der Blutalkoholkonzentration-Elimination vorgestellt. Die modellspezifischen Eigenschaften werden diskutiert und ihre Relevanz in der rechtsmedizinischen Praxis geprüft. An einem bisher in der Literatur nicht genannten Modell, einer nichtlinearen Extension der von Widmark (1932) postulierten linearen Elimination, werden die Tatzeit-BAK-Schätzungen a posteriori und ihre statistisch erfaßbaren Fehler beschrieben. Das vorgeschlagene nichtlineare Modell ist durch die geringe Anzahl zu bestimmender Parameter praktikabel und erlaubt damit eine einfache Abgrenzung der linearen von der nichtlinearen Eliminationsphase.

Summary

Various methods for mathematical modeling of blood alcohol concentration (BAC) elimination are presented. The characteristics of the models are discussed and their relevance examined in medicolegal practice. With a model which has so far not been described in the literature, a nonlinear extension of the linear elimination postulated by Widmark (1932), the time of action BAC-estimations a posteriori, and statistically recorded failures are described. The proposed nonlinear model is practicable because of the low number of determinative parameters, therefore allowing a simple differentiation of the linear from the nonlinear elimination phase.

Einleitung

Der Pharmakokinetik des Äthanols im Menschen kommt in der rechtsmedizinischen Praxis große Bedeutung zu. Die Kenntnis der Eliminationskinetik des oral eingenommenen Alkohols ermöglicht bereits aus einer einzigen Alkoholbestimmung Schätzungen von zeitlich vor der Beobachtung liegenden Blutalkoholkonzentrationen - Verfahren, die als "Rückrechnung" in die Rechtsprechung Eingang gefunden haben.

In den 30er Jahren begann Widmark, die Elimination des Äthanols im menschlichen Körper mathematisch zu beschreiben, indem er eine konstante Eliminationsrate annahm (Abb. 32 Widmark 1932).

$$(1.1) \quad A(t) = -\beta t + C_0 \quad 0 \leq t \leq T, \quad \beta > 0$$

Hierbei bezeichnen $A(t)$ die momentane Alkoholkonzentration zur Zeit t , β und C_0 die zu bestimmenden Parameter, die natürlich einer interindividuellen Variation unterliegen. C_0 beschreibt den Konzentrationswert zur Zeit $t = 0$. β ist die instantane Abbaurate, in (1.1) konstant und proportional zu dem in der Literatur bekannten β_{60} , welches den mittleren Abbau über 1 h angibt.

Aufgrund der Linearität des Modells (1.1) ist es evident, daß über die Zeit $t = C_0/\beta$ hinaus unsinnige, negative Konzentrationswerte von dem

Modell (1.1) präzisiert würden. Eine Extension des Widmark-Modells (1.1) über $t=C_0/\beta$ hinaus ist also nicht möglich. Wir werden in diesem Papier mehrere mathematische Methoden und Modelle vorstellen, die

1. den Konzentrationsverlauf in der späten Eliminationsphase erkennen, und
2. eine BAK-Schätzung a posteriori aus dieser Phase heraus erlauben.

Methoden, Modelle

Wir nehmen vorerst keinen bestimmten funktionalen Zusammenhang zwischen der Zeit t und der Konzentration A an, d.h., wir lassen jede beliebige Art von Konzentrationszeitverlauf zu und spezifizieren schrittweise. Wir nehmen weiterhin an, daß wir mehrere Messungen Y_i des BAK in einem Versuchskollektiv gemacht haben:

$$(2.1) \quad Y_i = A(t_i) + \epsilon_i, \quad i = 1, \dots, n$$

d.h., es wurden n Messungen vorgenommen zu den Zeitpunkten t_1, t_2, \dots, t_n . $A(t)$ bezeichnet wieder die tatsächlich vorliegende Alkoholkonzentration und die $\epsilon_i, i = 1, \dots, n$ repräsentieren den Meßfehler, von dem wir annehmen, daß zu verschiedenen Messungen zu Zeiten $t_i \neq t_j$ unabhängige Fehler ϵ_i, ϵ_j vorliegen.

Das Ziel der Analyse der vorliegenden Werte Y_1, \dots, Y_n ist es, die Kurve $A(t)$, die die gesamte Information über den BAK-Abbau in der Eliminationsphase enthält, zu bestimmen. Um die in der Einleitung angeschnittenen Fragen nach der funktionellen Form von $A(t)$ in der späten Eliminationsphase beantworten zu können, müssen wir natürlich annehmen, daß einige der Konzentrationswerte Y_i in der späten Eliminationsphase gemessen wurden. Wir stellen zuerst nichtparametrische Methoden vor, die keine spezifische funktionale, von Parametern abhängige Form (z.B. $A(t) = -\beta t + C_0$ wie bei Widmark) voraussetzen.

Nichtparametrische Methoden

Es wird von der Funktion $A(t)$ lediglich ein gewisser Grad von Glattheit vorausgesetzt, d.h. ein Vorrücken um eine kleine Zeiteinheit Δt sollte den BAK-Wert nicht abrupt verändern. Eine nichtparametrische Schätzung, die die Existenz der zweiten Ableitung $A''(t)$ voraussetzt, ist der Glättungsspline (De Boor 1978). Die Schätzkurve $\hat{A}(t)$ ist durch Lösung des folgenden Minimierungsproblems

$$(2.2) \quad p \sum_{i=1}^n \left(\frac{Y_i - A(t_i)}{\delta_i} \right)^2 + (1-p) \int_{t_1}^{t_n} (A''(t))^2 dt \stackrel{!}{=} \min$$

bestimmt. Der erste Term dieser Gleichung stellt ein Maß für die Datentreue dar, der zweite Term ist ein Maß für die Glattheit der Funktion. Hierbei ist δ_i proportional zur Varianz des Fehlers ϵ_i zu wählen. Unter der Annahme, daß alle Fehler ϵ_i gleich verteilt sind, kann $\delta_i = 1$ gesetzt werden. Der Parameter p balanciert das Verhältnis zwischen Glattheit und Datentreue aus, d.h. wenn wir den Glättungsparameter p nahe bei Null wählen, wird eine sehr glatte, nahezu lineare Funktion entstehen. Ist umgekehrt p nahe bei 1, wird eine sehr wellige Kurve entstehen, da wir zuviel Datentreue verlangen. Eine Einstellung von p muß nach Erfahrungswerten unter Berücksichtigung der physiologischen Zusammenhänge und Beobachtungen geschehen.

Eine weitere nichtparametrische Methode zur Kurvenschätzung ist die Kernschätzung (Gasser u. Rosenblatt 1979). Hierbei wird $A(t)$ durch den folgenden Ausdruck geschätzt:

$$(2.3) \quad A(t) = \frac{\sum_{i=1}^n K((t-t_i)/h_n) \cdot y_i}{\sum_{i=1}^n K((t-t_i)/h_n)}, \quad h_n > 0$$

K bezeichnet hier einen "Kern", eine symmetrische, stetige Funktion mit $\int K^2(t) dt < \infty$. h_n spielt in (2.3) dieselbe Rolle wie p in (2.2), nämlich die Rolle des Austarierens zwischen Datentreue und Glattheit von $A(t)$. Ein h_n zu nahe bei 0 führt zu welligen Funktionen, wählt man hingegen h_n zu groß, ergäbe sich eine nahezu konstante Funktion.

Die Anwendung der Methode (2.2) auf reale BAK-Daten führt zu einer klaren Ablehnung des Widmark-Modells (1.1) über einen gewissen Zeitpunkt in der späten Eliminationsphase hinaus (Abb. 1).

Parametrische Methoden

In diesen Methoden wird eine festgelegte funktionale Form von $A(t)$, abhängig von einem (evtl. mehrdimensionalen) Parameter θ angenommen:

$$(2.4) \quad A(t) = f(t; \theta)$$

wobei θ der zu schätzende Parameter ist und $f(t; \theta)$ eine feste Modellfunktion bezeichnet. Der Widmark-Ansatz ist z.B. parametrisch, hier ist $\theta = (-\beta, C_0)$ ein zweidimensionaler Parameter und $f(t; \theta) = -\beta t + C_0$.

Ebenso ist die Beschreibung der Pharmakokinetik durch die Michaelis-Menten-Enzymkinetik von parametrischer Form (Wilkinson 1980, Formel (4)):

$$(2.5) \quad C_0 - A(t) + K \log(C_0/A(t)) = V \cdot t,$$

mit $\theta = (C_0, V)$, wobei C_0 die Anfangskonzentration zur Zeit $t = 0$, K die Michaelis-Menten-Konstante, V die maximale Abbaugeschwindigkeit darstellen. Nach Wilkinson wird in der BAK-Literatur das Modell (2.5) als nichtlineare Kinetik bezeichnet. Wir wollen darauf hinweisen, daß dies nicht eine nichtlineare Abhängigkeit der Parameter meint. Denn durch die Transformation

$$\begin{aligned} \tilde{A}(t) &= - (A(t) + K \log A(t)) \\ \tilde{C}_0 &= - (C_0 + K \log C_0) \end{aligned}$$

gelangt man zu

$$\tilde{A}(t) = v \cdot t + \tilde{C}_0,$$

einer linearen Beziehungsgleichung.

Da die Michaelis-Menten-Enzymkinetik nur eine idealisierte Annäherung an die tatsächliche Äthanoeliminierung im menschlichen Individuum darstellt, ist es evident, daß zur Beschreibung der komplexen metabolischen Vorgänge die Beziehung (2.5) nur eine ungefähre Beziehung zwischen zeitlichem Ablauf und Elimination (über die Zeit beobachtet) beschreibt. Insbesondere kann die in forensischer Hinsicht wichtige Frage nach dem Übergang von annähernd konstanten Abbauraten ($\beta_{60} = \text{const.}$, d.h. im linearen Bereich) zu abnehmender Eliminationsrate nicht befriedigend beantwortet werden, denn es gibt im Modell (2.5) keinen ins Auge

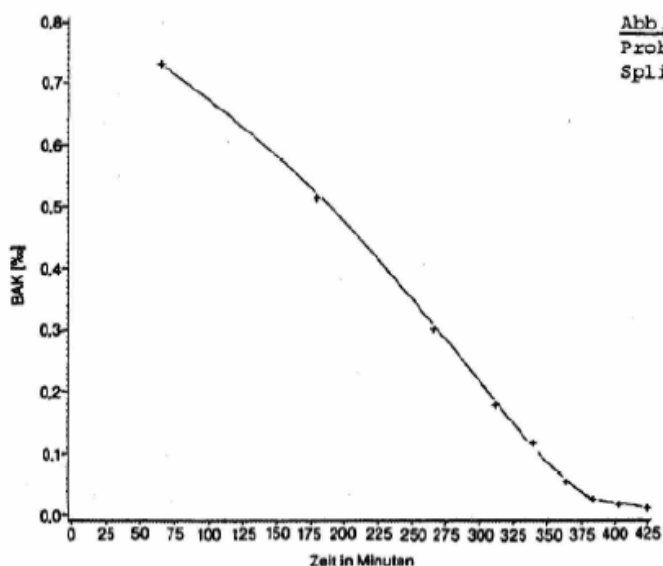


Abb. 1. Blutalkoholkurve eines Probanden. Meßwerte mit Smoothing Spline verbunden

springenden "Knickpunkt" der Eliminationskurve (vgl. Abb.7 Formel (8) Wilkinson 1980).

Wir haben deshalb im Hinblick auf den Verlauf der Enzym-Substratsättigung das folgende Modell näher untersucht und an 24 Probanden angepaßt (vgl. Mattern et al. im Druck). Es basiert auf dem Widmark-Modell und enthält einen Parameter zusätzlich, es gilt ab $t = B$, also erst nach Abschluß der Resorption. Die Phase der Resorption haben z.B. Mallach u. Stärk (1977) mit mathematischen Modellen beschrieben.

$$(2.6) \quad A(t) = \begin{cases} -\beta t + C_B, & t \leq t_1 \\ \alpha \exp(-\gamma(t-t_1)), & t \geq t_1 \end{cases}$$

wobei $\gamma = \beta / (-\beta t_1 + C_B)$, $\alpha = -\beta t_1 + C_B$, die Stetigkeit und Differenzierbarkeit von $A(t)$ in $t = t_1$ garantieren. Das Modell wird also durch 3 Parameter β , C_0 und t_1 bestimmt, die aus Abb.2 ersichtlich sind.

Da der "Knickpunkt t_1 " ein Parameter des Modells ist, bestimmt er sich selbst aus den Beobachtungen. Der Parametervektor θ wurde durch eine nichtlineare Kleinste-Quadrate-Anpassung ermittelt (Prozedur NLIN von SAS 1980). Explizit ausgeführt, bedeuten:

- t_1 = Umschwenkpunkt der Konzentrationskurve, d.h. Übergangspunkt von linearer in verlangsamte, ausklingende Elimination
- α = Konzentration bei t_1
- γ = Skalierungsparameter
- C_B = Konzentrationsniveau nach Beendigung der Resorptionsphase, d.h. bei $t = B$
- β = (positive) zeitunabhängige Abbaurate im linearen Bereich

Im Vergleich zu einem rein linearen Modell wie $A(t) = -\beta t + C_B$ ($t_1 = \infty$) ergab sich für Modell (2.6) bei der erwähnten Probandengruppe eine weit- aus bessere Anpassung an die Meßwerte.

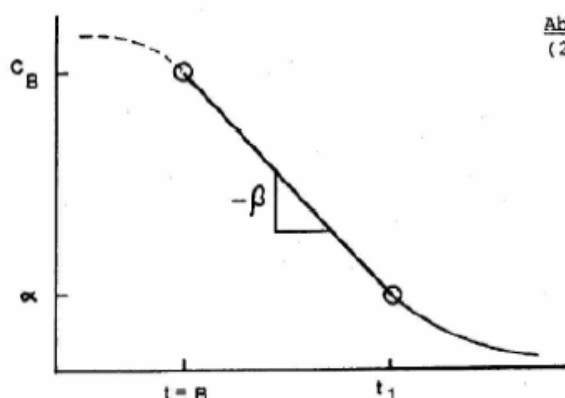


Abb. 2. Graphische Darstellung des Modells (2.6). Definition der Parameter s. Text

Schätzung von BAK-Werten, Rückrechnungsmöglichkeiten

Wir werden uns im folgenden nur mit dem Modell (2.6) beschäftigen. Es ist klar, daß die Abbauraten, mehr noch: der gesamte Parametervektor $\theta = (\beta, C_B, t_1)$, individuell verschieden sein werden. Ein individuelles θ ist vor allem abhängig von der Alkoholverteilung und dem Körpergewicht (Wilkinson 1980). Wir wollen jedoch hier nicht auf diese Problematik eingehen, sondern vielmehr annehmen, daß die interindividuelle Variabilität gering ist, d.h., daß Messungen an einem hinreichend homogenen Probandenkollektiv ausgeführt wurden. Auf der Basis der bereits erwähnten Gruppe von 24 Probanden erhielten wir einen bestimmten Parametervektor $\hat{\theta}$ und eine gewisse Streuung um die so ermittelte Modellkurve. Die in der rechtsmedizinischen Praxis wesentliche "Rückrechnung" soll anhand des Modells (2.6) exemplifiziert werden (Abb. 3).

Aus dem Kollektiv erhielten wir eine mittlere Konzentrationskurve $A(t)$ und eine durch die interindividuelle Variation bedingte Streuung. Wird nun eine Konzentration $C' = A(t')$ gemessen, so kann zu jedem Zeitpunkt t'' die zugehörige mittlere BAK C'' bestimmt werden. Die erwähnte Variation der Probandengruppe erlaubt natürlich nur eine probabilistische Aussage. Nehmen wir an, in Abb. 2 wären 95% Konfidenzintervalle eingezeichnet, so erhält man $P(\underline{C} \leq C'' \leq \bar{C}) = 95\%$, d.h. die Annahme, $C'' < \underline{C}$ oder $C'' > \bar{C}$ ist mit 5% Irrtumswahrscheinlichkeit abzulehnen.

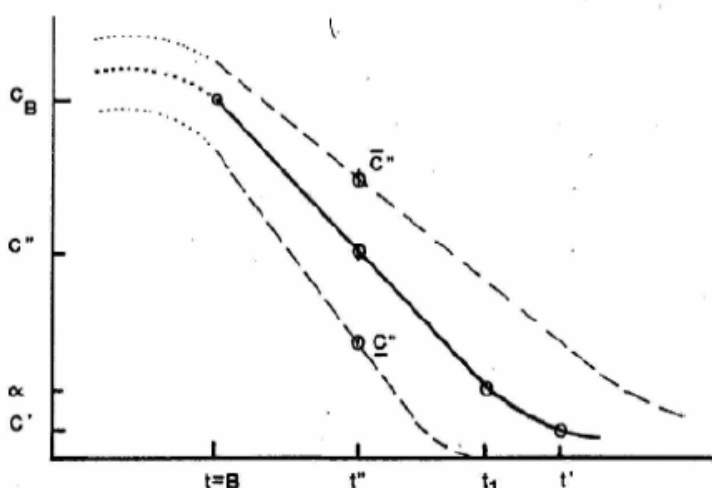


Abb. 3. "Rückrechnung" mit Modell (2.6) unter Beachtung der Konfidenzintervalle (s. Text)

Das hier vorgeschlagene Modell (2.6) erlaubt auf der Basis des erwähnten Kollektivs eine Rückrechnung aus dem nichtlinearen Bereich heraus und gestattet die Schätzung des Umschlagpunktes t_1 sowohl aus der Kenntnis der verstrichenen Zeit nach Resorptionsende, als auch aus der Konzentration. Eine statistisch ausreichend gesicherte Festlegung des Modellvektors θ erfordert die Anpassung des Modells an ein repräsentatives Probandenkollektiv.

Bei einem gegebenen Meßwert in der späten Eliminationsphase läßt sich somit entscheiden, ob er noch im linearen, oder schon im exponentiellen Bereich der Elimination liegt. Nach diesen Ergebnissen sollten in der Praxis unterhalb des Umschlagpunktes t_1 zeitabhängige Rückrechnungswerte verwendet werden. Die Schätzwerte, berechnet auf der Basis von Modell (2.6), dürften nur im Ausnahmefall mehr als 0,1%o niedriger liegen, als bei einer rein linearen Rückrechnung.

Literatur

- Boor C de (1978) A practical guide to splines. Springer, Berlin Heidelberg New York
Gasser T, Rosenblatt M (eds) (1979) Smoothing techniques for curve estimation. Lecture notes 757. Springer, Berlin Heidelberg New York
Mallach HJ, Stärk M (1977) Über Mathematische Funktionen zur näherungsweise Beschreibung von Blutalkoholkurven. Blutalkohol 14:161-171
Mattern R, Bösche J, Birk M, Härdle W (im Druck) Experimentelle Untersuchungen zum Verlauf der Alkoholkurve in der späten Eliminationsphase. Springer, Heidelberg Berlin New York
SAS Users Guide (1980) Statistical Analysis System. SAS Institute, North Carolina, USA
Widmark E (1932) Die theoretischen Grundlagen und die praktische Verwendbarkeit der gerichtlich-medizinischen Alkoholbestimmung. Urban & Schwartzberg, Berlin Wien
Wilkinson P (1980) Pharmacokinetics of ethanol: A review. Alcoholism (N4)4:1

THE NONEXISTENCE OF MOMENTS OF SOME KERNEL REGRESSION ESTIMATORS

by

Wolfgang Härdle¹

Universität Heidelberg, Sonderforschungsbereich 123

and

James Stephen Marron²

University of North Carolina, Chapel Hill

ABSTRACT

In the setting of nonparametric density estimation, it is seen that the moments of kernel based estimators (with high order kernels) may not exist. Thus the popular error criterion of mean square error may be useless in this setting.

KEY WORDS AND PHRASES: Nonparametric regression, kernel estimation, nonexistence of moments.

¹Research partially supported by "Deutsche Forschungsgemeinschaft" and Scientific Research Contract AFOSR-F49620 82 C 0009.

²Research partially supported by Office of Naval Research, Contract N00014-75-C-0809.

Please send all correspondence to:

James Stephen Marron
Department of Statistics
University of North Carolina
Chapel Hill, North Carolina 27514

The (stochastic design) regression estimation problem may be defined as follows. Let $f_{X,Y}(x,y)$ be a joint probability density, and let $f_X(x)$ denote the marginal density of X . Define the regression function,

$$r(x) = E[Y|X=x] = \int y f_{X,Y}(x,y) f_X(x)^{-1} dx .$$

The object is to estimate the function $r(x)$ using a iid sample,

$$(X_1, Y_1), \dots, (X_n, Y_n), \text{ from } f_{X,Y}(x,y).$$

Nadaraya (1964) and Watson (1964) have proposed "kernel estimators" of $r(x)$. These are defined as follows. Given a "kernel function", $K(x)$, and a "bandwidth", h , let

$$\hat{r}(x) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)} .$$

A discussion of this estimator and some related estimators may be found in the survey by Collomb (1981).

The most common means of assessing the accuracy of statistical regression estimators is the Mean Square Error, given by

$$MSE = E[\hat{r}(x) - r(x)]^2 .$$

This paper demonstrates that, under commonly occurring circumstances, MSE is a poor error criterion, because the moments of $\hat{r}(x)$, and hence MSE may fail to exist.

To help put these results in proper perspective, using the notation

$$Z_i = K\left(\frac{x-X_i}{h}\right) ,$$

note that

$$E[\hat{r}(x) | X_1, \dots, X_n] = \frac{\sum_{i=1}^n Z_i r(X_i)}{\sum_{i=1}^n Z_i} . \quad (1)$$

Thus, if the Z_i are nonnegative, $r(X_i)$ is bounded, and $0/0$ is appropriately defined, then $E\hat{r}(x)$ is easily seen to exist. However it is well known, see for example Parzen (1962) or Gasser and Müller (1979) that the rate of convergence for kernel estimators in either regression or density estimation can be greatly improved by allowing K to take on negative values. The rest of this paper is devoted to showing how this practice can easily cause nonexistence of $E\hat{r}(x)$.

Problems arise when the denominator in (1) is very close to 0 but the numerator is not. For example, consider the case $n = 2$. Routine computations show that, if Z_1 and Z_2 are absolutely continuous with respect to Lebesgue measure, if there is a point z_0 such that the densities of Z_1 and Z_2 are bounded above 0 on neighborhoods of both z_0 and $-z_0$, and if $r(x) - r(-x)$ is nonzero on some neighborhood of z_0 , then $E\hat{r}(x)$ fails to exist.

Similar examples may be easily constructed where Z_1 and Z_2 are not absolutely continuous (for example when K is compactly supported), but have an absolutely continuous component which satisfies the above conditions. In the case where Z_1 and Z_2 are discrete (corresponding to K a step function) counter-examples of the above type arise much less naturally, since it is required that for some $c > 0$, K takes on both the values c and $-c$.

For $n > 2$, analogous (but more complicated) examples can be constructed. It should be noted that for reasonable choices of K (ie: more "positive" than "negative") the probability that the denominator of (1) is close to 0 will decrease as n increases. Thus the difficulties discussed in this paper tend to disappear in the limit. However, for each n , MSE may still be undefined and so will not be a reasonable error criterion.

In the case of K a step function (considered by Serfling (1980)), note that for $n > 2$, counterexamples arise much more easily. Indeed, if K takes on the values c_1, \dots, c_k , then little more is required than $\sum_{j=1}^k n_j c_j = 0$ where n_1, \dots, n_k are nonnegative integers whose sum is \bar{n} .

Having seen that MSE can be a treacherous error criterion, one might look for substitutes. A first choice would probably be some sort of truncated MSE. Other approaches may be found in Härdle and Marron (1983) and Marron and Härdle (1983).

REFERENCES

- COLLOMB, G. (1981), Estimation nonparamétrique de la régression: revue bibliographique, Int. Statist. Rev. 49, 75-93.
- GASSER, T. and MÜLLER, H.-G. (1979), Kernel estimation of regression functions, in: T. Gasser and M. Rosenblatt, eds. Smoothing Techniques for Curve Estimation (Lecture Notes in Mathematics 757, Springer Verlag) pp. 23-68.
- HÄRDLE, W. and MARRON, J.S. (1983), Optimal bandwidth selection in nonparametric regression function estimation. North Carolina Institute of Statistics Mimeo Series #1530.
- MARRON, J.S. and HÄRDLE, W. (1983), Random approximations to an error criterion of nonparametric statistics. North Carolina Institute of Statistics Mimeo Series # 1538.
- NADARAYA, E.A. (1964), On estimating regression, Theory Prob. Applic. 9, 141-1
- PARZEN, E. (1962), On estimation of a probability density and mode, Ann. Math. Statist. 35, 1065-1076.
- SERFLING, R.J. (1980), Properties and applications of metrics of nonparametric density estimators, in: Coll. Math. Soc. János Bolyai, 32 Nonparametric Statistical Inference, Budapest.
- WATSON, G.S. (1964), Smooth regression analysis, Sankhyā A, 26 359-372.

Robust Regression Function Estimation*

WOLFGANG HÄRDLE

University of Heidelberg, Heidelberg, Federal Republic of Germany

Communicated by M. Rosenblatt

A robust estimator of the regression function is proposed combining kernel methods as introduced for density estimation and robust location estimation techniques. Weak and strong consistency and asymptotic normality are shown under mild conditions on the kernel sequence. The asymptotic variance is a product from a factor depending only on the kernel and a factor similar to the asymptotic variance in robust estimation of location. The estimation is minimax robust in the sense of Huber (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* 33 73-101.

1. INTRODUCTION

Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ be i.i.d. bivariate random variables with joint distribution function $F(x, y)$ and joint density $f(x, y)$. Let $g(x) = \int f(x, y) dy$ be the marginal density of X and $m(x) = \int yf(x, y) dy/g(x) = E(Y|X=x)$ be the regression function of Y on X . Nadaraya [10] and Watson [20] independently proposed nonparametric estimators of $m(x)$ based on kernel methods as introduced by Rosenblatt [14] and Parzen [12] for density estimation. Specifically the estimates have the form

$$m_n^*(x) = n^{-1}h_n^{-1} \sum_{i=1}^n K((x - X_i)/h_n) Y_i \left/ \left[n^{-1}h_n^{-1} \sum_{j=1}^n K((x - X_j)/h_n) \right] \right., \quad (1)$$

where $K(\cdot)$ is a kernel function and $\{h_n\}$ is a sequence of positive numbers ("bandwidths") tending to zero as n tends to infinity.

A more general estimate as defined in (1) is given by

$$m_n^*(x) = n^{-1} \sum_{i=1}^n \delta_n(x - X_i) Y_i \left/ n^{-1} \sum_{j=1}^n \delta_n(x - X_j) \right., \quad (2)$$

Received June 1, 1982.

AMS subject classifications: 62F35; 62G05, 62J02.

Keywords and phrases: Nonparametric regression, kernel estimation, robust smoothing.

* This work was made possible through the Deutsche Forschungsgemeinschaft and is part of the author's doctoral dissertation.

(A2) Let $f(y|x) = f(x,y)/g(x)$ the conditional probability density function of Y given X be symmetric and having bounded partial derivative $(\partial^2/\partial x^2)f(y|x)$, $x \in I$. From $g(x)$, the marginal density of X , assume that $\inf_{x \in I} g(x) \geq c_0 > 0$ and $(\partial^2/\partial x^2)g(x)$, $x \in I$, exists. The set I here is supposed to be a compact interval of the real line.

(A3) Let $\alpha_n = \int \delta_n^2(u) du < \infty$ for each n and let $\alpha_n/n \rightarrow 0$ as $n \rightarrow \infty$, $\{\delta_n(\cdot)\}$ denoting a positive DFS.

In all statements that follow, x is assumed to be in the interval I . The robust estimator, defined by Eq. (3) for functions satisfying (A1) will be denoted by $m_n(x)$. Various choices of ψ functions may be used in defining $m_n(x)$, such as Huber's ψ function [7]

$$\psi(u) = \max\{-\kappa, \min\{u, \kappa\}\}, \quad \kappa > 0,$$

or an arctan-like curve. Many more examples may be found in Andrews *et al.* [1] or Hampel [6].

Assumption (A1) excludes for the moment those ψ functions which bend down to zero again as $|u| \rightarrow \infty$, such as, $\psi(u) \cong u/(1+u^2)$. It will be shown in the results below that nonmonotone ψ functions will also produce consistent estimators, provided some additional requirements are fulfilled. In the next section the consistency and the asymptotic normality of $m_n(x)$ is shown. A short discussion of the asymptotic variance of $m_n(x)$ under minimax optimality considerations is carried out in Section 3.

2. CONSISTENCY AND ASYMPTOTIC NORMALITY

The use of delta function sequences in regression function estimation goes back to Watson and Leadbetter [21-23], also the following lemmas are due to Watson [20].

LEMMA 2.1. Let $\{\delta_n(\cdot)\}$ be a DFS, such that, $\alpha_n(p) = \int |\delta_n(u)|^p du < \infty$ for all n . Then $\alpha_n(p) \rightarrow \infty$ and

$$\{\delta_{n,p}(u)\} = \{|\delta_n(u)|^p/\alpha_n(p)\},$$

is again a DFS.

LEMMA 2.2. Suppose that $h(u)$ is integrable and continuous at $u=0$, and let $\{\delta_n(\cdot)\}$ be a DFS, then $h(\cdot)\delta_n(\cdot)$ is integrable and

$$\int h(u)\delta_n(u) du \rightarrow h(0) \quad \text{as } n \rightarrow \infty.$$

LEMMA 2.3. Suppose that $h(u)$ is an integrable function, continuous at x and x' , where $x \neq x'$, then $\delta_n(x - \cdot) \delta_n(x' - \cdot) h(\cdot)$ is integrable and

$$\int \delta_n(x - u) \delta_n(x' - u) h(u) du \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Define $H_n(x, s) = n^{-1} \sum_{i=1}^n \delta_n(x - X_i) \psi(Y_i - s)$ and $H(x, s) = E(\psi(Y - s) | X = x) \cdot g(x)$. We first show that $H_n(x, s)$ converges to $H(x, s)$ in probability and almost surely. The weak and strong consistency of $m_n(x)$ will then follow by Huber's technique [7], using the monotony of the ψ function.

LEMMA 2.4. Suppose that assumptions (A1) to (A3) hold, then

$$H_n(x, s) \xrightarrow{P} H(x, s) \quad \text{for each } x \in I \text{ and } s \in \mathbb{R}.$$

If in addition

$$\sum_{n=1}^{\infty} \alpha_n n^{-2} < \infty \quad \text{for the DFS } \{\delta_n(\cdot)\},$$

then $H_n(x, s) \rightarrow H(x, s)$ a.s.

Proof. Using Chebyshev's inequality and the boundedness of ψ it follows that

$$P(|H_n(x, s) - EH_n(x, s)| > \varepsilon) \leq C \cdot \alpha_n/n,$$

C denoting a constant depending on ε and the upper bound of ψ . Since $EH_n(x, s) = E\delta_n(x - X) \psi(Y - s) = \int \delta_n(x - u) E(\psi(Y - s) | X = u) g(u) du$, it follows from Lemma 2.2 and the smoothness assumption (A2) that $EH_n(x, s) \rightarrow H(x, s)$ as $n \rightarrow \infty$. So the first assertion of the lemma is shown.

To show the strong convergence of $H_n(x, s)$ to $H(x, s)$, define

$$\sigma_n^2 = \int \delta_n^2(x - u) E(\psi^2(Y - s) | X = u) g(u) du,$$

$$e_n = \int \delta_n(x - u) E(\psi(Y - s) | X = u) g(u) du = E[\delta_n(x - X) \psi(Y - s)],$$

and

$$Z_{i,n} = \delta_n(x - X_i) \psi(Y_i - s) - e_n.$$

The $\{Z_{i,n}\}$ are a mean zero i.i.d. triangular sequence, so if we use $EZ_{i,n}^2 = \sigma_n^2 - e_n^2 \leq \alpha_n \cdot C$ (Lemmas 2.1, 2.2, (A3)) it is clear from the assumption of the lemma that the SLLN applies (Serfling [18, p. 27]). As

already shown $e_n \rightarrow H(x, s)$, hence also the second assertion of the lemma is shown.

Since Eq. (3) may have several solutions we will take the estimate $m_n(x)$ as one member of the set of solutions. By assumption (A3), the positivity of the DFS, Lemma 1 of Huber [7] applies and we have

LEMMA 2.5. *The set of solutions of (3), denoted by $\{m_n(x)\}$ is nonempty and compact and convex.*

Using the same proof as for Lemma 3 in Huber [7] we get the consistency of $m_n(x)$, noting that $g(x)$ is always positive.

THEOREM 2.1. *Suppose that (A1) to (A3) hold, then $m_n(x)$ is weakly consistent, i.e., $m_n(x) \rightarrow^p m(x)$. If, in addition, $\sum_{n=1}^{\infty} \alpha_n/n^2$ is finite, then $m_n(x)$ is strongly consistent, i.e., $m_n(x) \rightarrow m(x)$ a.s.*

For nonmonotone ψ -functions, i.e., ψ functions which are monotone around $u = 0$ but return back to zero as $|u| \rightarrow \infty$, Huber's original proof does not work. Those rebending ψ functions have strong robustness properties since they really cut off bad observations, for instance, Hampel's "three part redescendor" [6],

$$\begin{aligned} \psi(u) &= u, & |u| &\leq a, \\ &= a \cdot \text{sign}(u), & a < |u| &\leq b, \\ &= \frac{c - |u|}{c - b} a, & b < |u| &\leq c, \\ &= 0, & |u| &> c. \end{aligned}$$

It is desirable to obtain consistency for those robust smoothers also. This is done by coupling the solutions of (3) for nonmonotone ψ functions, $\{\tilde{m}_n(x)\}$, together with $m_n(x)$, the robust estimator for monotone ψ . That is, define $\tilde{m}_n(x)$ as that solution of (3) which is nearest to $m_n(x)$, i.e.,

$$|m_n(x) - \tilde{m}_n(x)| = \inf\{|t - m_n(x)| : H_n(x, t) = 0, \psi \text{ not necessarily monotone}\}. \tag{4}$$

By standard arguments $\tilde{m}_n(x)$ will also be strongly (weakly) consistent and we have

COROLLARY 2.1. *Suppose that the assumptions of Theorem 2.1 hold, then $\tilde{m}_n(x)$, defined in (4), is strongly (weakly) consistent.*

For delta function sequences of kernel type we have $\alpha_n \simeq h_n^{-1}$, the inverse of the bandwidth. Assumption (A3) is for DFS of kernel type now, $nh_n \rightarrow \infty$ as $n \rightarrow \infty$. Note that in (4), defining $\tilde{m}_n(x)$, another consistent candidate is provided by $m_n^*(x)$, the Nadaraya-Watson estimate. (Noda [11], Johnston [8]). Coupling $\tilde{m}_n(x)$ to $m_n^*(x)$ would give us an estimator with similar properties as the so-called "M15" (Andrews *et al.* [1]). To formulate the result of the asymptotic normality let us define

$$Z_n(x) = c_1(x) \cdot \left(m_n(x) - m(x) - \frac{B_n(x)}{c_1(x) \cdot g(x)} \right) / [(\alpha_n/n) \sigma^2(x) g^{-1}(x)]^{1/2},$$

where $\sigma^2(x) = E(\psi^2(Y - m(x)) | X = x)$, $c_1(x) = E(\psi'(y - m(x)) | X = x)$, $B_n(x) = EH_n(x)$, and $H_n(x) = H_n(x, m(x))$.

THEOREM 2.2. *Let $\{\delta_n(\cdot)\}$ be a DFS with the properties*

- (1) $\gamma_n = \int |\delta_n(u)|^{2+\eta} du < \infty$ for some $\eta > 0$,
- (2) $\gamma_n = o(n^{\eta/2} \alpha_n^{1+\eta/2})$ as $n \rightarrow \infty$.

Further let x_1, \dots, x_p be p distinct design points, then the random vector

$$(Z_n(x_1), \dots, Z_n(x_p)),$$

converges in distribution to a normally distributed random vector with zero mean and identity covariance matrix.

The proof will be clear from an expansion of $H_n(x, m_n(x))$ around $m(x)$, because by the mean value theorem we have

$$m_n(x) - m(x) = H_n(x)/D_n(x), \tag{5}$$

where $D_n(x) = n^{-1} \sum_{i=1}^n \delta_n(x - X_i) \psi'(Y_i - m(x) + w_i(m_n(x) - m(x)))$, $w_i \in (0, 1)$. From the WLLN, Theorem 2.1, and the boundedness of ψ' it is clear that $D_n(x) \rightarrow^p E(\psi'(y - m(x)) | X = x) \cdot g(x) = c_1(x) \cdot g(x)$. We therefore only have to prove that $(W_n(x_1), \dots, W_n(x_p))'$, where

$$W_n(x) = (H_n(x) - B_n(x)) / \left[\left(\frac{\alpha_n}{n} \right) \sigma^2(x) g(x) \right]^{1/2}, \tag{6}$$

is asymptotically normally distributed with mean zero and identity covariance matrix.

PROPOSITION 2.1. *Suppose that the assumptions of Theorem 2.2 hold, then*

$$\mathbf{W} = (W_n(x_1), \dots, W_n(x_p))',$$

when $W_n(x)$ as defined in (6), converges in distribution to a standard normal vector.

Proof. The random vector \mathbf{W} is asymptotically normally distributed if and only if each linear combination of its coordinate random variables is (one dimensional) asymptotically normally distributed. (Cramer-Wold device, Billingsley [2]). So if we show that

$$\sum_{k=1}^p t_k W_n(x_k) \xrightarrow{L} N\left(0, \sum_{k=1}^p t_k^2\right),$$

for each set of real numbers t_1, \dots, t_p , the proposition follows.

By definition (6) this is equivalent to showing

$$\sum_{k=1}^p t_k [H_n(x_k) - B_n(x_k)] \Big/ \left[(\alpha_n/n) \sum_{i=1}^p \sigma^2(x_i) \cdot g(x_i) \right]^{1/2} \xrightarrow{L} N\left(0, \sum_{k=1}^p t_k^2\right).$$

From Lemmas 2.1 and 2.2 it is clear that

$$(n/\alpha_n) \text{var}(H_n(x_j) - B_n(x_j)) \rightarrow \sigma^2(x_j) g(x_j). \tag{7}$$

And that for $j \neq k$,

$$(n/\alpha_n) \text{cov}(H_n(x_j) - B_n(x_j), H_n(x_k) - B_n(x_k)) \rightarrow 0 \quad \text{as } n \rightarrow \infty. \tag{8}$$

So, defining $\tilde{H}_n(x) = H_n(x) - B_n(x)$, it remains to show that

$$Z_n = \sum_{k=1}^p t_k \tilde{H}_n(x_k) \Big/ \left(\text{var} \sum_{k=1}^p t_k \tilde{H}_n(x_k) \right)^{1/2} \xrightarrow{L} N(0, 1).$$

Interchanging the sums in this expression gives

$$Z_n = \sum_{i=1}^n Z_{n,i},$$

where $Z_{n,i} = n^{-1/2} s_n^{-1} \sum_{k=1}^p t_k [\delta_n(x_k - X_i) \psi(Y_i - m(x_k)) - B_n(x_k)]$ and $s_n^2 = \text{var}(\sum_{k=1}^p t_k \delta_n(x_k - X) \psi(Y - m(x_k)))$. Since the random variables $Z_{n,i}$ are independent identically distributed it remains to show for an application of the Lindebergh-Feller CLT that for some $\eta > 0$,

$$nE |Z_{n,1}|^{2+\eta} \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Using Loève's C_r inequality (Loève [9]) we obtain

$$\begin{aligned} nE |Z_{n,1}|^{2+\eta} &\leq \sum_{k=1}^p C_k(\eta) E |\psi(y - m(x_k)) \delta_n(x_k - X) - B_n(x_k)|^{2+\eta} \\ &\quad \times (n^{n/2} s_n^{2+\eta})^{-1} \\ &= \sum_{k=1}^p c_k(\eta) R_{k,n}. \end{aligned}$$

Since $s_n^2 = n \text{var}(\sum_{k=1}^p t_k \tilde{H}_n(x_k)) \sim \alpha_n \sum_{k=1}^p t_k^2 \sigma^2(x_k) g(x_k)$ by the asymptotic relations (7), (8) it suffices to consider the numerator of $R_{k,n}$.

$$\begin{aligned} &E |\delta_n(x - X) \psi(Y - m(x)) - B_n(x)|^{2+\eta} \\ &= \gamma_n \int \delta_n^*(x - u) E(|\psi(y - m(x)) - B_n(x)|^{2+\eta} | X = u) g(u) du \\ &\leq \gamma_n \cdot C, \quad \delta_n^*(\cdot) = |\delta_n(\cdot)|^{2+\eta} / \gamma_n, \end{aligned}$$

since ψ is bounded and $\delta_n^*(\cdot)$ is again a DFS by Lemma 2.1. So finally

$$R_{k,n} = O \left(\frac{\gamma_n}{n^{n/2} \alpha_n^{1+n/2} \sum_{k=1}^p t_k^2 \sigma^2(x_k) g(x_k)^{1+n/2}} \right),$$

and by assumption (2) of the theorem $R_{k,n} \rightarrow 0$ as $n \rightarrow \infty$. This completes the proof.

In practical applications, if we are interested in constructing confidence bans, Theorem 2.2 does not help us since we neither know the bias $B_n(x)$ nor $f(y|x)$. In order to drop the bias term $B_n(x)$ when constructing asymptotic confidence intervals we have to ensure that $(n/\alpha_n)^{1/2} B_n(x)$ vanishes as $n \rightarrow \infty$. For this purpose the following condition on a DFS $\{\delta_n(\cdot)\}$ will be convenient. Let

$$(B') \quad \int [\delta_n(x - u) l(u) - l(x)] du = o((\alpha_n/n)^{-1/2}),$$

for each twice differentiable function $l(\cdot)$ with bounded second derivative. Let us note that this condition reduces to $\int u^2 \delta_n(u) du = o((\alpha_n/n)^{-1/2})$, if we use a symmetric DFS, i.e., $\delta_n(u) = \delta_n(-u)$.

Reading through the proof of Proposition 2.1 it will be clear that the term $B_n(x)$ may be dropped under assumption (B'). The asymptotic relation (7) changes now to

$$(n/\alpha_n) E(H_n^2(x)) \rightarrow \sigma^2(x) g(x) \quad \text{as } n \rightarrow \infty. \tag{9}$$

Let $V_n(x) = c_1(x)(m_n(x) - m(x))/[(\alpha_n/n) \sigma^2(x) g(x)^{-1}]^{1/2}$, we then obtain

the following corollary, which states the asymptotic normality of $V_n(x)$ without the bias term $B_n(x)$.

COROLLARY 2.2. *Suppose that in addition to the assumption of Theorem 2.1 condition (B') holds, then $(V_n(x_1), \dots, V_n(x_p))$ is asymptotically normally distributed with zero mean and identity covariance matrix.*

In the case that DFS of kernel type are used, assumption (B') is easily translated into expressions involving only the sample size n and the bandwidth h_n . Assume that the kernel satisfies

(K) K is a continuous function with compact support $[-A, A]$,

$$\int K(u) du = 1, \quad \int uK(u) du = 0, \quad \int u^2K(u) du < \infty.$$

Under this assumption it follows by Taylor expansion of $l(\cdot)$ that (B') is equivalent to $nh_n^5 \rightarrow 0$ for DFS of kernel type. This condition is evidently necessary for the asymptotic negligence of the bias $B_n(x)$. By Taylor expansion one obtains from (K) that $B_n(x) = O(h_n^2)$ (Stützle and Mittal [19]), so we have to assume that $(nh_n)^{1/2} h_n^2 \rightarrow 0$ which is equivalent to the above-mentioned condition. This condition also occurs in the work of Schuster [17] and Johnston [8]. Intuitively, it would seem that the bias becomes important if the regression curve has a large second derivative $m''(x)$. But as Stützle and Mittal [19] show, the bias is $h_n^2 \cdot m''(x) \cdot c_1(x)$, i.e., the bias and the rate of convergence are the same as one would obtain with $m_n^*(x)$. Summarizing we obtain the following theorem involving DFS of kernel type.

THEOREM 2.3. *Suppose that the following conditions hold:*

- (1) $nh_n^5 \rightarrow 0$,
- (2) $\gamma = \int |K(u)|^{2+\eta} du < \infty$,

then $(V_n(x_1), \dots, V_n(x_p))$ converges in distribution to a multivariate normal random vector with zero mean and identity covariance matrix, where

$$V_n(x) = c_1(x)(m_n(x) - m(x)) \left/ \left[(nh_n g(x))^{-1} \cdot \int K^2(u) du \sigma^2(x) \right]^{1/2} \right.$$

The proof is clear by the previous remarks and the equality $\alpha_n = (\int K^2(u) du) h_n^{-1}$ for DFS of kernel type. The asymptotic variance $V_x(\psi, f)$ admits an intuitive interpretation of what robust smoothing is doing. The asymptotic variance is

$$V_x(\psi, f) = R_1(x) \cdot \int K^2(u) du / g(x), \quad (10)$$

where

$$R_1(x) = \frac{E(\psi^2(y - m(x)) | X = x)}{(E(\psi'(y - m(x)) | X = x))^2},$$

is due to the robustness of the estimate $m_n(x)$ and $\int K^2(u) du/g(x)$ is due to the smoothing property of $m_n(x)$. (see Schuster [17], Nadaraya [10], and Collomb [4]). As far as the Nadaraya-Watson estimate $m_n^*(x)$ is concerned, the optimization of the asymptotic variance of $m_n^*(x)$ was concentrated on the "smoothing part" of the asymptotic variance $\int K^2(u) du$. From Table 1 in Rosenblatt [16] it is evident that the use of optimal kernels does not gain very much in relative efficiency. For instance, the ratio of the asymptotic variance of the optimal kernel $K(u) = 0.75(1 - u^2)$ (Epanechnikov [5]) to the asymptotic variance obtained from the simple uniform window

$$K(u) = \frac{1}{2}I_{[-1,1]}(u),$$

is 1.077. It becomes more important to optimize $R_1(x)$ in the asymptotic variance, since this factor may dominate the "smoothing part" $\int K^2(u) du$ in the case of heavy-tailed conditional distributions. It is clear from Table 1 in Huber [7] that in the case of extreme outlier contamination $R_1(x)$ may be the half of $\text{var}(Y | X = x)$ which is the corresponding factor to $R_1(x)$ if $m_n^*(x)$ is used. The optimization and minimax consideration $R_1(x)$ with respect to a contamination model is the topic of the next section.

3. MINIMAX ROBUSTNESS

The contamination model (for fixed x) is formalized as

$$\begin{aligned} \mathcal{M}(x) = \{ & f(x, y) = f(y | x) \cdot g(x) \mid g(\cdot) \text{ fixed,} \\ & f(y | x) = (1 - \varepsilon(x))\tilde{f}(y - m(x)) + \varepsilon(x)h(y - m(x)), \\ & \tilde{f}, h \text{ symmetric, } -\log \tilde{f}(\cdot - m(x)) \text{ convex,} \\ & 0 < \varepsilon(x) < 1, \tilde{f} \text{ fixed, } h \text{ arbitrary} \}, \end{aligned} \quad (11)$$

which is exactly the same contamination model that is used in robust estimation of location, except that here the contamination rate $\varepsilon(x)$ depends on x and a marginal density $g(x)$ is involved. Noting that the asymptotic variance $V_x(\psi, f)$ splits up into the factors $R_1(x)$ and $\int K^2(u) du/g(x)$ where the latter is independent of ψ and f , we obtain from Huber's theory the following result.

THEOREM 3.1. *Let $\mathcal{M}(x)$ be the class of distributions as defined in (11), with a fixed marginal density $g(x)$. Then the asymptotic variance has a saddlepoint. There is an $f_0(x, y)$ and a ψ_0 such that*

$$\sup_{\mathcal{M}(x)} V_x(\psi_0, f) = V_x(\psi_0, f_0) = \inf_{\psi} V(\psi, f_0).$$

Let $t_0(x) < t_1(x)$ be the endpoints of the interval where

$$\left| \frac{\partial \tilde{f}(y - m(x))}{\partial y} / \tilde{f}(y - m(x)) \right| \leq \kappa(x),$$

and $\kappa(x)$ is related to $\varepsilon(x)$ by

$$(1 - \varepsilon(x))^{-1} = \int_{t_0(x)}^{t_1(x)} \tilde{f}(y - m(x)) dy + [\tilde{f}(t_0(x) - m(x)) + \tilde{f}(t_1(x) - m(x))] / \kappa(x).$$

Then $f_0(x, y)$ can be computed as

$$\begin{aligned} f_0(x, y) &= (1 - \varepsilon(x)) \tilde{f}(t_0(x) - m(x)) g(x) \\ &\quad \cdot \exp(\kappa(x)(y - m(x) - t_0(x))), \quad y \leq t_0(x), \\ &= (1 - \varepsilon(x)) \tilde{f}(y - m(x)) g(x), \quad t_0(x) < y < t_1(x), \\ &= (1 - \varepsilon(x)) \tilde{f}(t_1(x) - m(x)) g(x) \\ &\quad \cdot \exp(-\kappa(x)(y - m(x) - t_1(x))), \quad y \geq t_1(x), \end{aligned}$$

and $\psi_0(y, x) = -(\partial f_0(x, y) / \partial y) / f_0(x, y)$, which is monotone and bounded.

The proof of the theorem is the same as in Huber [7]. We only have to cope with the dependence on x . The same minimax calculus may also be carried out with other contamination models (Portnoy [13], Collins [3]) leading to asymmetric or nonmonotone ψ functions. For given x , $m_n(x)$ is in fact a robust "location" estimate, therefore after the computation of the asymptotic variance the theory on robust estimation of location applies. Instead of minimizing $V_x(\psi, f)$ one might also use a weighted uniform loss such as

$$R(\psi, f) = \int V_x(\psi, f) g(x) dx,$$

but this functional can be optimized in the same way as in Theorem 3.1, provided the dependence of $V_x(\psi, f)$ on x is smooth.

ACKNOWLEDGMENTS

I would like to thank Professor Gasser and Professor Carroll for many helpful suggestions and remarks. I would also like to thank the referee for helpful remarks on the paper.

REFERENCES

- [1] ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H., AND TUKEY, I. W. (1972). *Robust Estimation of Location*. Princeton Univ. Press, Princeton, N.J.
- [2] BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.
- [3] COLLINS, J. R. (1976). Robust estimation of a location parameter in the presence of asymmetry. *Ann. Statist.* **4** 68-85.
- [4] COLLOMB, H. (1981). Estimation nonparamétrique de la regression. *Revue Bibliographique. Internat. Statist. Rev.* **49** 75-93.
- [5] EPANECHNIKOV, V. A. (1969). Nonparametric estimation of a multivariate probability density. *Theory Probab. Appl.* **14** 153-158.
- [6] HAMPEL, F. R. (1973). Robust estimation. A condensed partial survey. *Z. Wahrsch. Verw. Gebiete* **27** 87-104.
- [7] HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **33** 73-101.
- [8] JOHNSTON, G. J. (1979). Smooth nonparametric regression analysis. Ph.D. thesis, Univ. of North Carolina.
- [9] LOÈVE, M. (1977). *Probability Theory I*. Springer-Verlag, Berlin/New York.
- [10] NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141-142.
- [11] NODA, K. (1976). Estimation of a regression function by the Parzen kernel type density estimators. *Ann. Inst. Math. Statist.* 221-234.
- [12] PARZEN, E. (1962). On estimation of a probability density function. *Ann. Math. Statist.* **31** 1065-1076.
- [13] PORTNOY, S. L. (1956). Robust estimation in dependent situation. *Ann. Statist.* **5** 22-43.
- [14] ROSENBLATT, M. (1956). Remarks on some nonparametric estimators of a density function. *Ann. Math. Statist.* **27** 832-837.
- [15] ROSENBLATT, M. (1968). Conditional probability density and regression estimators. In *Multivariate Analysis, II*, (P. R. Krishnaiah, Ed.), Academic Press, New York.
- [16] ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815-1842.
- [17] SCHUSTER, E. F. (1972). Joint asymptotic distribution of the estimated regression function at a finite number of distinct points. *Ann. Math. Statist.* **43** 84-88.
- [18] SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.
- [19] STÜTZLE, W., AND MITTAL, Y. (1979). Some comments on the asymptotic behaviour of robust smoothers. In *Smoothing Techniques for Curve Estimation* (T. Gasser and M. Rosenblatt, Eds.), Lecture Notes in Mathematics, No. 757, Springer-Verlag, Heidelberg.
- [20] WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā, Ser. A* **26** 359-372.
- [21] WATSON, G. S., AND LEADBETTER, M. R. (1964). On the estimation of the probability density I. *Ann. Math. Statist.* **34** 480-491.
- [22] WATSON, G. S., AND LEADBETTER, M. R. (1964). Hazard analysis I. *Biometrika* **51** 175-184.
- [23] WATSON, G. S., AND LEADBETTER, M. R. (1964). Hazard analysis II. *Sankhyā* **26** 101-116.

Printed by the St. Catherine Press Ltd., Tempelhof 41, Bruges, Belgium

EEG-RESPONSIVENESS TO EYE OPENING AND CLOSING IN MILDLY RETARDED CHILDREN COMPARED TO A CONTROL GROUP *

Wolfgang HÄRDLE

Institut für Angewandte Mathematik, Universität Heidelberg, D-6900 Heidelberg 1, FRG

Theo GASSER and Petra BÄCHER

Zentralinstitut für Seelische Gesundheit, D-6800 Mannheim 1, FRG

Accepted for publication 31 August 1983

Changes in the ongoing EEG when repeatedly closing and opening the eyes are quantified and compared for a group of mildly retarded children and a matched control group. The most prominent changes occur for the α -rhythm at posterior derivations. Blocking of the α is faster than its restriction, but there is no group difference in this respect. Many of the amplitude changes are also quite similar for the two groups. The differences found are associated with a lowered arousal by the experimental group.

1. Introduction

There is general agreement that mentally retarded children constitute an etiologically heterogeneous group. It is, however, a matter of debate to what extent some kind of neurophysiological dysfunction contributes to or is even decisive for the assignment to special schools, and how far socio-cultural factors are of major influence. In a previous study (Gasser, Möcks, Lenard, Bächer and Verleger, 1983a) a group of mildly retarded children differed significantly from a matched control group in a number of parameters of the EEG at rest which are known to be of developmental relevance. The same EEG parameters also showed sizable correlations within the mildly retarded group with performance in intelligence tests indicating that retarded children

* This work has been performed as part of the research program of the Sonderforschungsbereich 116 (project M2) and the Sonderforschungsbereich 123 (project B1), both at the University of Heidelberg, and was made possible by financial support from the Deutsche Forschungsgemeinschaft.

** Address for correspondence and reprint requests: Theo Gasser, Zentralinstitut für Seelische Gesundheit, Postfach 5970, D-6800 Mannheim 1, FRG.

with EEG parameters within the normal range achieve on the average higher IQ scores (Gasser, von Lucadou-Müller, Verleger and Bächer, 1983b).

When recording a clinical EEG, it is a common procedure to investigate the effect of eye opening and closing. This was also done in our study with the aim of quantifying the accompanying changes in the EEG; the generation of the α -rhythm and changes in its characteristics are associated anatomically with the reticulo-thalamo-cortical ascending axis (Andersen and Andersson, 1968) and psychophysiologically with the regulation of arousal (Moruzzi and Magoun, 1949). It was considered to be of interest how far mildly retarded children differ in these aspects from a control group, but also to what extent they have a similar pattern of change in their ongoing EEG. Attempts to find differences in these neurophysiological mechanisms have already been made in the sixties (Berkson, Hermelin and O'Connor, 1961; Baumeister, Spain and Ellis, 1963; Wolfensberger and O'Connor, 1965; Baumeister and Hawkins, 1967). Flashes of light were used as a provocation method rather than opening of the eyes. On the whole, no differences to non-retarded subjects could be found. Baumeister and Hawkins (1967), however, showed that the probability of α -blocking is lower with a lower IQ. The work mentioned so far relied on an analysis made by eye which may be too gross to describe adequately these changes in the EEG.

Computerized methods for studying α -attenuation and enhancement have been used in recent years for normal subjects: Aranibar and Pfurtscheller (1978) describe significant time-dependent changes of power within the α -band under 1 sec photic stimulation. Nogawa, Katayama, Tabata, Oshio and Kawahara (1976) quantified changes in the α -amplitude of the individual EEG by means of the demodulation technique, and Kawabata (1972) used nonstationary power spectrum analysis in order to identify time-dependent changes in frequency. The attenuation of the α during problem solving was studied for learning disabled children by Fuller (1977). When comparing the changes in the ongoing EEG due to consecutive eye opening and closing, we had the following hypotheses and questions in mind:

- (1) The latency of suppressing and of restituting the alpha is greater for the experimental group.
- (2) The amplitude changes from eyes opened to eyes closed are more drastic for the control group.
- (3) The question arises then how changes in amplitude are distributed over frequency bands.
- (4) Differences in rhythmic activity (dominant frequency of the α and its relative power, and also of the 'driving force' of the EEG) between the two groups will be quantified and a comparison with the EEG at rest is sought in these parameters.
- (5) A habituation effect is expected to be more pronounced for the experimental group when repeatedly passing through the on and off epochs.

Whenever appropriate, a comparison was made with parameters characterizing the EEG at rest; in a system-theoretic framework the EEG with eyes closed in the on-off experiment is equivalent to a transient state while the EEG at rest corresponds to a steady state. Complex demodulation (to be described in the next section) at parieto-occipital derivations was used for answering (1). EEG amplitude was quantified by total power (from 1.5 to 25.0 Hz) and by broad band spectrum parameters (Matoušek and Petersen, 1973). In order to characterize properties of rhythmic activity, autoregressive model building was used in a way suggested by Zetterberg (1969) with modifications introduced in Steinberg, Gasser and Franke (1983).

2. Subjects and methods

2.1. Subjects

The experimental group (EG) of this investigation consisted of 25 children – 14 children attending a school for the mentally retarded (MR) and 11 children attending a school for the learning disabled (LD). Out of 35 children identified in an epidemiological survey as being 10–13 years old and having IQs of 50–70 (Lipmann, 1979), these 25 children participated in our study. According to the ICD classification (WHO, 1977), this experimental group coincides with the subclass 'mild mental retardation'. The control group (CG) was individually matched for sex (11 boys, 14 girls) and age, and across the sample for socioeconomic status, as measured by a prestige score (Treiman, 1975) of the parents' occupation. The mean age was 12 years 6 months and the mean prestige score was in the range of the lower social class for both groups (it was somewhat higher for the CG). The control group was drawn from the general population and a small reward was paid to increase participation. Children receiving medication affecting the EEG were excluded from the study and no child had previous neurological treatment as judged from an interview with the mother.

2.2. Electrodes and recording

Beckmann miniature Ag/AgCl electrodes were fixed with GRASS EC2 cream at F₄, F₃, C₄, C₃, C_z, P_z, O₂, O₁ and at both earlobes as linked reference. In addition the vertical electrooculogram (EOG) was recorded bipolarly below and above the right eye. Resistances as measured on the amplifier Schwarzer encephaloscript 1630 were on the average 13 K Ω in the EG and 11.5 K Ω in the CG.

The on-off experiment consisted of six 10 sec blocks starting with 10 sec of eyes opened, followed by 10 sec of closed eyes, and so on. The EEG technician

set a marker on the paper trace at the moment she gave the order, these time points were also registered on a timer channel of the analog tape. The analog data was later digitized with a frequency of 68 Hz, with analog filtering (Krohn-Hite, 32 Hz low pass) as an intermediate step.

2.3. Data processing

All the computations were done off-line on an IBM 370/168 at the computing center of the University of Heidelberg, using our own software. We first rejected the blocks with gross artifacts which led to the exclusion of 2 subjects of the EG. The time of eye opening and closing was determined by displaying the EOG channel on a Tektronix 4014 terminal in the interactive mode. It proved to be inadequate to use the time when the order was given since the EG took a longer time to react than the CG. The time needed for the attenuation and restitution of the α -rhythm was determined by computer via complex demodulation (Walter, 1969); the demodulated signal was determined with respect to the individual α main frequency. The reaction time for blocking was then defined as the time needed to suppress the demodulated signal to 25% of the average amplitude in the preceding block with eyes closed. Restitution of the α after closing the eyes by definition occurred when 75% of the average amplitude of the demodulated signal of the respective block were reached. Here and in the following, individual results were obtained by averaging over the three consecutive phases of eyes opening and closing.

The next step consisted of spectral analysis via the Fast Fourier Transform separately for each phase, yielding total power from 1.5 to 25.0 Hz, and also the conventional broad band parameters δ (1.5–3.5 Hz), θ (3.5–7.5 Hz), α_1 (7.5–9.5 Hz), α_2 (9.5–12.5 Hz), β_1 (12.5–17.5 Hz) and β_2 (17.5–25.0 Hz), as defined by Matoušek and Petersén (1973). Results for relative power only will be presented.

Next an autoregressive (AR) process was fitted to each of the six blocks separately for each individual (Zetterberg, 1969; Koopmans, 1974). For simplicity of notation, the model is defined for the first two blocks only:

$Y_1(t)$ = observed EEG for eyes opened

$Y_2(t)$ = observed EEG for eyes closed

$$Y_j(t) + \alpha_1 Y_j(t-1) + \dots + \alpha_p Y_j(t-p_j) = \eta_j(t),$$

where:

$j = 1, 2$ for model for eyes opened response and eyes closed respectively,

α_i = autoregressive coefficients,

p_j = order of the AR-process, and

$\eta_j(t)$ = white noise process, with innovation variance σ_j^2

The white noise process η can be loosely interpreted as the driving process – with power σ^2 – of an active filter characterized by the coefficients $\alpha_1, \dots, \alpha_p$. At this stage of research this cannot be considered to be but a crude phenomenological model of the EEG. The order p was determined from the data following a criterion by Schwarz (1978) (which should avoid over- and under-fitting, see appendix 1). Criteria suggested by Akaike (1969) and Hannan and Quinn (1979) were used for comparison. Statistics dealing with the estimation of the innovation variance are quoted in appendix 2. Note that the driving process η was fully determined by its variance if it had a Gaussian structure. The autoregressive parameters have been transformed such as to yield the dominant frequency and the power content of the rhythms which they describe. For an α -rhythm at parieto-occipital derivations we required that the peak frequency falls in the interval (7.5–12.5 Hz) and also a minimal power content of 10%.

Habituation was studied by computing the ratios between phases 1 and 2, 1 and 3 and also 2 and 3 in the above quantities and by defining the average of these three ratios to be a habituation index.

To check for the statistical significance of differences between groups the two-sample Wilcoxon-test was used and a p -value of 0.05 was considered significant. Within group differences were tested by the one-sample Wilcoxon test.

3. Results

Phases contaminated by artifacts were excluded from further analysis; for two subjects from the EG no phase with adequate quality was left which reduced the number to $n = 23$.

The systematic changes in the spectral components of the EEG for eyes opened/eyes closed are illustrated in a 'running spectral plot' (fig. 1). From these slightly smoothed spectra the drastic changes to be quantified below, in total power, and specifically in the α -band, can be readily seen.

Results for the reaction times (RT) for α -blocking and α -restitution at derivations P_2 , O_2 and O_1 are assembled in table 1 (they were determined automatically by complex demodulation, i.e. by suitable filtering around the individual α -frequency, see methods). Means and standard deviations are of comparable size for the two groups, and, as a consequence, differences were far from statistical significance. The RT for enhancing the α -rhythm are larger than for attenuating and this difference is statistically significant for both groups.

Let us now focus on measures of amplitude of the EEG: Table 2 contains the results for total power (from 1.5 to 25.0 Hz) for eyes opened and eyes closed in the 'on-off experiment' and also for the EEG at rest with eyes closed. The largest and most significant differences between the EG and the CG occur

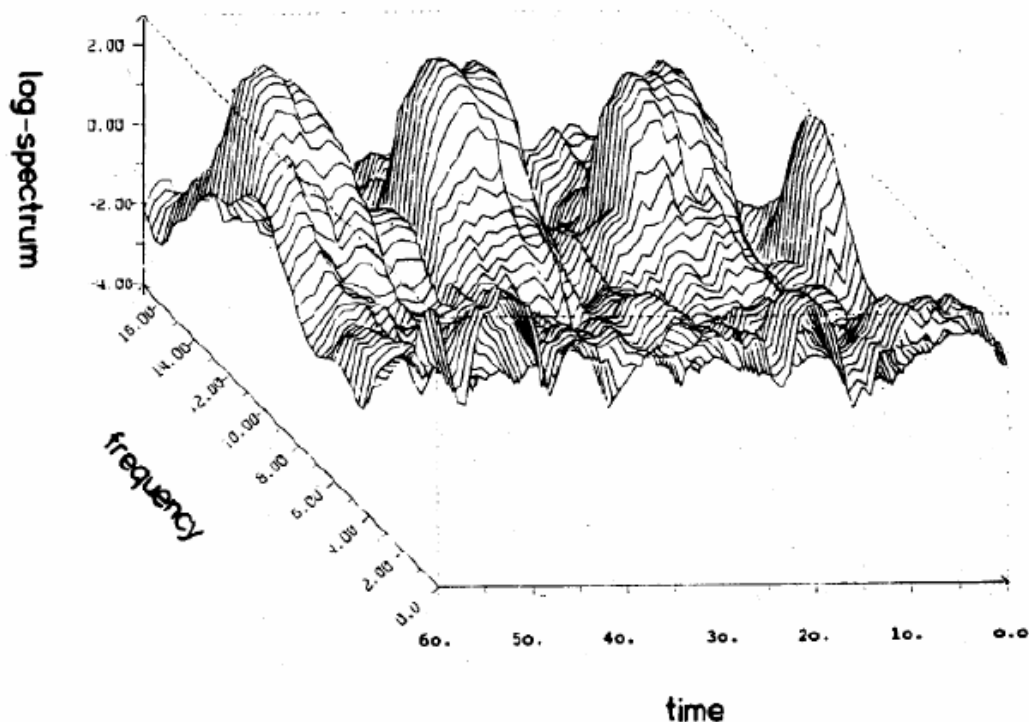


Fig. 1. Running spectra of 6×10 sec of eyes opened and eyes closed, derivation O_2 , example of a subject of the CG.

for the eyes-opened condition; for eyes closed and for the EEG at rest they are found primarily for the frontal derivations. It is important that in none of the three conditions a significant group difference could be found with respect to EOG power. The standard deviations are much higher for the EG. The individual ratio of the power with eyes closed to the power with eyes opened ('reactivity to change of conditions') averages approximately 1.5 for the EG and 2.5 for the CG and this group difference is highly significant. The ratio of the power with eyes closed to the power of the EEG at rest, decreasing from

Table 1

Time (sec) needed for α -blocking and α -restitution, determined by complex demodulation (React = reactivity)

		P_2		O_1		O_2	
		EG	CG	EG	CG	EG	CG
Open	\bar{x}	1.5	1.4	1.4	1.4	1.6	1.3
	s	0.6	0.4	0.5	0.4	0.4	0.4
Closed	\bar{x}	2.2	2.1	2.1	2.1	2.3	2.1
	s	0.9	0.7	0.8	0.9	0.9	1.0
React	\bar{x}	1.5	1.6	1.6	1.6	1.5	1.8
	s	0.5	0.6	0.8	0.8	0.7	1.1

Härdle, W., Gasser, T. and Bächer, P. (1984) EEG-responsiveness to eye opening and closing in mildly retarded children compared to a control group.

Table 2

Total power μV^2 (restricted to 1.5–25.0 Hz) for eyes open, eyes closed and the EEG at rest; mean, standard deviation and *p*-value for tests of group difference

Derivation	Eyes open			Eyes closed			At rest		
	EG	<i>p</i> -value	CG	EG	<i>p</i> -value	CG	EG	<i>p</i> -value	CG
F ₄ \bar{x}	484	0.0002	267	405	0.02	252	143	0.02	69
<i>s</i>	249		156	366		108	145		25
F ₃ \bar{x}	496	0.002	251	328	0.03	216	137	0.017	67
<i>s</i>	367		148	185		71	137		31
C ₄ \bar{x}	388	0.002	154	264	> 0.05	164	140	> 0.05	73
<i>s</i>	620		91	184		88	150		46
C ₃ \bar{x}	274	0.002	135	224	0.003	153	143	0.03	70
<i>s</i>	259		69	107		79	143		40
C _z \bar{x}	279	0.004	156	272	> 0.05	186	164	> 0.05	88
<i>s</i>	255		66	162		79	159		39
P _z \bar{x}	221	0.023	143	291	> 0.05	211	210	> 0.05	120
<i>s</i>	163		50	232		90	216		54
O ₂ \bar{x}	246	> 0.05	154	334	> 0.05	402	234	> 0.05	209
<i>s</i>	186		55	272		236	239		129
O ₁ \bar{x}	223	0.037	137	299	> 0.05	335	251	> 0.05	198
<i>s</i>	161		54	256		198	354		142

anterior to posterior, has a mean value roughly the same for both groups.

How total power (between 1.5 and 25.0 Hz) is distributed over the conventional frequency bands δ , θ , α_1 , α_2 , β_1 and β_2 is given in terms of relative power. In table 3 the average relative power for eyes opened and closed and at rest is given for all derivations; and table 4 contains the *p*-values when testing for statistical significance of differences found between the EG and the CG. Contrary to popular opinion, β -activity is not augmented when eyes are opened (the slow β -band, β_1 , shows then even a reduction compared to eyes closed or to the rest condition). The differences between groups for β_1 and β_2 are small and this holds also for the α_1 -band. There, an increase is found from eyes opened to eyes closed and even more with respect to the condition at rest. The most dramatic changes occur for the fast α -band (α_2): compared to eyes opened there is a drastic increase in the average power when closing the eyes and also at rest for both groups. The average power at parieto-occipital derivations is even larger for eyes closed compared to the rest conditions for the CG. For all three conditions and for most derivations, the relative α_2 -power is significantly larger for the CG and the *p*-values are smallest for the eyes closed and not for the rest conditions. For the slow bands δ and θ ,

Table 3

Relative broad band parameter under eyes open and closed conditions in an 'on-off experiment' and of the EEG at rest; means for EG ($n = 23$) and CG ($n = 25$)

Loca- tion	Group	δ			θ			α_1		
		Eyes open	Eyes shut	At rest	Eyes open	Eyes shut	At rest	Eyes open	Eyes shut	At rest
F ₄	EG	0.86	0.50	0.32	0.25	0.24	0.33	0.06	0.09	0.14
	CG	0.55	0.52	0.34	0.27	0.22	0.29	0.06	0.08	0.13
F ₃	EG	0.56	0.51	0.32	0.26	0.24	0.32	0.06	0.09	0.14
	CG	0.54	0.50	0.33	0.27	0.22	0.31	0.06	0.08	0.13
C ₄	EG	0.49	0.43	0.28	0.27	0.28	0.34	0.08	0.11	0.16
	CG	0.40	0.38	0.27	0.27	0.24	0.29	0.09	0.11	0.16
C ₃	EG	0.46	0.39	0.28	0.28	0.28	0.34	0.09	0.12	0.15
	CG	0.43	0.34	0.26	0.27	0.25	0.30	0.09	0.12	0.17
C ₂	EG	0.46	0.38	0.29	0.30	0.30	0.37	0.08	0.12	0.16
	CG	0.43	0.35	0.27	0.31	0.29	0.35	0.09	0.12	0.16
P ₂	EG	0.44	0.35	0.27	0.31	0.29	0.32	0.10	0.15	0.21
	CG	0.43	0.31	0.22	0.27	0.22	0.27	0.10	0.13	0.22
O ₂	EG	0.45	0.33	0.25	0.28	0.22	0.25	0.09	0.13	0.21
	CG	0.40	0.22	0.16	0.22	0.12	0.17	0.09	0.11	0.25
O ₁	EG	0.46	0.33	0.26	0.28	0.22	0.23	0.09	0.13	0.22
	CG	0.40	0.22	0.17	0.23	0.13	0.16	0.09	0.11	0.25

	α_2			β_1			β_2		
	Eyes open	Eyes shut	At rest	Eyes open	Eyes shut	At rest	Eyes open	Eyes shut	At rest
F ₄	0.04	0.07	0.09	0.04	0.05	0.07	0.05	0.05	0.05
	0.04	0.09	0.11	0.04	0.05	0.07	0.04	0.04	0.05
F ₃	0.04	0.07	0.09	0.04	0.05	0.07	0.04	0.04	0.06
	0.04	0.10	0.12	0.04	0.06	0.07	0.04	0.05	0.05
C ₄	0.06	0.09	0.10	0.05	0.06	0.07	0.04	0.05	0.05
	0.12	0.15	0.16	0.05	0.07	0.07	0.05	0.05	0.05
C ₃	0.06	0.10	0.11	0.06	0.06	0.07	0.05	0.05	0.05
	0.10	0.16	0.16	0.06	0.08	0.08	0.04	0.04	0.05
C ₂	0.06	0.10	0.09	0.05	0.06	0.06	0.04	0.04	0.04
	0.07	0.14	0.12	0.05	0.07	0.06	0.04	0.04	0.04
P ₂	0.06	0.11	0.12	0.05	0.06	0.05	0.03	0.03	0.03
	0.10	0.22	0.19	0.06	0.08	0.06	0.04	0.04	0.03
O ₂	0.08	0.20	0.20	0.06	0.08	0.06	0.04	0.04	0.03
	0.15	0.40	0.31	0.09	0.10	0.07	0.05	0.04	0.04
O ₁	0.08	0.20	0.20	0.06	0.07	0.06	0.03	0.04	0.03
	0.16	0.41	0.31	0.08	0.09	0.07	0.04	0.04	0.04

Härde, W., Gasser, T. and Bächer, P. (1984) EEG-responsiveness to eye opening and closing in mildly retarded children compared to a control group.

Table 4

Relative broad band parameters for EEG in different conditions: *p*-values for two-sample Wilcoxon tests when comparing EG and CG (tabulated when *p* < 0.05)

Location	δ			θ			α_1		
	Eyes open	Eyes shut	At rest	Eyes open	Eyes shut	At rest	Eyes open	Eyes shut	At rest
F ₄	-	-	-	-	-	-	-	-	-
F ₃	-	-	-	-	-	-	-	-	-
C ₄	0.03	-	-	-	-	-	-	-	-
C ₃	-	-	-	-	-	-	-	-	-
C _z	-	-	-	-	-	-	-	-	-
P _z	-	-	-	0.03	0.01	-	-	-	-
O ₂	-	0.02	0.006	0.001	0.0001	0.01	-	-	-
O ₁	0.005	0.002	0.008	0.01	0.0004	0.01	-	-	-

	α_2			β_1			β_2		
	Eyes open	Eyes shut	At rest	Eyes open	Eyes shut	At rest	Eyes open	Eyes shut	At rest
F ₄	-	-	0.029	-	-	-	-	-	-
F ₃	-	0.018	0.02	-	-	-	-	-	-
C ₄	0.002	0.001	0.004	-	-	-	-	-	-
C ₃	0.009	0.002	0.01	-	-	-	-	-	-
C _z	-	0.006	-	-	-	-	-	-	-
P _z	0.005	0.0001	0.03	-	-	-	-	-	-
O ₂	0.0009	0.0001	0.047	-	-	-	-	-	-
O ₁	0.0002	0.0014	0.03	-	-	-	-	-	-

significant differences between the EG and the CG arise with respect to the occipital derivations. The ratio between absolute α_2 -power for eyes closed to eyes opened ('reactivity') is larger for the CG and this difference becomes significant for 4 derivations.

Let us now turn to a description of the α -rhythm at parieto-occipital derivations in the conditions with eyes closed and at rest (table 5); children for whom an α -rhythm in both conditions could be identified are included (compare methods). The proportion of children with an α -rhythm is lower for the EG than for the CG. The average peak frequency shows the same pattern for both groups, being significantly faster for eyes closed and also for occipital derivations compared to the parietal one. Standard deviations are higher for the EG. The average power (in percentage) of the α -rhythm is about the same for both groups when at rest; after closing the eyes, roughly the same average α -peak power is reached for the CG, whereas it stays below that power for the EG.

The results for the innovation variance, i.e. the power of the driving process

Table 5
Peak frequency and peak power (%) of α -rhythm in conditions of rest and after closing the eyes (mean and standard deviation of EG and CG)

		P_2		O_1		O_2	
		EG <i>n</i> = 12	CG <i>n</i> = 21	EG <i>n</i> = 15	CG <i>n</i> = 19	EG <i>n</i> = 13	CG <i>n</i> = 18
Peak frequency							
\bar{x}	Closed	9.41	9.85	10.70	10.80	10.68	10.56
	Rest	9.11	9.27	9.42	9.63	9.37	9.53
<i>s</i>	Closed	1.19	0.95	1.00	0.71	1.20	0.95
	Rest	1.21	1.15	1.39	0.84	1.02	0.75
Peak power							
\bar{x}	Closed	42	51	55	62	51	65
	Rest	56	55	65	64	63	69
<i>s</i>	Closed	16	18	18	13	13	18
	Rest	22	12	13	11	13	14

Table 6
Innovation variance (logarithmic scale) when fitting an autoregression separately to phases of eyes closed and eyes opened; mean standard deviation and *p*-value (if < 0.05) of test between groups (*n* = 23; *n* = 25)

		P_2		O_1		O_2	
		EG	CG	EG	CG	EG	CG
Open	\bar{x}	4.0	3.7	4.0	3.8	4.2	4.0
	<i>s</i>	0.4	0.3	0.5	0.4	0.5	0.4
	<i>p</i>	0.001		0.035		-	
Closed	\bar{x}	4.1	4.0	4.2	4.3	4.4	4.5
	<i>s</i>	0.5	0.3	0.5	0.5	0.5	0.5
	<i>p</i>	-		-		-	
Reactivity	\bar{x}	1.02	1.07	1.0	1.0	1.0	1.0
	<i>s</i>	0.07	0.04	0.08	0.08	0.08	0.07
	<i>p</i>	0.001		0.006		0.0002	

Härdle, W., Gasser, T. and Bächer, P. (1984) EEG-responsiveness to eye opening and closing in mildly retarded children compared to a control group.

in the autoregressive model, are given in table 6. Higher averages are reached for the EG in the eyes opened but not in the eyes closed condition, and the resulting reactivities are significantly different with low error probabilities. Let us note that consistently for both groups the innovation variance is larger for eyes closed. When fitting an autoregression, the average order (a measure of complexity of the spectrum) needed was higher for the CG, and this difference was statistically significant for eyes closed (no table given).

There was a slight trend for the EG to habituate more over the three phases of eyes closing and opening, but this trend did not reach statistical significance.

4. Discussion

Opening and closing the eyes is a simple activation procedure when recording an EEG which has the advantage that it does not need too much cooperation or comprehension for mildly retarded children. To achieve our goal of investigating changes in the ongoing EEG, discrete visual stimuli might also be used; their primary purpose is, however, the determination of the brain potential elicited.

In order to counter-balance a tendency to look for differences only between the EG and the CG, it should be stressed how similar the pattern of change is for the two groups from eyes opened to eyes closed (transient state) to the rest condition (steady state); tables 3 and 5 are rather illuminating in this respect. Regarding the latency for attenuating and restituting the α -rhythm there was qualitative similarity in so far as both groups took a longer time to reconstitute the α than to attenuate it. With respect to latency we found, moreover, quantitative agreement for the EG and the CG in both conditions. Allowing for differences in technique, this is in accordance with Yasui (1975) and Psatta (1981) who reported no latency differences for the visual evoked potentials of mildly retarded children. The latencies found in our study are in good agreement with those reported earlier (Nogawa et al., 1976; Grünwald-Zuberbier, Grünwald and Rasche, 1975; Kemp and Blom, 1981). The findings of Berkson et al. (1961), confirmed by Baumeister et al. (1963), of shorter α -blocking duration times for retarded subjects, are in some contradiction to ours. One has, however, to keep in mind that their visual determination of α -return (at least 5 continuous α -waves) differ from our computer-aided method and also that they used flashes of light for activation. Furthermore, they reported a considerable amount of variability and for the last 6 out of 10 trials the average latencies were similar for the two groups.

Measures of amplitude of the EEG may be substantially influenced by eye movements and blinks. It is, therefore, reassuring that neither for eyes closed nor for eyes opened significant differences in EOG-power could be found between the EG and the CG (and the same was true for the EEG at rest, see

Gasser et al., 1983a). Total power (restricted to 1.5–25.0 Hz to exclude some common types of extracerebral potentials) showed insignificant group differences for the eyes closed and the rest condition, except for frontal derivations, but differences became significant for eyes opened at all derivations except for O₂. The lack of significance in the first two conditions is due to a confounding of higher power in the slow bands for the EG and of stronger α -activity for the CG at central, parietal and occipital derivations (table 3). This implies that a simple quantity like total power is more meaningful for the eyes opened condition. The higher average total power for the EG can be interpreted as a developmental lag (as in Gasser et al., 1983 a, b). The significantly smaller changes in EEG-amplitude for the EG, when passing through phases of eyes opened and closed, are tentatively interpreted as a lowered performance of the arousal system of the EG (changes were quantified as ratios of total power in the two conditions and also of power in the α_2 -band, and differences were significant at parieto-occipital derivations). Synchronization and desynchronization of EEG-activity as well as the sleep-wakefulness-cycle, are mediated by the reticulo-thalamo-cortical ascending system (Andersen and Andersson, 1968; Moruzzi and Magoun, 1949). The changes from an inattentive state, accompanied by a synchronized EEG, to an aroused state with a desynchronized EEG, can be provoked by sensory stimuli as simple as eyes opening and closing, and these changes are regulated by the same system. More recent results have confirmed and differentiated the validity of such an interpretation (see Yingling and Skinner, 1977, and the literature cited there). Relative power in the 6 frequency bands introduced and under eyes opened and closed and at rest condition yields again a pattern which is rather similar for the two groups. It is of interest to see a higher percentage of α_2 -power after closing the eyes ('transient state') compared to the rest condition ('steady state') for the CG but not for the EG. The widely held belief that β -activity is enhanced by opening the eyes is not confirmed by our data analysis for either group; this might be explained by the fact that in a visual inspection β -activity becomes better visible due to the suppression of the α -rhythm.

Autoregressive model building in the way proposed by Zetterberg (1969), and further modified by Steinberg et al. (1983), allows a quantification of rhythmic and of diffuse EEG activity. A lower number of children of the EG as compared to the CG had an α -rhythm in both the eyes closed and the rest conditions. There is an insignificant tendency for a slower α -rhythm for the EG in both conditions; it is remarkable that the α is faster – for both groups – in the transient state compared to the steady state condition. The fact that the relative power of the α -rhythm is roughly the same for both groups at rest, but is lower for the EG in the eyes closed condition, is in line with the arousal hypothesis formulated above. The same is true for the highly significant finding of lower reactivities for the EG in the innovation variance, i.e. in the driving power of the EEG as modeled by an autoregression.

For the neurophysiological quantities considered, we often found a higher standard deviation for the EG; this is interpreted as indicating heterogeneity of the group of mildly retarded children in neurophysiological terms, confirming earlier results obtained for the EEG at rest (Gasser et al., 1983a).

Appendix 1

Autoregressive scheme in the on-off experiment

Let $Y_1(t)$, $t = 1, \dots, T_1$ be the EEG under opened eyes and $Y_2(t)$, $t > T_1$ the EEG when the eyes are closed. We assume that the following model holds:

$$\sum_{i=1}^p \alpha_i^{(j)} Y_j(t-i) = \eta_j(t), \quad j = 1, 2$$

indicating the block, where $\alpha_i^{(j)}$ are the autoregressive coefficients, $\alpha_0^{(j)} = 1$, without restriction on generality, and $\eta_j(t)$ denotes white noise with distribution function $G(x/\sigma_j)$, p_j the order of the AR-scheme.

We also require $\int x dG(x) = 0$, $\int x^2 dG(x) = 1$, $\int x^4 dG(x) = \gamma_2 + 3$.

The autoregressive coefficients α_i are estimated by the Yule-Walker equations (see Kopmans, 1974) and we obtain from those estimates a residual process $\hat{\eta}_j(t)$ from which the innovation variance σ_j^2 is estimated.

The order p_j is chosen according to one of the following criteria, namely as the minimum of the functions

$$AIC(p) = T \log(\hat{\sigma}_j) + 2p \quad (\text{Akaike, 1969}),$$

$$HAN(p) = T \log(\hat{\sigma}_j) + (\theta \log \log T) p, \theta > 2 \quad (\text{Hannan and Quinn, 1979}),$$

$$SHW(p) = T \log(\hat{\sigma}_j) + p \log T \quad (\text{Schwarz, 1978}).$$

$\hat{\sigma}_j^2$ here denotes the following estimate of the innovation variance

$$\hat{\sigma}_j^2 = \sum_{t=1}^{T_j} \hat{\eta}_j^2(t) / (T_j - 1).$$

Appendix 2

Jack-knife estimate of the innovation variance

The Jack-knife estimate of the innovation variance is based on the pseudovalues

$$\hat{\theta}_{i(j)} = T_j \log(\hat{\sigma}_j^2) - (T_j - 1) \log(v_{i(j)}), \quad i = 1, \dots, T_j, \quad j = 1, 2,$$

where;

$$v_{i(j)} = \sum_{t \neq i} \hat{\eta}_j^2(t) / (T_j - 1).$$

The Jack-knife estimate is the average of the pseudovalues:

$$\hat{\theta}_{(j)} = \sum_{i=1}^{T_j} \hat{\theta}_{i(j)} / T_j.$$

This Jack-knife estimate was considered by Davis (1978, 1979), who also established a device to test the differences of the innovation variances in the on/off phase.

References

- Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21, 243-247.
- Andersen, P. and Andersson, S.A. (1968). *Physiological Basis of the Alpha Rhythm*. Appleton-Century-Crofts: New York.
- Aranibar, A. and Pfurtscheller, G. (1978). On and off effects in the background EEG activity during one-second photic stimulation. *EEG and Clinical Neurophysiology*, 44, 307-316.
- Baumeister, A.A. and Hawkins, W.F. (1967). Alpha responsiveness to photic stimulation in mental defectives. *American Journal of Mental Deficiency*, 71, 783-786.
- Baumeister, A.A., Spain, C.J. and Ellis, N.R. (1963). A note on alpha-block duration of normals and retardates. *American Journal of Mental Deficiency*, 67, 723-725.
- Berkson, G., Hermelin, G. and O'Connor, N. (1961). Physiological responses of normals and institutionalized mental defectives to repeated stimuli. *Journal of Mental Deficiency Research*, 5, 30-39.
- Davis, W.W. (1977). Robust interval estimation of the innovation variance of an ARMA model. *Annals of Statistics*, 5, 700-708.
- Davis, W.W. (1979). Robust methods for detection of shifts of innovation variance of a time series. *Technometrics*, 21, 313-320.
- Fuller, P.W. (1977). Computer estimated alpha attenuation during problem solving in children with learning disabilities. *EEG and Clinical Neurophysiology*, 42, 149-156.
- Gasser, T., Möcks, J., Lenard, H.G. and Verleger, R. (1983a). The EEG of mildly retarded children: Developmental, classificatory and topographic aspects. *EEG and Clinical Neurophysiology*, 55, 131-144.
- Gasser, T., von Lucadou-Müller, I., Verleger, R. and Bäcker, P. (1983b). Correlating EEG and IQ: A new look at an old problem using computerized EEG parameters. *EEG and Clinical Neurophysiology*, 55, 493-504.
- Grünwald-Zuberbier, E., Grünwald, G. and Rasche, A. (1975). Hyperactive behaviour and EEG arousal reactions in children. *EEG and Clinical Neurophysiology*, 38, 149-159.
- Hannan, E.J. and Quinn, B.G. (1979). Estimating the dimension of a model. *Journal of the Royal Statistical Society, Series B*, 41, 190-195.
- Kawabata, N. (1972). Nonstationary power spectrum analysis of the photic alpha blocking. *Kybernetik*, 12, 40-44.
- Kemp, B. and Blom, H.A.P. (1981). Optimal detection of the alpha state in a model of the human electroencephalogram. *EEG and Clinical Neurophysiology*, 52, 222-225.
- Koopmans, L.H. (1974). *The Spectral Analysis of Time Series*. Academic Press: New York.

Härdle, W., Gasser, T. and Bäcker, P. (1984) EEG-responsiveness to eye opening and closing in mildly retarded children compared to a control group.

- Liepmann, M.C. (1979). *Geistig Behinderte Kinder and Jugendliche*. Hans Huber Verlag: Bern.
- Matoušek, M. and Petersén, I. (1973). Frequency analysis of the EEG in normal children and adolescents. In: Kellaway, P. and Petersén, I. (Eds.). *Automation of Clinical Electroencephalography*. Raven Press: New York, 75–102.
- Moruzzi, G. and Magoun, H.W. (1949). Brain stem reticular formation and activation of the EEG. *EEG and Clinical Neurophysiology*, 1, 455–473.
- Nogawa, T., Katayama, K., Tabata, Y., Oshio, T. and Kawahara, T. (1976). Changes in the amplitude of the EEG induced by a photic stimulus. *EEG and Clinical Neurophysiology*, 40, 78–88.
- Psatta, D.M. (1981). Visual evoked potential habituation in mental deficiency. *Biological Psychiatry*, 16, 729–740.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Mathematical Statistics*, 6, 461–464.
- Steinberg, H.-W., Gasser, T. and Franke, J. (1983). Fitting autoregressive processes to EEG time series: Comparison of estimates of the order. Manuscript.
- Treiman, G.J. (1975). Problems of concept and measurement in the comparative study of occupational mobility. *Social Science Research*, 4, 183.
- Walter, D.O. (1969). The method of complex demodulation. *EEG and Clinical Neurophysiology, Suppl.*, 27, 53–57.
- W.H.O. (1977). *Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death, 9th Revision*. World Health Organization: Geneva.
- Wolfensberger, W. and O'Connor, N. (1965). Stimulus intensity and duration effects on EEG and GSR responses of normals and retardates. *American Journal of Mental Deficiency*, 70, 21–37.
- Yasui, M. (1975). Visual evoked responses of mentally retarded children. *Wakayama Medical Report*, 17, 57–64.
- Yingling, C.D. and Skinner, J.E. (1977). Gating of thalamic input to cerebral cortex by nucleus reticularis thalami. In: Desmedt, J.E. (Ed.). *Attention, Voluntary Contraction and Eventuated Cerebral Potentials*. Karger: Basel, 70–96.
- Zetterberg, L.H. (1969). Estimation of parameters for a linear difference equation with application to EEG analysis. *Mathematical Biosciences*, 6, 227–275.

Härdle, W., Gasser, T. and Bächer, P. (1984) EEG-responsiveness to eye opening and closing in mildly retarded children compared to a control group.

Contribution to the Discussion of the
Paper by Silverman, October 1984

Wolfgang Härdle
Johann Wolfgang Goethe - Universität
D-6000 Frankfurt am Main

Professor Silverman's article on the spline smoothing approach to curve fitting is an excellent contribution to the understanding of data smoothing. He points out the various attractive features and shows in a variety of examples the wide applicability of spline smoothing. I found the clear and elegant discussion of Section 3, pointing out the relationships among spline smoothing and kernel regression, very stimulating.

My comments will address (a) the generalized cross-validation method (Section 4) and (b) the proposal of an automatic choice of the smoothing parameter in the case of robust spline smoothing (Section 8.1).

The generalized cross-validation method can be considered as a member of the smoothing parameter selection procedures :

"Choose α to minimize the score

$$S_n(\Xi; \alpha) = \text{RSS}(\alpha) \Xi(n^{-1} \text{tr } A(\alpha)). "$$

Here Ξ denotes a "selection penalty" with expansion $\Xi(u) = 1 + 2u + \Xi''(\xi)u^2$. The generalized cross-validation score GXVSC(α) has penalty $\Xi(u) = (1-u)^{-2}$. A FPE-Type penalty $\Xi(u) = (1+u)/(1-u)$ (Akaike, 1970) or Shibata's (1981) $\Xi(u) = (1+2u)$ are also possible.

Note that

$$\begin{aligned} E S_n(\Xi; \alpha) &= E \left\{ n^{-1} \sum_{i=1}^n \varepsilon_i^2 + n^{-1} \sum_{i=1}^n (\hat{g}(t_i) - g(t_i))^2 + 2n^{-1} \sum_{i=1}^n \varepsilon_i (\hat{g}(t_i) - g(t_i)) \right\} \\ &\quad \times [1 + 2n^{-1} \text{tr } A(\alpha) + O((n^{-1} \text{tr } A(\alpha))^2)] \\ &= \sigma^2 + E n^{-1} \sum_{i=1}^n (\hat{g}(t_i) - g(t_i))^2 - 2n^{-1} \text{tr } A(\alpha) \sigma^2 \\ &\quad + 2n^{-1} \text{tr } A(\alpha) \sigma^2 + O((n^{-1} \text{tr } A(\alpha))^2). \end{aligned}$$

So asymptotically minimizing $S_n(\Xi; \alpha)$ is the same as to minimize $n^{-1} \sum_{i=1}^n (\hat{g}(t_i) - g(t_i))^2$.

This expansion also suggests that all possible selectors $S_n(\Xi; \alpha)$ are asymptotically equivalent. However, a Ξ with a large second derivative in a neighborhood of zero could be preferred in order to penalize more for undersmoothing.

Which of the possible penalties Ξ should be applied in practice ?

Denote the robust spline by \hat{g}_R . An automatic choice of the smoothing parameter by means of (8.3) is a natural extension of the cross-validation score $XVSC(\alpha)$.

Let $\rho(s;t) = \rho(s-t)$, $\psi = \rho'$ and let $V_n(\psi)$ be a consistent estimate of $E\psi^2/E\psi'$. I propose the following score

$$W_n(\alpha) = 2n^{-1} \sum_{i=1}^n \rho(Y_i - \hat{g}_R(t_i)) + 2n^{-1} \text{tr } A(\alpha) V_n(\psi)$$

as a smoothing parameter selector. The idea behind $W_n(\alpha)$ is that by Taylor expansion

$$\begin{aligned} E W_n(\alpha) &= 2E\rho(\varepsilon) + 2\{n^{-1} \sum_{i=1}^n E[\psi(\varepsilon_i)(g(t_i) - \tilde{g}(t_i))]\} \\ &\quad + n^{-1} \text{tr } A(\alpha) E V_n(\psi) + n^{-1} \sum_{i=1}^n E[\psi'(\varepsilon_i)(g(t_i) - \tilde{g}(t_i))^2], \end{aligned}$$

where $\tilde{g}(s)$ is the linear approximation to \hat{g}_R , as given by *Cox* (1983). The first term on the right hand side is independent of α , the second vanishes and the third term is the quantity of interest.

How is $W_n(\alpha)$ related to $XVSC$?

REFERENCES :

- AKAIKE, H. (1970) : Statistical predictor identification.
Ann. Inst. Statist. Math., 22, 203-217 .
- COX, D. (1983) : Asymptotics for M-type smoothing splines.
Ann. Statist. 11, 530-551 .
- SHIBATA, R. (1981) : An optimal selection of regression variables.
Biometrika, 68, 45-54 .

BANDWIDTH CHOICE IN NONPARAMETRIC REGRESSION FUNCTION ESTIMATION

W.Härdle

J.S.Marron

Received: Revised version: March 8, 1985

Abstract. It is shown that a crossvalidatory choice of the bandwidth in nonparametric kernel regression yields an estimator of the regression function that solves an open problem of C.J.Stone.

Let (X, Y) be a pair of random variables which are respectively d and 1 dimensional and let $m(\cdot)$ denote the regression curve of the response Y on X , i.e. $m(x) = E(Y|X=x)$. Suppose that a random sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of size n has been observed and it is desired to estimate the function $m(x)$ by nonparametric estimators $T_n(x)$ based on that random sample. Stone [3] showed that, in a L_2 -sense, the optimal rate of convergence of $T_n(x)$ to $m(x)$ depends, roughly speaking, on the amount of smoothness subscribed to m . More precisely, let \mathcal{H} denote the collection of functions on \mathbb{R}^d with Hölder-continuous bounded k -th derivative with exponent β , i.e. $\mathcal{H} = \{g \in C^k, |g^{(k)}(x) - g^{(k)}(x')| \leq H |x - x'|^\beta\}$ and let, with $p = k + \beta$, $r = p / (2p + d)$. C.Stone showed that n^{-r} is the optimal rate of convergence and constructed an estimator that achieved this optimal rate (see his definition of optimality and achievability). The construction of that estimator required the knowledge of k and β . Therefore, in his question 3, he asks if there exists a single estimator which achieves the optimal rate independently of r .

AMS 1980 Subject Classification: Primary 62G05, Secondary 62G20
keywords and phrases: nonparametric regression, cross-validation

In this note we show that kernel estimators with a bandwidth selected by a crossvalidatory technique are providing an answer to Stone's question.

The kernel estimators $m_n(x)$ are defined with a kernel function $K: \mathbb{R}^d \rightarrow \mathbb{R}$ and a bandwidth $h=h(n) > 0$,

$$m_n(x) = n^{-1} h^{-d} \sum_{i=1}^n K((x-X_i)/h) Y_i / f_n(x),$$

where $f_n(x)$ is the familiar Rosenblatt-Parzen density estimator of $f(x)$, the marginal density of X . The well-known cross-validation technique is based on the leave-one-out estimators

$$m_n^j(x) = (n-1)^{-1} h^{-d} \sum_{i \neq j} K((x-X_i)/h) Y_i / f_n(x)$$

that are used to define the following estimate of the prediction error

$$p_n(h) = n^{-1} \sum_{j=1}^n (Y_j - m_n^j(X_j))^2 w(X_j),$$

where $w: \mathbb{R}^d \rightarrow \mathbb{R}$ is a weight function. The so-called cross-validatory choice of the bandwidth is that \hat{h} that minimizes $p_n(h)$ over a certain interval $[\underline{h}, \bar{h}]$ to be defined below. We will make use of the following measures of accuracy for $m_n(x)$,

$$d_I(h) = \int (m_n(x) - m(x))^2 f(x) w(x) dx$$

$$d_M(h) = E(d_I(h) \mathbb{I}(f_n(x) > \gamma/2)),$$

where γ is a lower bound on $f(x)$ on the support of the weight function w .

Theorem. Assume that

- (i) $\underline{h} = n^{-d^{-1}+\delta}$, $\bar{h} = n^{-\delta}$, $\delta > 0$;
- (ii) K is compactly supported with $\int K(u) du = 1$ and is Hölder-continuous;
- (iii) f is Hölder-continuous and $f(x) > \gamma$ on $\text{supp}\{w\}$;
- (iv) m is Hölder-continuous;
- (v) $S(x) = E(Y^2 | X=x)$ is Hölder-continuous;

$$(vi) \sup_{x \in \text{supp}\{w\}} E(|Y|^l | X=x) \leq M_l < \infty, l=1,2,\dots$$

Let $r_n = n^{-1} \sum_{j=1}^n (Y_j - m(X_j))^2 w(X_j)$ be an estimate of the integrated residual variance and let P_m denote the dependence of P on m as in Stone [3]. Then for all $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \sup_{m \in \Theta_r} P_m \left\{ \sup_{h \in [\underline{h}, \bar{h}]} \left| \frac{p_n(h) - r_n}{d_M(h)} - 1 \right| > \epsilon \right\} = 0,$$

where Θ_r denotes the dependence of Θ on r as described above.

Remark. This result shows that the function $p_n(h)$ approximates $d_M(h)$ (up to a constant) uniformly over $[\underline{h}, \bar{h}]$. Since this approximation is of lower order than $\inf_{\underline{h}} d_M(h)$, which itself tends to zero, the cross-validatory choice \hat{h} asymptotically minimizes $d_M(h)$ over $[\underline{h}, \bar{h}]$.

We first indicate why this result yields a solution to Stone's question. Define \hat{h} as the cross-validatory choice of the bandwidth and note that \hat{h} is defined independently of r (or the smoothness of m). From the theorem we have

$$(I) \lim_{n \rightarrow \infty} \sup_{\Theta_r} P_m \left\{ \left| \frac{d_M(\hat{h})}{\inf_{\underline{h}} d_M(h)} - 1 \right| > \epsilon \right\} = 0.$$

In Marron and Härdle [2] it is seen that d_I uniformly approximates d_M , more precisely

$$(II) \lim_{n \rightarrow \infty} \sup_{\Theta_r} P_m \left\{ \sup_{\underline{h}} \left| \frac{d_I(h)}{d_M(h)} - 1 \right| > \epsilon \right\} = 0.$$

By straightforward computations using the fact that $m \in \Theta_r$ for some r we obtain that there exists a $C > 0$ such that

$$(III) \lim_{n \rightarrow \infty} \sup_{\Theta_r} \inf_{\underline{h}} d_M(h)/n^{-2r} \leq C.$$

Putting (I), (II), (III) together gives, if $m \in \Theta_r$ for some r , with a constant c

$$\lim_{n \rightarrow \infty} P_m \left\{ d_I(\hat{h}) \geq c n^{-2r} \right\} = 0,$$

which answers Stone's question since \hat{h} is defined independently of r .

Proof. We only give an idea of the proof, the detailed analysis can be found in Härdle and Marron [1]. The key step is to decompose the prediction error in the following way

$$p_n(h) = r_n + n^{-1} \sum_{j=1}^n (m(X_j) - m_n^j(X_j))^2 w(X_j) + 2n^{-1} \sum_{j=1}^n (Y_j - m(X_j)) (m(X_j) - m_n^j(X_j)) w(X_j).$$

It suffices to show

$$(IV) \quad \lim_{n \rightarrow \infty} \sup_{\mathbb{H}_n} P_m \left\{ \sup_{\mathbb{H}^p} \left| \frac{n^{-1} \sum_{j=1}^n (m(X_j) - m_n^j(X_j))^2 w(X_j)}{d_M(h)} - 1 \right| > \varepsilon \right\} = 0$$

$$(V) \quad \lim_{n \rightarrow \infty} \sup_{\mathbb{H}_n} P_m \left\{ \sup_{\mathbb{H}^p} \left| \frac{n^{-1} \sum_{j=1}^n (Y_j - m(X_j)) (m(X_j) - m_n^j(X_j)) w(X_j)}{d_M(h)} \right| > \varepsilon \right\} = 0.$$

These two statements are shown by splitting up the range of h 's into small balls centered at gridpoints in $[\underline{h}, \bar{h}]$. The gaps between the gridpoints are bridged by Hölder-continuity of K, f, S and m . The probability at the gridpoints is estimated by Bonferroni's inequality and the Marcinkievitch-Zygmund inequality using the moment conditions (vi).

References.

[1] Härdle, W. and Marron, J.S. Optimal bandwidth selection in nonparametric regression function estimation. Inst. of Stat. Mimeo Series #1546, Chapel Hill, North Carolina

- [2] Marron, J.S. and Härdle, W. Random approximations to an error criterion of nonparametric statistics. Inst. of Stat. Mimeo Series #1566, Chapel Hill, North Carolina
- [3] Stone, C.J. Optimal global rates of convergence for nonparametric regression. Ann. Statist. Vol. 10 (1982), 1040 - 1053

W. Härdle
FB Mathematik
Johann-Wolfgang-Goethe Universität
D-6000 Frankfurt

J.S. Marron
Dept. Statistics
University of North Carolina
Chapel Hill, NC 27514, USA

On Robust Kernel Estimation of Derivatives of Regression Functions

WOLFGANG HÄRDLE

Frankfurt

THEO GASSER

Mannheim

ABSTRACT. When estimating derivatives of regression functions from noisy data, a number of additional problems arise compared with the estimation of the regression function itself. Linear methods, such as kernel regression or smoothing splines, will be quite sensitive to outlying observations; this holds in particular for the estimation of derivatives where differences of consecutive data points are involved. In this paper, a robust kernel estimate for derivatives of regression functions is introduced and some of its asymptotic properties are investigated.

Key words: non-parametric regression, estimation of derivatives, robust smoothing, kernel estimators, non-linear smoothers, kernel estimation

1. Introduction

Let $Y_i^{(n)} = m(t_i^{(n)}) + Z_i^{(n)}$, $i = 1, 2, \dots, n$, be a sequence of independent observations with regression functions $m(t)$, $0 < t < 1$, recorded at $0 < t_1 < \dots < t_n < 1$ and with errors $\{Z_i^{(n)}\}_{i=1}^n$ being identically distributed with mean zero. The practical importance of obtaining a non-parametric estimate of $m(t)$ has led to several estimators for $m(t)$, among them the so-called kernel estimators (Priestley & Chao, 1972; Gasser & Müller, 1979). The presence of a small portion of outliers may, however, render difficult an interpretation of the estimated regression function. Robust alternatives to the kernel method, the latter operating linearly on the data, have been proposed by Härdle & Gasser (1984) and in a random design model by Härdle (1984a) and Tsybakov (1983). Robust spline smoothing was considered by Huber (1979) and by Cox (1983).

The estimation of derivatives from noisy data is of importance in many areas of engineering and physics, and also in biomedicine (compare, e.g., Largo *et al.* (1978) for applications to longitudinal growth using smoothing splines and Bahill & Stark (1979) for applications to saccadic eye movements using heuristic methods in engineering tradition). When using an estimator which acts as a linear operation on the data, such as the kernel estimators studied by Gasser & Müller (1984) and Gasser *et al.* (1985), single outliers might mimic peaks and troughs, corresponding to unexpected zeros in the estimated derivative of the regression function. The occurrence of outliers would, therefore, lead in these cases to qualitatively wrong conclusions.

The object of this paper is to introduce a robust kernel estimator of $m'(t)$, the derivative of $m(t)$. Robust estimators of higher derivatives will also be defined, although not discussed in full detail, since their statistical analysis appears to be straightforward, given the analysis of estimators for $m'(t)$. The proposed method is derived from M estimation (Huber, 1981, Ch. 3.2) and it will be seen that the robust kernel estimate of the first derivative is an ordinary (linear) kernel estimator operating on suitable transformed residuals.

As for the ordinary kernel estimate, the smoothing parameter of the kernel weights has to be selected in an application to real data. In this paper, the choice of the smoothing parameter

is not discussed, this topic will be investigated in a forthcoming paper. For the estimation of $m(t)$, a cross-validatory device has been proposed by Härdle (1984b).

In theorem 1 consistency of the robust kernel estimator is shown, but rates of convergence are not considered. A linearization argument as in theorem 1 reveals that the proposed estimator achieves the optimal rate in the sense of Stone (1980), when requiring additional smoothness assumptions on m .

2. Notation and formulation of the estimator

Let us assume the following model for the observations

$$\{Y_i^{(n)}\}_{i=1}^n \\ Y_i^{(n)} = m(t_i^{(n)}) + Z_i^{(n)} \quad (2.1)$$

where $Z_i^{(n)}$, $1 \leq i \leq n$, are i.i.d. with $E Z_i^{(n)} = 0$. Let now $K_0: \mathbb{R} \rightarrow \mathbb{R}$ be a continuous kernel function with compact support $[-A, A]$ and $\int K_0(u) du = 1$.

The kernel regression estimate

$$m_{n,0}^*(t) = \sum_{i=1}^n \alpha_{i,0}^{(n)}(t) Y_i^{(n)}, \quad 0 < t < 1 \quad (2.2)$$

is defined through the weights

$$\alpha_{i,0}^{(n)}(t) = h^{-1} \int_{s_{i-1}^{(n)}}^{s_i^{(n)}} K_0\{(t-u)/h\} du \quad (2.3)$$

where

$$\{s_i^{(n)}\}_{i=1}^n, \quad t_i^{(n)} \leq s_i^{(n)} \leq t_{i+1}^{(n)}, \quad i = 1, \dots, n-1.$$

$0 < s_0^{(n)} \leq t_1^{(n)}$, $t_n^{(n)} \leq s_n^{(n)} < 1$ and $h = h(n)$ is the so-called *bandwidth*.

The following assumption on the asymptotic spacing of the design variables $\{t_i^{(n)}\}_{i=1}^n$ will be convenient:

$$\sup_{1 \leq i \leq n} |s_i^{(n)} - s_{i-1}^{(n)}| \leq C_1 n^{-1}$$

with a generic constant $C_1 > 0$. Gasser & Müller (1984) introduced the following estimate of $m'(t)$:

$$m_{n,1}^*(t) = \sum_{i=1}^n \alpha_{i,1}^{(n)}(t) Y_i^{(n)}, \quad 0 < t < 1 \quad (2.4)$$

where the weights $\{\alpha_{i,1}^{(n)}(t)\}_{i=1}^n$ are to be computed from a kernel K_1 satisfying moment conditions up to order k ($k \geq 3$ and odd).

$$\alpha_{i,1}^{(n)}(t) = h^{-2} \int_{s_{i-1}^{(n)}}^{s_i^{(n)}} K_1\{(t-u)/h\} du \quad (2.5a)$$

$$\int_{-A}^A K_1(u) u^j du = 0, \quad j = 0, 2, \dots \\ = -1, \quad j = 1 \\ = \beta \neq 0, \quad j = k. \quad (2.5b)$$

Note that both $m_{n,0}^*$ and $m_{n,1}^*$ are local averages of the observations and have, therefore, the well-known sensitivity to outliers. A robust estimate $m_{n,0}(t)$ for $m(t)$, based on concepts of M estimation, was considered by Härdle & Gasser (1984). This estimate is defined through a non-linear function ψ , satisfying the following property.

Let $\psi: \mathbb{R} \rightarrow \mathbb{R}$ be a bounded, antisymmetric function with bounded (Z) second derivative, such that for each fixed $0 < t < 1$, the function $\theta \rightarrow H(t, \theta) = E_t \{Y(t) - \theta\}$ has a unique zero at $\theta = m(t)$. Here $Y(t)$ has c.d.f. $F\{y - m(t)\}$.

Under mild assumptions on K_n , m and h it was shown in Härdle & Gasser (1984) that the zero $m_{n,0}(t)$ of the function

$$\theta \rightarrow H_n(t, \theta) = \sum_{i=1}^n \alpha_{i,0}^{(n)}(t) \psi(Y_i^{(n)} - \theta), \quad 0 < t < 1$$

converges in probability to $m(t)$ as $n \rightarrow \infty$.

The proposed robust estimate of $m'(t)$ is defined as follows:

$$m_{n,1}(t) = \sum_{i=1}^n \alpha_{i,1}^{(n)}(t) \psi'(Y_i^{(n)} - m_{n,0}(t)) / D_n(t) \quad (2.6)$$

where

$$D_n(t) = \sum_{i=1}^n \alpha_{i,0}^{(n)}(t) \psi'(Y_i^{(n)} - m_{n,0}(t)). \quad (2.7)$$

Expand

$$\psi'(Y_i^{(n)} - m_{n,0}(t)) = \psi'(Z_i^{(n)}) + \{m(t) - m_{n,0}(t)\} \psi''(\xi_i^{(n)}) + \{m(t) - m_{n,0}(t)\} \psi''(\xi_i^{(n)}).$$

Summing now with the weights $\{\alpha_{i,0}^{(n)}(t)\}_{i=1}^n$ shows that for each fixed $t \in (0, 1)$:

$$(i) \sum_{i=1}^n \alpha_{i,0}^{(n)}(t) \psi'(Z_i^{(n)}) \xrightarrow{P} E_t \psi'(Z)$$

by the WLLN:

$$(ii) \sum_{i=1}^n \alpha_{i,0}^{(n)}(t) \{m(t) - m_{n,0}(t)\} \psi''(\xi_i^{(n)}) \xrightarrow{P} 0,$$

provided $m(t)$ is Hölder continuous, which follows from (C) below:

$$(iii) \{m(t) - m_{n,0}(t)\} \sum_{i=1}^n \alpha_{i,0}^{(n)}(t) \psi''(\xi_i^{(n)}) \xrightarrow{P} 0,$$

by consistency of $m_{n,0}(t)$ and the boundedness of ψ'' . This altogether yields that $D_n(t) \xrightarrow{P} q = E_t \psi'(Z)$.

Ignoring the effect of randomness of $D_n(t)$, the estimator $m_{n,1}(t)$ can therefore be interpreted as an ordinary (linear) kernel estimate of the first derivative of a regression function applied to the non-linearly transformed residual $\psi(Y_i^{(n)} - m_{n,0}(t))/q$. A heuristic justification of $m_{n,1}(t)$ is delayed to section 4 where we also consider the estimation of higher derivatives. It is well known from the theory of robust estimation (Huber, 1981) that the boundedness of ψ guarantees the bounded influence of $m_{n,0}(t)$ if an observation is moved to infinity. It is therefore interesting to note that setting $\psi(u) = u$ leads to the estimator (2.4). For notational convenience we will omit the superscript n from now on.

3. Results

The following assumptions are needed for our results

$$nh^3 \rightarrow \infty, \text{ as } h=h(n) \rightarrow 0 \text{ and } n \rightarrow \infty, \quad (B)$$

$$m'(t) \text{ is continuous for } 0 < t < 1. \quad (C)$$

We prove consistency and asymptotic normality of the robust estimator $m_{n,1}(t)$, as defined in (2.6). In our result about asymptotic normality (theorem 2) we centre $m_{n,1}(t)$ around $m(t)$ in order to avoid additional (asymptotic) bias considerations. However, the treatment of the bias terms as in our recent paper (Härdle & Gasser, 1984) suggests that it parallels the argumentations in the linear case. The consistency of $m_{n,1}(t)$ is established in the following theorem.

Theorem 1. Assume that (Z), (B), (C), (D) hold. Then, as $n \rightarrow \infty$, $m_{n,1}(t)$, the estimator as defined in (2.6), converges in probability to $m'(t)$ for all t , $0 < t < 1$.

Proof. We have already seen above that $D_n(t) \xrightarrow{p} q < 0$, so we only have to show that the numerator of $m_{n,1}(t)$ converges in probability to $qm'(t)$.

By Taylor's theorem we can write the numerator of $m_{n,1}(t)$ as

$$\begin{aligned} & \sum_{i=1}^n \alpha_{i,1}(t) \psi(Z_i) + \sum_{i=1}^n \alpha_{i,1}(t) \psi'(Z_i) \{m(t) - m_{n,1}(t)\} \\ & \quad + \sum_{i=1}^n \alpha_{i,1}(t) \psi''(Z_i) \{m(t) - m(t)\} \\ & \quad + \frac{1}{2} \sum_{i=1}^n \alpha_{i,1}(t) \psi''(\xi_i^{**}) \{m(t) - m_{n,1}(t)\}^2 \\ & = T_{1n} + T_{2n} + T_{3n} + T_{4n}. \end{aligned} \quad (3.1)$$

By assumption (Z) we have $E_F \psi(Z) = 0$ and therefore

$$P(|T_{1n}| > \varepsilon) \leq \varepsilon^{-2} \sum_{i=1}^n \{\alpha_{i,1}(t)\}^2 E_F \psi^2.$$

The RHS is of order $O(n^{-1} h^{-3})$ as is shown in the appendix of Gasser & Müller (1984), and therefore $T_{1n} \xrightarrow{p} 0$. In Härdle & Gasser (1984) it is shown that under the assumption of this theorem: $m_{n,1}(t) - m(t) = o_p(1)$. Now, since

$$\sum_{i=1}^n \alpha_{i,1}(t) \psi'(Z_i) = o_p(1)$$

it follows that $T_{2n} \xrightarrow{p} 0$.

Gasser & Müller (1984, formula (6)) have shown that

$$\begin{aligned} E m_{n,1}^*(t) &= h^{-1} \int_{-A}^A K_1'(u) m(t-uh) du + o(n^{-1} h^{-1}) \\ &= \int_{-A}^A K_1(u) m'(t-uh) du + o(n^{-1} h^{-1}). \end{aligned}$$

By continuity of m' , assumption (C), the RHS equals $m'(t) + o(1)$. This yields immediately, with

$$\sum_{i=1}^n \alpha_{i,1}(t) = 0,$$

$$ET_{1n} = qEm_{n,1}^*(t) = qm'(t) + o(1).$$

Define now $\eta = \psi(Z) - q$. By Chebyshev's inequality

$$P(|T_{1n} - ET_{1n}| > \varepsilon) = P\left\{ \left| \sum_{i=1}^n \alpha_{i,1}(t) \eta_i m(t) \right| > \varepsilon \right\} \leq \varepsilon^{-2} E \eta^2 \sum_{i=1}^n \{\alpha_{i,1}(t) m(t)\}^2$$

which tends to zero by Gasser & Müller, formula (7). It remains to show that $T_{2n} \xrightarrow{p} 0$. Note that by the assumption on K_1 and Lipschitz continuity of m , $m(t_i) = m(t) + o(h)$ for those indices i that contribute to the sum T_{2n} . We have therefore with a constant C_2 (bounding ψ''),

$$\begin{aligned} T_{2n} &= \frac{1}{2} \sum_{i=1}^n \alpha_{i,1}(t) \psi''(\xi_i^{(n)}) \{m_{n,0}^*(t) - m(t) + o(h)\}^2 \\ &\leq C_2 \sum_{i=1}^n \alpha_{i,1}(t) \cdot O_p\{m_{n,0}^*(t) - m(t)\} \end{aligned}$$

showing that $T_{2n} \xrightarrow{p} 0$.

The next theorem shows that $m_{n,1}(t)$ is asymptotically normally distributed.

Theorem 2. Assume that (Z), (B), (C), (D) hold and that furthermore $nh^2 \rightarrow 0$. Then, as $n \rightarrow \infty$,

$$\sqrt{nh^2} \{m_{n,1}(t) - m'(t)\}$$

is asymptotically normally distributed with mean zero and variance

$$V(\psi, F, K_1) = \int_{-A}^A \{K_1'(u)\}^2 du E_F \psi^2(Z) / \{E_F \psi'(Z)\}^2.$$

The following lemma will be needed.

Lemma 1. Under the assumptions of the theorem we have that with T_{1n} as defined in (3.1),

$$\sqrt{nh^2} \left[\sum_{i=1}^n \alpha_{i,1}(t) \psi\{Y_i - m_{n,0}(t)\} - ET_{1n} \right] \xrightarrow{d} N\{0, V_S(\psi, F, K_1)\}$$

where

$$V_S(\psi, F, K_1) = \int_{-A}^A \{K_1'(u)\}^2 du E_F \psi^2(Z).$$

Proof. The expansion (3.1) can be written as

$$\sum_{i=1}^n \alpha_{i,1}(t) \psi\{Y_i - m_{n,0}(t)\} - ET_{1n} = T_{1n} + T_{2n} + T_{3n} - ET_{1n} + T_{4n}.$$

Applying theorem 3 of Gasser & Müller (1984) yields that

$$\sqrt{nh^2} T_{1n} \xrightarrow{d} N\{0, V_S(\psi, F, K_1)\}.$$

It remains therefore to show that the remainder terms are of lower order than $\sqrt{nh^3}$. The asymptotic normality of $\sqrt{nh} \{M_{n,1}(t) - m(t)\}$, as shown in Härdle & Gasser (1984), yields that $T_{2,n} = o_p\{(nh^3)^{-1/2}\}$. The difference $T_{3,n} - ET_{3,n} = o_p(h)$ and by assumption $(\sqrt{nh^3}h)^2 = nh^5 \rightarrow 0$ which gives $T_{3,n} - ET_{3,n} = o_p\{(nh^3)^{-1/2}\}$. The term $T_{1,n}$ is handled similarly.

Proof of the theorem

Since $D_n(t) = q\{1 + o_p(1)\}$ we have the decomposition

$$m_{n,1}(t) - m'(t) = \left[\sum_{i=1}^n \alpha_{i,1}(t) \psi\{Y_i - m_{n,0}(t)\} - ET_{1,n} \right] / [q\{1 + o_p(1)\}] \\ + \{ET_{1,n} - qm'(t)\} / [q\{1 + o_p(1)\}].$$

The first term tends to the desired limit distribution and the second term is tending to zero in probability by the assumption $nh^5 \rightarrow 0$.

Remark

By assuming the existence of higher derivatives of $m'(t)$ it can be shown that $nh^5 \rightarrow 0$ can replace $nh^3 \rightarrow 0$. Specifically, the existence of a continuous third derivative yields a bias rate of $O(h^2)$ which, together with the unchanged rate for the variance $O(n^{-1}h^{-3})$, gives the optimal rate $n^{-4/7}$ in the sense of Stone (1980).

Example. When analysing longitudinal growth data, the estimation of derivatives is more important than estimating the growth curve itself (Gasser *et al.*, 1984). In order to check the efficacy of estimating derivatives robustly by the kernel method, the height data of a girl were analysed. An outlier was artificially produced by displacing the measurement at seven years (measurements were available yearly and during puberty halfyearly, from four weeks to 20 years). Fig. 1 shows a comparison of the velocity, obtained either linearly or robustly. For

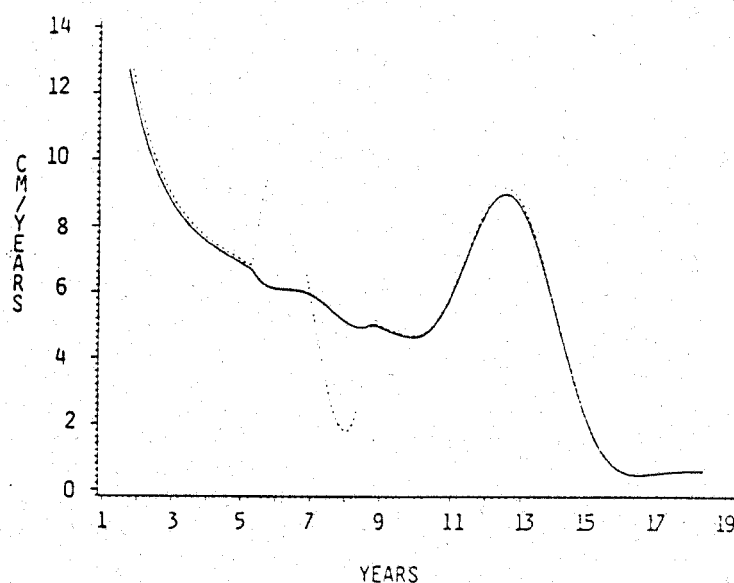


Fig. 1. Velocity of height of a girl (cm/year). Robust kernel estimate (dotted line) and linear kernel estimate (solid line).

both, a bandwidth of 1.7 years was used, following previous experience. For the robust part Huber's ψ -function was chosen and the optimal kernel of order (1.3) (Gasser *et al.* 1985). The outlier produces a large oscillation for the velocity curve of the ordinary kernel estimate, a pattern which is often of interest in practical applications. The robust estimate, on the other hand, suppresses the influence of the outlier to a large extent (a redescending ψ -function might be even more successful in doing so). For further examples of velocity of height growth see Gasser & Müller (1984) or Gasser *et al.* (1984). By visual judgement, velocity is affected much more by the outlier compared to the height curve itself when using the ordinary kernel estimate (the height curves are not displayed, since they are of minor interest for the present topic).

4. Estimation of higher derivatives

The introduction of $m_{n,1}(t)$ as an estimator for $m'(t)$ seemed to be somewhat odd since $m_{n,1}(t)$ operates linearly on non-linearly transformed estimated residuals. We will not give a heuristic justification for $m_{n,1}(t)$ which can be expanded to estimators for higher derivatives of $m(t)$. Recall the definition of $H_n(t, \theta)$. By definition of $m_{n,0}(t)$ we have $H_n(t, m_{n,0}(t))=0$, differentiating now formally w.r.t. t gives

$$0 = \sum_{i=1}^n \alpha_i^{(n)}(t) \psi\{Y_i^{(n)} - m_{n,0}(t)\} - m_{n,1}(t) \sum_{i=1}^n \alpha_i^{(n)}(t) \psi'\{Y_i - m_{n,0}(t)\} \tag{4.1}$$

which is just the definition of $m_{n,1}(t)$. Putting $\psi(u)=u$ and noting that

$$\sum_{i=1}^n \alpha_i^{(n)}(t) = 0$$

gives the linear estimate $m_{n,1}^*(t)$. Differentiating relation (4.1) once more gives

$$\begin{aligned} 0 &= \sum_{i=1}^n \alpha_i^{(n)}(t) \psi\{Y_i^{(n)} - m_{n,0}(t)\} - 2m_{n,1}(t) \sum_{i=1}^n \alpha_i^{(n)}(t) \psi'\{Y_i^{(n)} - m_{n,0}(t)\} \\ &\quad - m_{n,2}(t) \sum_{i=1}^n \alpha_i^{(n)}(t) \psi''\{Y_i^{(n)} - m_{n,0}(t)\} - \{m_{n,1}(t)\}^2 \sum_{i=1}^n \alpha_i^{(n)}(t) \psi''\{Y_i^{(n)} - m_{n,0}(t)\} \\ &= N_n(t) - 2m_{n,1}(t)R_{1n} - m_{n,2}(t)D_n(t) - \{m_{n,1}(t)\}^2R_{2n}. \end{aligned} \tag{4.2}$$

Here $\alpha_i^{(n)}(t)$ denote the kernel weights when estimating second derivatives and $m_{n,2}(t)$ is the (formal) derivative of $m_{n,1}(t)$.

Assume now that ψ'' exists, then by the same arguments that we used in the proof of theorem 1 with ψ'' in the place of ψ it follows that $R_{2n} = 0_p(1)$, provided $E_F \psi''(Z) = 0$. This condition can be easily met, for instance, by symmetry arguments. If ψ is antisymmetric, as was assumed, and F is symmetric, then $E_F \psi(Z) = 0$ and so is $E_F \psi''(Z)$. Expanding R_{1n} in a Taylor series as in (3.1) shows that the leading term is

$$\sum_{i=1}^n \alpha_i^{(n)}(t) \psi'(Z_i^{(n)})$$

which has mean zero and variance

$$\sum_{i=1}^n \{\alpha_i^{(n)}(t)\}^2 q^2 = 0(n^{-1} h^{-3}).$$

Hence, $R_{1n} = O_p(1)$ and equation (4.2) can now be rewritten as

$$m_{n,2}(t) = \sum_{i=1}^n \alpha_i^{(2)}(t) \psi(Y_i^{(2)} - m_{n,1}(t)) / D_n(t). \quad (4.3)$$

This definition of $m_{n,2}(t)$ as an estimator for $m''(t)$ has the same structure as $m_{n,1}(t)$. In both cases we are essentially applying an ordinary linear kernel estimate for derivatives to non-linearly transformed estimated residuals. That is, ignoring the effect of the randomness of $D_n(t)$, $m_{n,2}(t)$ is roughly a kernel estimate operating for each t on non-observable pseudo data

$$\tilde{Y}_i^{(2)} = \psi(Y_i^{(2)} - m_{n,1}(t)) / q.$$

Carrying out the same arguments as above for (4.2) we see that

$$m_{n,p}(t) = \sum_{i=1}^n \alpha_i^{(p)}(t) \{Y_i^{(p)} - m_{n,p-1}(t)\} / D_n(t), \quad (4.4)$$

with $\alpha_i^{(p)}(t)$ the weights for a linear kernel estimate of the p th derivative $m^{(p)}(t)$ of $m(t)$, will be a reasonable estimator for $m^{(p)}(t)$. We will not pursue the analysis of $m_{n,p}(t)$ since the technical details are straightforward given the arguments for $m_{n,1}$ and $m_{n,2}$.

Acknowledgements

This research was partially supported by the Deutsche Forschungsgemeinschaft, SFB 123 "Stochastische Mathematische Modelle". We would also like to thank an Associate Editor and an anonymous referee for helpful suggestions and improvement of the presentation.

References

- Bahill, A. T. & Stark, L. (1979). The trajectory of saccadic eye movements. *Scient. Am.* **240**, 108-117.
- Cox, D. D. (1983). Asymptotics for M -type smoothing splines. *Ann. Statist.* **11**, 530-551.
- Gasser, Th. & Müller, H. G. (1979). Kernel estimation of regression functions. In *Smoothing techniques for curve estimation* (ed. Th. Gasser & M. Rosenblatt), *Lecture Notes in Mathematics*, No. 757, pp. 23-68. Springer Verlag, Berlin.
- Gasser, Th. & Müller, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11**, 171-185.
- Gasser, Th., Müller, H. G., Köhler, W., Molinari, L. & Prader, A. (1984). Non-parametric regression analysis of growth curves. *Ann. Statist.* **12**, 210-229.
- Gasser, Th., Müller, H. G. & Mammitzsch, V. (1985). Kernels for non-parametric curve estimation. *J. R. Stat. Soc. B* (to appear).
- Härdle, W. & Gasser, Th. (1984). Robust non-parametric function fitting. *J. R. Statist. Soc. B* **46**, 42-51.
- Härdle, W. (1984a). Robust non-parametric regression function estimation. *J. Mult. Anal.* **14**, 169-180.
- Härdle, W. (1984b). How to determine the bandwidth of non-linear smoothers in practice? In *Robust and non-linear time series analysis* (ed. J. Franke, W. Härdle & D. Martin), *Lecture Notes in Statistics*, No. 26, pp. 163-184. Springer Verlag, Heidelberg.
- Huber, P. J. (1979). *Robustness in statistics*. Proceedings of a workshop 1978 (ed. Robert L. Launer & Graham N. Wilkinson). Academic Press, New York.
- Huber, P. J. (1981). *Robust statistics*. Wiley, New York.
- Largo, R. H., Gasser, Th., Prader, A., Stützel, W. & Huber, P. J. (1978). Analysis of the adolescent growth spurt using smoothing spline functions. *Ann. Hum. Biol.* **5**, 421-434.
- Priestley, M. B. & Chao, M. T. (1972). Non-parametric function fitting. *J. R. Statist. Soc. B* **34**, 385-392.
- Stone, C. J. (1980). Optimal rates of convergence for non-parametric estimators. *Ann. Statist.* **8**, 1348-1360.
- Tsybakov, A. B. (1983). Robust estimates of a function. *Problems of Information Transmission* **1**, 190-201.

Received May 1984, in final form April 1985

Theo Gasser, Zentralinstitut für Seelische Gesundheit, POB 5970, D-6800 Mannheim

OPTIMAL BANDWIDTH SELECTION IN NONPARAMETRIC REGRESSION FUNCTION ESTIMATION

BY WOLFGANG HÄRDLE¹ AND JAMES STEPHEN MARRON²

*Universität Heidelberg and University of North Carolina at Chapel Hill
and University of North Carolina at Chapel Hill*

Kernel estimators of an unknown multivariate regression function are investigated. A bandwidth-selection rule is considered, which can be formulated in terms of cross validation. Under mild assumptions on the kernel and the unknown regression function, it is seen that this rule is asymptotically optimal.

1. Introduction. Let $(X, Y), (X_1, Y_1), (X_2, Y_2), \dots$ be independent identically distributed \mathbb{R}^{d+1} valued random vectors with Y real valued. Consider the problem of estimating the regression function,

$$m(x) = E[Y|X = x],$$

using $(X_1, Y_1), \dots, (X_n, Y_n)$. In this paper, kernel estimators with a data-driven bandwidth are investigated. Asymptotic optimality is established for a bandwidth-selection rule which can be interpreted in terms of cross validation. The results address two issues. First, they are important in exploratory data analysis, [see, for example, the Projection Pursuit Regression algorithm given in Friedman and Stuetzle (1981).] Second, they settle an open problem of Stone (1982).

Kernel estimators, as introduced by Nadaraya (1964) and Watson (1964), are a local weighted average of the Y_i given by

$$\hat{m}(x) = \hat{m}_h(x) = n^{-1} \sum_{i=1}^n h^{-d} K\left(\frac{x - X_i}{h}\right) Y_i / \hat{f}_h(x),$$

where $K: \mathbb{R}^d \rightarrow \mathbb{R}$ is a kernel (i.e., window) function, $h = h(n) \in \mathbb{R}^+$ is the bandwidth (i.e., smoothing parameter), and $\hat{f}_h(x)$ is the familiar Rosenblatt-Parzen kernel density estimator,

$$\hat{f}(x) = \hat{f}_h(x) = n^{-1} \sum_{i=1}^n h^{-d} K\left(\frac{x - X_i}{h}\right),$$

of the marginal density $f(x)$ of X . A slight generalization of this estimator may be obtained by allowing h to be a d -dimensional vector or even a $d \times d$ matrix. The results of this paper extend to that case in a straightforward fashion, although for simplicity of presentation, only scalar h is treated here.

Received May 1983; revised June 1984, December 1984, and May 1985.

¹Research partially supported by Air Force Office of Scientific Research Contract AFOSR-F49620-82-C0009 and the Deutsche Forschungsgemeinschaft, SFB123.

²Research partially supported by Office of Naval Research Contract N00014-81-K-0373.

AMS 1980 subject classifications. Primary 62G05; secondary 62G20.

Key words and phrases. Nonparametric regression estimation, kernel estimators, optimal bandwidth, smoothing parameter, cross validation.

One of the crucial points in applying \hat{m}_h is the choice of the bandwidth h . Suppose that h is in some set $H_n \subseteq \mathbb{R}_n^+$ of interest. A *bandwidth-selection rule* $\hat{h} = \hat{h}(n)$ is an H_n -valued function of $(X_1, Y_1), \dots, (X_n, Y_n)$. Let the distance $d(\hat{m}_h, m)$ denote a given measure of accuracy for the estimator \hat{m}_h . Following Shibata (1981), the bandwidth-selection rule \hat{h} is said to be *asymptotically optimal with respect to d* when

$$\lim_{n \rightarrow \infty} \left[\frac{d(\hat{m}_{\hat{h}}, m)}{\inf_{h \in H_n} d(\hat{m}_h, m)} \right] = 1,$$

with probability one.

In this paper, a bandwidth-selection rule is given, which is then shown to be asymptotically optimal with respect to the distances:

Averaged Squared Error:

$$d_A(\hat{m}, m) = n^{-1} \sum_{j=1}^n [\hat{m}(X_j) - m(X_j)]^2 w(X_j);$$

Integrated Squared Error:

$$d_I(\hat{m}, m) = \int [\hat{m}(x) - m(x)]^2 w(x) f(x) dx;$$

Conditional Mean Integrated Squared Error:

$$d_C(\hat{m}, m) = E [d_I(\hat{m}, m) | X_1, \dots, X_n],$$

where $w(x)$ is a nonnegative weight function.

A bandwidth-selection rule \hat{h} will now be motivated. Write

$$d_I(\hat{m}_h, m) = \int \hat{m}_h^2 w f - 2 \int \hat{m}_h m w f + \int m^2 w f.$$

Since the last summand is independent of h , the goal of minimizing this loss is equivalent to that of minimizing

$$(1.1) \quad \int \hat{m}_h^2 w f - 2 \int \hat{m}_h m w f.$$

But this cannot be realized in practice because this quantity depends on the unknowns m and f . Observe, however, that the second term, for instance, may be written as

$$\int \hat{m}_h m w f = E_{(X, Y)} [\hat{m}_h(X) Y w(X)].$$

This motivates estimating the second term by

$$n^{-1} \sum_{j=1}^n [\hat{m}_j(X_j) Y_j w(X_j)],$$

where \hat{m}_j is the "leave-one-out" estimator given by

$$(1.2) \quad \begin{aligned} \hat{m}_j(x) &= (n-1)^{-1} \sum_{i \neq j} h^{-d} K\left(\frac{x-X_i}{h}\right) Y_i / \hat{f}_j(x), \\ \hat{f}_j(x) &= (n-1)^{-1} \sum_{i \neq j} h^{-d} K\left(\frac{x-X_i}{h}\right). \end{aligned}$$

Similarly, the first term of (1.1) may be approximated by

$$n^{-1} \sum_{j=1}^n [\hat{m}_j^2(X_j) w(X_j)].$$

Thus, it seems reasonable to take h to minimize the sum of the estimates of the first two terms. Adding a term which is independent of h does not change the bandwidth-selection rule, which is then:

Choose \hat{h} to minimize

$$CV(h) = n^{-1} \sum_{j=1}^n [Y_j \hat{m}_j(X_j)]^2 w(X_j).$$

The above motivation is related to some ideas of Rudemo (1982) and Bowman (1984).

Note that the bandwidth-selection rule \hat{h} may also be thought of in terms of choosing h to make $\hat{m}_j(X_j)$ an effective predictor of Y_j . This approach, based on the idea of cross validation, was taken by Clark (1975) and Wahba and Wold (1975) in the setting of spline estimation. See Rice (1984) and Härdle and Marron (1985) for a discussion of other asymptotically optimal bandwidth selectors.

In Section 2, a theorem is stated which shows that this bandwidth-selection rule is asymptotically optimal with respect to the distances d_A, d_I, d_C . In Section 3 it is seen how the theorem of Section 2 provides an answer to Question 3 of Stone (1982). Section 4 demonstrates an application of these results. The rest of the paper consists of proofs.

2. Asymptotic optimality. Assume the weight function w is bounded and supported on a compact set with nonempty interior. Assumptions to be made on the bandwidth, the kernel, and the probability distribution of (X, Y) are:

(A.1) For $n = 1, 2, \dots$ $H_n = [\underline{h}, \bar{h}]$ where

$$\underline{h} \geq C^{-1} n^{\delta-1/d}, \quad \bar{h} \leq C n^{-\delta},$$

for some constants $C, \delta > 0$.

(A.2) K is Hölder continuous, ie,

$$|K(x) - K(t)| \leq C \|x - t\|^\xi,$$

where $\|\cdot\|$ denotes Euclidean norm on \mathbb{R}^d , and also

$$\int K(u) du = 1,$$

$$\int \|u\|^\xi |K(u)| du < \infty.$$

(A.3) The regression function m and the marginal density j are Hölder continuous.

(A.4) The conditional moments of Y given $X = x$ are bounded in the sense that there are positive constants C_1, C_2, \dots so that for $i = 1, 2, \dots$

$$E[|Y|^i | X = x] \leq C_i \quad \text{for all } x.$$

(A.5) The marginal density $f(x)$ of X is bounded from below on the support of w .

(A.6) The marginal density $f(x)$ of X is compactly supported.

THEOREM 1. *Under the assumptions (A.1)–(A.6), the bandwidth-selection rule, “choose h to minimize $CV(h)$,” is asymptotically optimal with respect to the distances $d_A, d_I,$ and d_C .*

Condition (A.1) may appear somewhat restrictive because minimization is being performed over an interval whose length tends to zero. This is not a severe restriction because in order to obtain the consistency of \hat{m} , the bandwidth must satisfy some similar condition.

The condition (A.4) is substantially weaker than the boundedness conditions on Y that have been imposed by a number of authors, starting with Nadaraya (1964). This condition may be weakened to only a certain finite number of conditional moments being bounded.

Condition (A.5) allows handling of the random denominator of $\hat{m}(x)$. Also, since by (A.3), f and m are assumed to be continuous beyond the support of w , any concern about “boundary effects,” such as those described by Gasser and Müller (1979), and Rice and Rosenblatt (1983) is eliminated.

The assumption (A.6) is added for convenience in the proof. It may be weakened to either the existence of any moment of X , or to the compact support of K .

The techniques of this paper may also be applied to estimators related to \hat{m} . For example, if the marginal density f is known, as in the “fixed-design” (ie, X not random) case, it makes sense to consider the estimator

$$n^{-1} \sum_{i=1}^n h^{-d} K\left(\frac{x - X_i}{h}\right) Y_i / f(X_i),$$

as studied by Johnston (1982).

3. Stone’s Question 3. Stone (1982) investigates the way in which the rate of convergence of nonparametric regression estimators depends on the smoothness of the regression functions. In particular, Stone defines smoothness classes Θ_r , indexed by $r \in \mathbb{R}^+$, and finds an estimator \hat{m} , depending on r , which “achieves the rate of convergence r ” in the sense that there is a constant C so that

$$\lim_{n \rightarrow \infty} \sup_{m \in \Theta_r} P_m [d_I(\hat{m}, m) \geq Cn^{-r}] = 0,$$

where the notation P_m is used to indicate parametrization by m . [See Stone (1982) for the details.] Stone then shows that the rate of convergence r is "optimal" by showing that no estimator of any type can have a faster rate of convergence uniformly over Θ_r . Stone's Question 3 may be expressed as: Is there an estimator \hat{m} , independent of r , which achieves the optimal rate uniformly over the smoothness classes?

Under an additional assumption on the smoothness of the marginal density of X , an estimator having this property can be obtained by using a kernel estimator with bandwidth selected as above:

THEOREM 2. *Given $\eta \in (0, \frac{1}{2})$, there is a kernel K and a constant $C_r > 0$ so that, under the assumptions (A.1)–(A.6),*

$$\lim_{n \rightarrow \infty} \sup_{r \in [\eta, 1-\eta]} \sup_{f, m \in \Theta_r} P_{f, m} [d_f(\hat{m}_h, m) \geq C_r n^{-r}] = 0.$$

The proof of Theorem 2 is in Section 10.

4. An application. In this section it is seen how the proposed kernel regression estimator performs in a real life example. The data consist of 300 pairs of variables where Y denotes liver weight and X denotes age (note here $d = 1$), gathered by the Institute of Forensic Medicine, Universität Heidelberg. It is apparent from the scatter diagram (Figure 1) that the data are quite nonlinear and heteroscedastic, so that a nonparametric approach seems reasonable.

The above theorems make the choice of the smoothing parameter automatic, but there are several quantities that still must be chosen. It is well known [see Table 1 of Rosenblatt (1971)] that the choice of the kernel function, K , is of relatively small importance. We used the kernel of Epanechnikov (1969) given by

$$K(u) = 3(1 - u^2)1_{[-1,1]}(u)/4.$$

Of more concern is the choice of the weight function, w , and through w the choice of its support S . To study the effect on our estimators of different choices of S , we chose

$$w(x) = 1_{[\Delta x, 100 - \Delta x]}(x),$$

where several different values of Δx were considered. Figure 2 shows the graph of the cross-validation function for several choices of Δx . Note the minimum is roughly at $h = 22$ except in the extreme case $\Delta x = 10$ where about 20% of the data has been deleted.

Since this is a real data set, it is impossible to show that $h = 22$ optimizes any of d_A , d_f , or d_C , but Figure 3 allows some comparison. The bandwidths 14 and 30 give regression estimates $\hat{m}(x)$ which seem under (and over, respectively) smoothed. For a final comparison, Figure 1 shows how $\hat{m}(x)$ with $h = 22$ fits the data.

Thus, at least in this example, the techniques of this paper seem relatively independent of the choice of S .

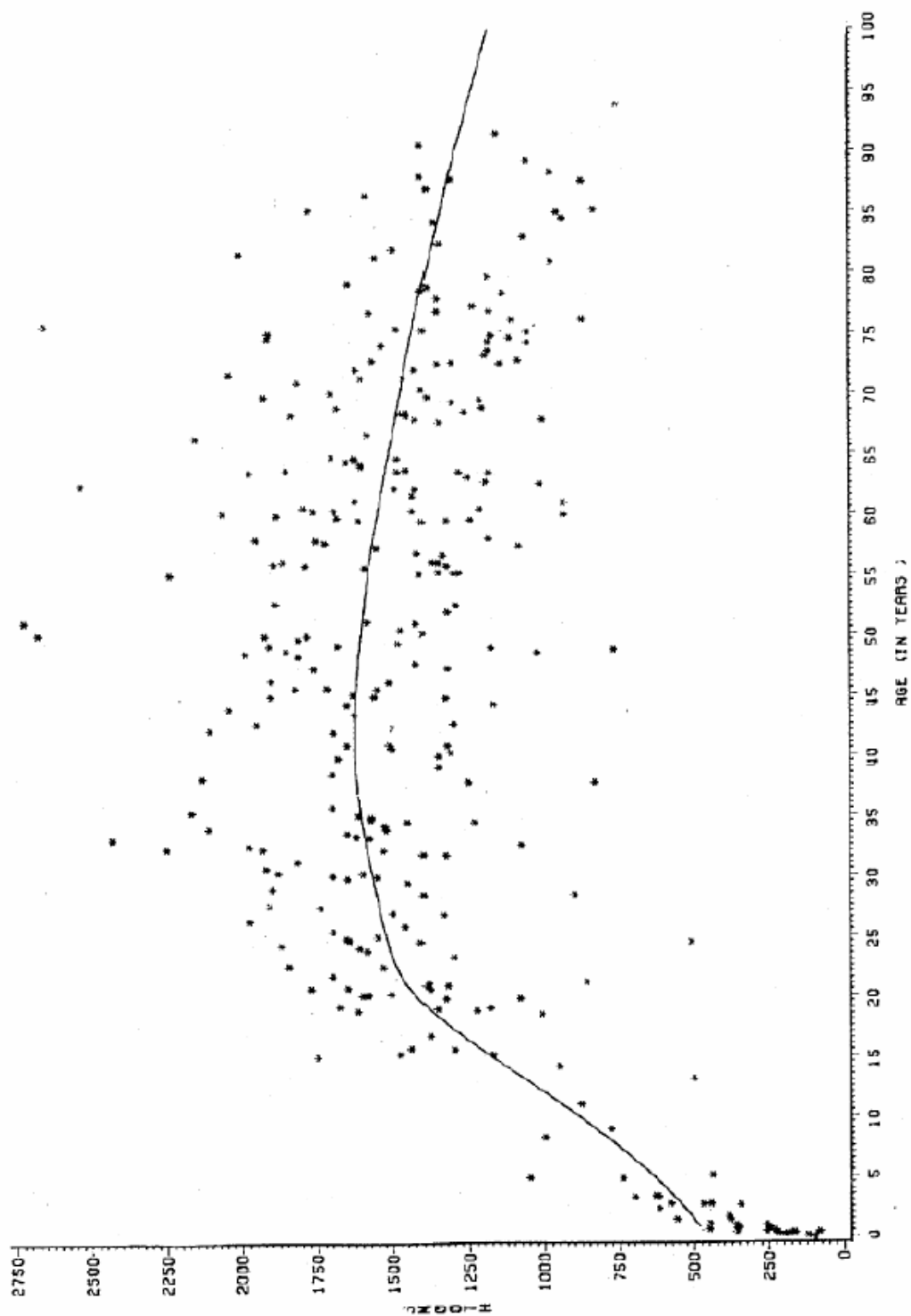


FIG. 1. Liver weights * age of 300 female persons (smoothed with kernel estimate $h = .22$).

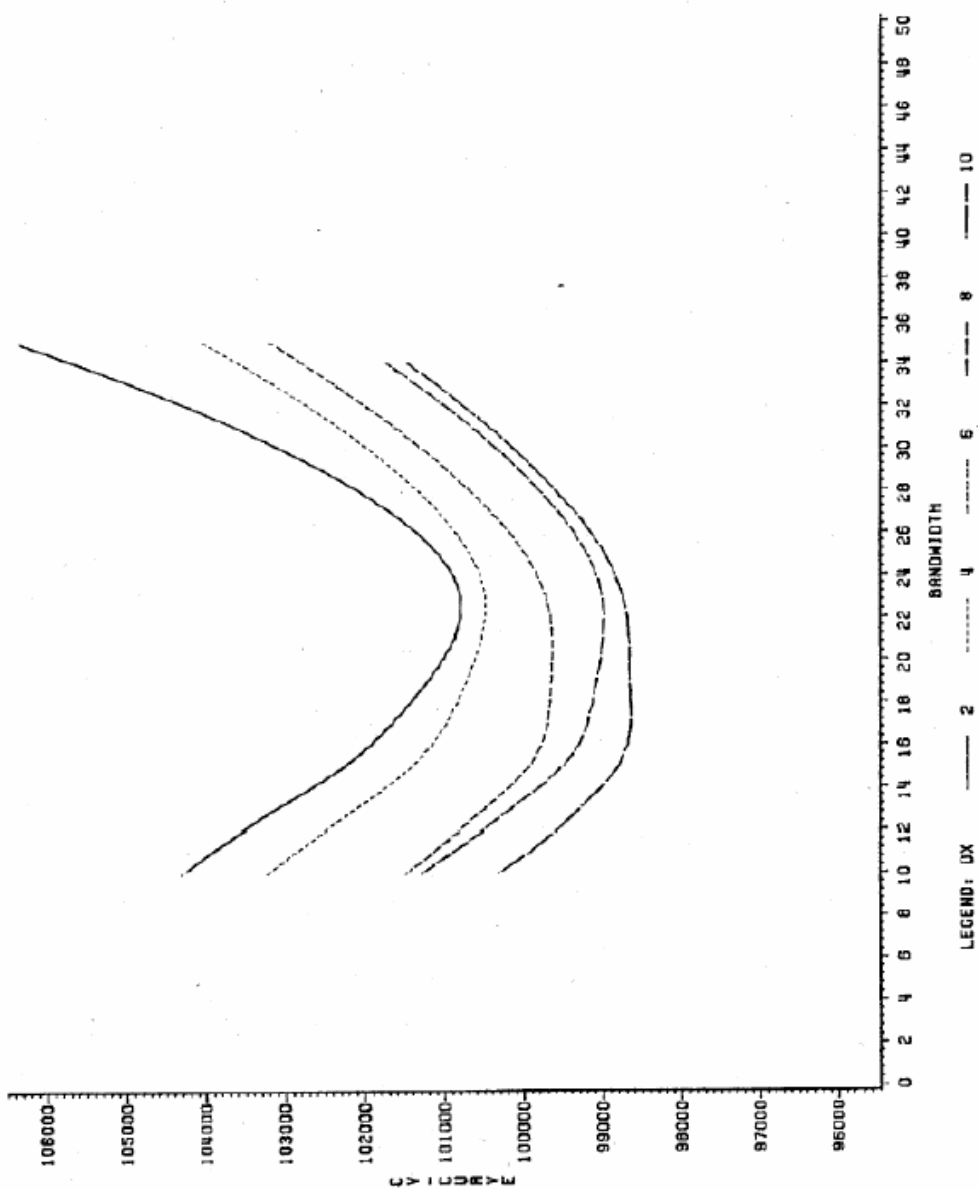


FIG. 2. CV-graph for liver weight data ($n = 300$, $\Delta x = 2, 4, 6, 8, 10$, spacing = .01).

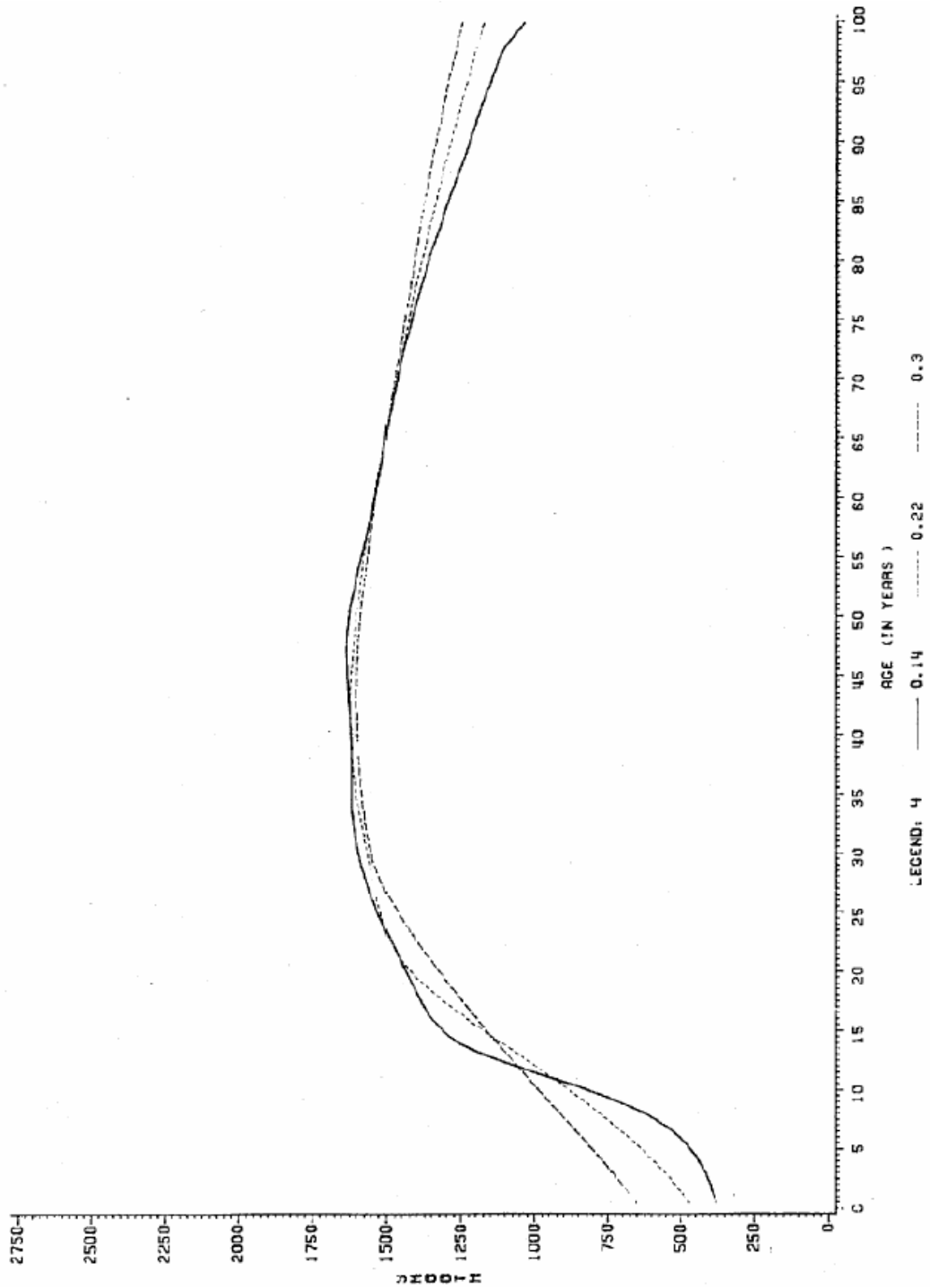


FIG. 3. Liver weights * age of 300 female persons.

Härdle, W. and Marron, S. (1985) Optimal Bandwidth Selection in Nonparametric Regression Function Estimation.

5. Proof of Theorem 1. A difficult feature, from an analytical point of view, of the estimator \hat{m} is that it has a random denominator. This will be dealt with by the following device. For x in the compact support of w , write

$$(5.1) \quad \hat{m} - m = (\hat{m} - m) \hat{f}/f + (\hat{m} - m)(f - \hat{f})/f.$$

Note that by the uniform consistency of \hat{f} to f (see Lemma 1 below), the second term is negligible compared to the first [in a sense that is made precise in (5.3) below]. Hence the following distances will be considered

$$\begin{aligned} d_A^*(\hat{m}, m) &= d_A(\hat{m}\hat{f}/f, m\hat{f}/f), \\ d_I^*(\hat{m}, m) &= d_I(\hat{m}\hat{f}/f, m\hat{f}/f), \\ d_C^*(\hat{m}, m) &= E[d_I^*(\hat{m}, m)|X_1, \dots, X_n], \end{aligned}$$

and also

$$d_M^*(\hat{m}, m) = E[d_I^*(\hat{m}, m)].$$

[The unstarred analogue of d_M^* is not considered here because it may fail to exist, see Härdle and Marron (1983).]

Marron and Härdle (1984) have shown that, under the assumption of Theorem 1,

$$(5.2) \quad \begin{aligned} \sup_h \left| \frac{d_A^*(\hat{m}_h, m) - d_M^*(\hat{m}_h, m)}{d_M^*(\hat{m}_h, m)} \right| &\rightarrow 0 \text{ a.s.}, \\ \sup_h \left| \frac{d_I^*(\hat{m}_h, m) - d_M^*(\hat{m}_h, m)}{d_M^*(\hat{m}_h, m)} \right| &\rightarrow 0 \text{ a.s.}, \end{aligned}$$

where \sup_h denotes supremum over H_n . (Actually, this is shown for h in a finite set H'_n , whose cardinality grows only algebraically fast, but that can be easily extended to $H_n = [h, \bar{h}]$ by a Hölder continuity argument like that used in the proof of the following Lemma 1.) In the rest of this paper, H'_n will denote a finite subset of H_n whose cardinality is bounded by n^ρ , for some $\rho > 0$. The fact that d_A , d_I , d_C , and d_C^* are also similar to d_M^* in the sense (5.2) is the key to the proof.

A substantial part of this is the verification of:

LEMMA 1. *If (A.1), (A.2), (A.3), and (A.6) hold, then for any compact set $S \subset \mathbb{R}^d$*

$$\sup_{x \in S} \sup_h |f_h(x) - f(x)| \rightarrow 0 \text{ a.s.}$$

The proof of Lemma 1 is in Section 6.

It follows immediately from Lemma 1, (5.1), and (5.2) that

$$(5.3) \quad \begin{aligned} \sup_h \left| \frac{d_A(\hat{m}_h, m) - d_M^*(\hat{m}_h, m)}{d_M^*(\hat{m}_h, m)} \right| &\rightarrow 0 \text{ a.s.} \\ \sup_h \left| \frac{d_I(\hat{m}_h, m) - d_M^*(\hat{m}_h, m)}{d_M^*(\hat{m}_h, m)} \right| &\rightarrow 0 \text{ a.s.} \end{aligned}$$

In a similar spirit,

$$\sup_h \left| \frac{d_C(\hat{m}_h, m) - d_M^*(\hat{m}_h, m)}{d_M^*(\hat{m}_h, m)} \right| \rightarrow 0 \text{ a.s.}$$

follows from:

LEMMA 2. *Under the assumptions of Theorem 1*

$$\sup_h \left| \frac{d_C^*(\hat{m}_h, m) - d_M^*(\hat{m}_h, m)}{d_M^*(\hat{m}_h, m)} \right| \rightarrow 0 \text{ a.s.}$$

The proof of Lemma 2 is in Section 7.

Next, to bridge the gap between d_A and $CV(h)$, using the notation (1.2), define

$$\begin{aligned} \bar{d}_A(\hat{m}, m) &= n^{-1} \sum_{j=1}^n [\hat{m}_j(X_j) - m(X_j)]^2 w(X_j), \\ \bar{d}_A^*(\hat{m}, m) &= n^{-1} \sum_{j=1}^n [\hat{m}_j(X_j) - m(X_j)]^2 f_j(X_j)^2 f(X_j)^{-2} w(X_j). \end{aligned}$$

Note that, for $j = 1, \dots, n$,

$$(5.4) \quad \hat{f}_j(x) - \hat{f}(x) = (n-1)^{-1} \hat{f}(x) - (n-1)^{-1} h^{-d} K\left(\frac{x - X_j}{h}\right).$$

This relationship and (8.1) allow expressions containing the leave-one-out estimators to be approximated by the same expressions in terms of the ordinary estimators. Thus, by Lemma 1 and (A.1)

$$(5.5) \quad \sup_{j=1, \dots, n} \sup_x \sup_h |\hat{f}_j(x) - \hat{f}(x)| \rightarrow 0 \text{ a.s.},$$

where \sup_x denotes supremum over the support of w . So, as above, with \hat{m} and \hat{f} replaced by \hat{m}_j and \hat{f}_j in (5.1),

$$\sup_h \left| \frac{\bar{d}_A(\hat{m}, m) - d_M^*(\hat{m}, m)}{d_M^*(\hat{m}, m)} \right| \rightarrow 0 \text{ a.s.}$$

follows from:

LEMMA 3. *Under the assumption of Theorem 1*

$$\sup_h \left| \frac{\bar{d}_A^*(\hat{m}, m) - d_M^*(\hat{m}, m)}{d_M^*(\hat{m}, m)} \right| \rightarrow 0 \text{ a.s.}$$

The proof of Lemma 3 is in Section 8.

Let d denote any of $d_A, d_I, d_C, d_A^*, d_I^*, d_C^*, d_M^*, \bar{d}_A$, or \bar{d}_A^* . To show

$$(5.6) \quad \frac{d(\hat{m}_h, m)}{\inf_h d(\hat{m}_h, m)} \rightarrow 1 \text{ a.s.},$$

it is enough to check that

$$\sup_{h, h'} \frac{|d(\hat{m}_h, m) - d(\hat{m}_{h'}, m) - (CV(h) - CV(h'))|}{d(\hat{m}_h, m) + d(\hat{m}_{h'}, m)} \rightarrow 0 \text{ a.s.}$$

But in view of the above equivalences, this may be done by showing

$$(5.7) \quad \sup_{h, h'} \left| \frac{\bar{d}_A(\hat{m}_h, m) - \bar{d}_A(\hat{m}_{h'}, m) - (CV(h) - CV(h'))}{d_M^*(\hat{m}_h, m) + d_M^*(\hat{m}_{h'}, m)} \right| \rightarrow 0 \text{ a.s.}$$

To check this write

$$(5.8) \quad \bar{d}_A(\hat{m}_h, m) - CV(h) = 2 \text{Cross}(h) + n^{-1} \sum_{j=1}^n [m(X_j) - Y_j]^2 w(X_j),$$

where

$$\text{Cross}(h) = n^{-1} \sum_{j=1}^n (\hat{m}_j(X_j) - m(X_j))(m(X_j) - Y_j)w(X_j).$$

Note that the last term on the right of (5.8) is independent of h . So the proof of (5.7) and hence of Theorem 1 will be finished when it is seen that:

LEMMA 4. *Under the assumptions of Theorem 1*

$$\sup_h |\text{Cross}(h) / d_M^*(\hat{m}_h, m)| \rightarrow 0 \text{ a.s.}$$

The proof of Lemma 4 is in Section 9.

6. Proof of Lemma 1. Given $\eta > 0$, for $n = 1, 2, \dots$, find a set $H'_n \subset H_n$ and a set $C'_n \subset C$ so that for any $h \in H_n$ and any $x \in C$, there is $h' \in H'_n$ and $x' \in C'_n$ with

$$|h - h'| \leq n^{-\eta} \text{ and } |x - x'| \leq n^{-\eta}.$$

Note that H'_n and C'_n can be chosen so that their cardinality increases algebraically fast in $n \rightarrow \infty$.

Given $\epsilon > 0$,

$$P \left[\sup_{h \in H_n} \sup_{x \in C} |\hat{f}(x, h) - f(x)| > \epsilon \right] \leq I_n + II_n,$$

where

$$I_n = P \left[\sup_{h' \in H'_n} \sup_{x' \in C'_n} |\hat{f}(x', h') - f(x')| > \frac{\epsilon}{2} \right],$$

$$II_n = P \left[\sup_{h, h', x, x'} |\hat{f}(x, h) - f(x) - (\hat{f}(x', h') - f(x'))| > \frac{\epsilon}{2} \right],$$

and where $\sup_{h, h', x, x'}$ denotes supremum over $h \in H_n, h' \in H'_n, x \in C$, and

$x' \in C'_n$. By the Borel–Cantelli Lemma, the proof of Lemma 1 is complete when it is seen that

$$(6.1) \quad \sum_{n=1}^{\infty} I_n < \infty,$$

$$(6.2) \quad \sum_{n=1}^{\infty} II_n < \infty.$$

An argument based on Bernstein’s Inequality (Hoeffding, 1963), quite similar to the proof of Lemma 2 of Stone (1984), may be used to establish (6.1). The verification of (6.2) follows in a straightforward fashion from the Hölder continuity of f and K .

7. Proof of Lemma 2. Write

$$d_C^*(\hat{m}_h, m) = \int \left[n^{-1} \sum_{i=1}^n \delta_h(x, X_i) \right]^2 f(x)^{-2} w(x) dx,$$

where

$$\delta_h(x, X_i) = h^{-d} K\left(\frac{x - X_i}{h}\right) [m(X_i) - m(x)].$$

Under the assumptions of Theorem 1,

$$n^{-1} \sum_{i=1}^n \delta_h(x, X_i)$$

is a so called delta sequence estimator [of $g(x) \equiv 0$] which satisfies the conditions of Theorem 1 in Marron and Härdle (1984). Hence,

$$\sup_{h \in H'_n} \left| \frac{d_C^*(\hat{m}_h, m) - d_M^*(\hat{m}_h, m)}{d_M^*(\hat{m}_h, m)} \right| \rightarrow 0 \quad \text{a.s.}$$

The above supremum may be easily extended to $H_n = [\underline{h}, \bar{h}]$ by taking the points of H'_n to be sufficiently close together and then using a Hölder continuity argument.

8. Proof of Lemma 3. First note that, as in (5.4), for $j = 1, \dots, n$

$$(8.1) \quad \begin{aligned} & \hat{m}_j(x) \hat{f}_j(x) - \hat{m}(x) \hat{f}(x) \\ &= (n-1)^{-1} \hat{m}(x) \hat{f}(x) - (n-1)^{-1} h^{-d} K\left(\frac{x - X_j}{h}\right) Y_j. \end{aligned}$$

In the following the functions $m, \hat{m}, \hat{m}_j, f, \hat{f}, \hat{f}_j$, and w will be always evaluated at X_j , so it is to be understood that “ m ” means “ $m(X_j)$ ”, and so on. Write

$$\bar{d}_A^*(\hat{m}_h, m) = n^{-1} \sum_{j=1}^n [A_j + (\hat{m} \hat{f} - m \hat{f})]^2 f^{-2} w,$$

where

$$\begin{aligned} A_j &= \hat{m}_j \hat{f}_j - m \hat{f}_j - (\hat{m}f - mf) \\ &= (n-1)^{-1} [\hat{m}f - mf - h^{-d}K(0)(Y_j - m)]. \end{aligned}$$

Then

$$\begin{aligned} \bar{d}_A^*(\hat{m}_h, m) - d_A^*(\hat{m}_h, m) &= n^{-1} \sum_{j=1}^n (A_j^2 + 2A_j(\hat{m}f - mf)) f^{-2}w \\ &= ((n-1)^{-2} + 2(n-1)^{-1}) d_A^*(\hat{m}_h, m) \\ &\quad - 2((n-1)^{-2} + (n-1)^{-1}) n^{-1} \\ &\quad \cdot \sum_{j=1}^n (\hat{m}f - mf) h^{-d}K(0)(Y_j - m) f^{-2}w \\ &\quad + (n-1)^{-2} n^{-1} \sum_{j=1}^n h^{-2d}K(0)^2(Y_j - m)^2 f^{-2}w. \end{aligned} \tag{8.2}$$

But, by the Schwartz Inequality,

$$\begin{aligned} &\left| n^{-1} \sum_{j=1}^n (\hat{m}f - mf) h^{-d}K(0)(Y_j - m) f^{-2}w \right| \\ &\leq (d_A^*(\hat{m}_h, m))^{1/2} h^{-d}K(0) \left(n^{-1} \sum_{j=1}^n (Y_j - m)^2 f^{-2}w \right)^{1/2}, \end{aligned} \tag{8.3}$$

and by the Strong Law of Large Numbers,

$$n^{-1} \sum_{j=1}^n (Y_j - m)^2 f^{-2}w \rightarrow E((Y_j - m)^2 f^{-2}w) \quad \text{a.s.} \tag{8.4}$$

By a variance-bias² decomposition [see, for example, Parzen (1962), Rosenblatt (1969, 1971)], $d_M^*(\hat{m}_h, m)$ can be written

$$\begin{aligned} d_M^*(\hat{m}_h, m) &= n^{-1} h^{-d} \left[\int V(x) w(x) dx \right] \left[\int K(u)^2 du \right] \\ &\quad + o(n^{-1} h^{-d}) + b^2(h), \end{aligned} \tag{8.5}$$

where the o is uniform over $h \in H_n$, where $V(x)$ denotes the conditional variance

$$V(x) = E[Y^2 - m(x)^2 | X = x],$$

and where the part analogous to squared bias is denoted

$$\begin{aligned} b^2(h) &= \int \left[\int K(u) [m(x - hu) - m(x)] \right. \\ &\quad \left. \cdot f(x - hu) du \right]^2 f(x)^{-1} w(x) dx. \end{aligned} \tag{8.6}$$

It follows from (5.2), (8.2), (8.3), (8.4), and (8.5) that

$$\sup_h \frac{|d_A^*(\hat{m}_h, m) - d_M^*(\hat{m}_h, m)|}{d_M^*(\hat{m}_h, m)} \rightarrow 0 \text{ a.s.}$$

This completes the proof of Lemma 3.

9. Proof of Lemma 4. By the expansion (5.1), with \hat{m} and f replaced by \hat{m}_j and \hat{f}_j , and by (5.5), the proof of Lemma 4 will be complete when it is shown that

$$(9.1) \quad \sup_h \left| \frac{n^{-1} \sum_{j=1}^n (\hat{m}_j(X_j) - m(X_j)) \hat{f}_j(X_j) (Y_j - m(X_j)) f(X_j)^{-1} w(X_j)}{d_M^*(\hat{m}_h, m)} \right| \rightarrow 0 \text{ a.s.}$$

The numerator of (9.1) may be written as

$$n^{-2} \sum_{i \neq j} U_{i,j} + n^{-2} \sum_{i \neq j} V_{i,j},$$

where

$$U_{i,j} = \left(\frac{n}{n-1} \right) \frac{1}{h^d} K \left(\frac{X_j - X_i}{h} \right) (Y_i - m(X_i)) (Y_j - m(X_j)) f(X_j)^{-1} w(X_j),$$

$$V_{i,j} = \left(\frac{n}{n-1} \right) \frac{1}{h^d} K \left(\frac{X_j - X_i}{h} \right) (m(X_i) - m(X_j)) (Y_j - m(X_j)) f(X_j)^{-1} w(X_j).$$

Hence (9.1) and the Lemma 4 will be established when it is shown that

$$(9.2) \quad \sup_h \left| \frac{n^{-2} \sum_{i \neq j} U_{i,j}}{d_M^*(\hat{m}_h, m)} \right| \rightarrow 0 \text{ a.s.},$$

$$(9.3) \quad \sup_h \left| \frac{n^{-2} \sum_{i \neq j} V_{i,j}}{d_M^*(\hat{m}_h, m)} \right| \rightarrow 0 \text{ a.s.}$$

To verify (9.2), note that by Hölder-continuity considerations, it is enough to show that, for H'_n as above,

$$\sup_{h \in H'_n} \left| \frac{n^{-2} \sum_{i \neq j} U_{i,j}}{d_M^*(\hat{m}_h, m)} \right| \rightarrow 0 \text{ a.s.}$$

For this, note that given $\varepsilon > 0$, $k = 1, 2, \dots$

$$P \left[\sup_{h \in H'_n} \left| \frac{n^{-2} \sum_{i \neq j} U_{i,j}}{d_M^*(\hat{m}_h, m)} \right| > \varepsilon \right] \leq \varepsilon^{-2k} \#(H'_n) \sup_{h \in H'_n} E \left[\frac{n^{-2} \sum_{i \neq j} U_{i,j}}{d_M^*(m_h, m)} \right]^{2k},$$

so that the proof of (9.2) will be complete when it is seen that there is a constant $\tau > 0$, so that for $k = 1, 2, \dots$, there are constants C_k so that

$$(9.4) \quad \sup_h E \left[\frac{n^{-2} \sum_{i \neq j} U_{i,j}}{d_M^*(\hat{m}_h, m)} \right]^{2k} \leq C_k n^{-\tau k},$$

Similarly (9.3) will be verified by showing that

$$(9.5) \quad \sup_h E \left[\frac{n^{-2} \sum_{i \neq j} V_{i,j}}{d_M^*(\hat{m}_h, m)} \right]^{2k} \leq C_k n^{-\tau k}.$$

To check (9.4), for $i, j = 1, \dots, n$ define

$$(9.6) \quad \begin{aligned} Z_i &= Y_i - m(X_i), \\ \alpha_{ij} &= (n-1)^{-1} h^{-d} K\left(\frac{X_j - X_i}{h}\right) f^{-1}(X_j) w(X_j) 1_{(i \neq j)}. \end{aligned}$$

In the following, C will denote a generic constant which may depend on k and may take on different values even in the same formula. From Theorem 2 of Whittle (1960) and (A.4), it follows that

$$\begin{aligned} E \left[\left(n^{-1} \sum_{i \neq j} U_{i,j} \right)^{2k} \mid X_1, \dots, X_n \right] &= E \left[\left(\sum_{i,j} \alpha_{ij} Z_i Z_j \right)^{2k} \mid X_1, \dots, X_n \right] \\ &\leq C \left(\sum_{i,j} \alpha_{ij}^2 \right)^k. \end{aligned}$$

Thus, by (A.5) and integration by substitution,

$$\begin{aligned} E \left[n^{-1} \sum_{i \neq j} U_{i,j} \right]^{2k} &\leq CE \left[(n-1)^{-2} \sum_{i \neq j} h^{-2d} K\left(\frac{X_i - X_j}{h}\right)^2 \right]^k \\ &\leq Cn^{-2k} h^{-2dk} \sum_{l=2}^{2k} n^l h^{dl/2} \leq Ch^{-dk}. \end{aligned}$$

The inequality (9.4) follows easily from this and (8.5).

To check (9.5), in addition to the notation (9.6), define

$$b_j = (n-1)^{-1} \sum_{i=1}^n h^{-d} K\left(\frac{X_j - X_i}{h}\right) (m(X_i) - m(X_j)) f(X_j)^{-1} w(X_j) 1_{(i \neq j)}.$$

Again using Theorem 2 of Whittle (1960) and (A.4),

$$\begin{aligned} E \left[\left(n^{-1} \sum_{i \neq j} V_{i,j} \right)^{2k} \mid X_1, \dots, X_n \right] &= E \left[\left(\sum_{j=1}^n b_j Z_j \right)^{2k} \mid X_1, \dots, X_n \right] \\ &\leq C \left(\sum_{j=1}^n b_j^2 \right)^k. \end{aligned}$$

Nonparametric Sequential Estimation of Zeros and Extrema of Regression Functions

WOLFGANG K. HÄRDLE AND RAINER NIXDORF

Abstract—Let $(X, Y), (X_1, Y_1), (X_2, Y_2), \dots$ be independent identically distributed pairs of random variables, and let $m(x) = E(Y|X=x)$ be the regression curve of Y on X . The estimation of zeros and extrema of the regression curve via stochastic approximation methods is considered. Consistency results of some sequential procedures are presented and termination rules are defined providing fixed width confidence intervals for the parameters to be estimated.

I. INTRODUCTION

LET $(X, Y), (X_1, Y_1), (X_2, Y_2), \dots$ be a sequence of independent identically distributed (i.i.d.) bivariate random variables with joint probability density function $f(x, y)$. In this paper we consider the sequential estimation of zeros and extrema of $m(x) = E(Y|X=x)$ using a combination of the nonparametric kernel and stochastic approximation methods. The structure of our sampling scheme is different from the one considered by Robbins and Monro [8] since the experimenter observing the bivariate data has no control over the design variables $\{X_i\}$, as is assumed in classical stochastic approximation algorithms.

The proposed sequential procedure is based on the principal idea of non-parametric kernel estimation of $m(x)$, i.e., to construct a weighted average of those observations (X_i, Y_i) whose X_i fall into an asymptotically shrinking neighborhood of x . The shrinkage of such a neighborhood is usually parameterized by a sequence of *bandwidths* h_n tending to zero, whereas the shape of the neighborhoods is given by a real *kernel* function K .

Motivated by classical procedures, we define the following sequential estimator of a zero of m ,

$$Z_{n+1} = Z_n - a_n h_n^{-1} K((Z_n - X_n)/h_n) Y_n, \quad n \geq 1. \quad (1)$$

Here Z_1 denotes an arbitrary starting random variable with finite second moment, and $\{a_n\}$ is a sequence of positive constants tending to zero. In fact, the sequence $\{Z_n\}$ will converge under our conditions to the (unique)

zero of

$$\tilde{m}(x) = \int y f(x, y) dy = m(x) f'_X(x),$$

where $f_X(x)$ denotes the marginal density of X , but an assumption about f_X ensures that the zero of the two functions m and \tilde{m} is identical.

Under mild conditions we show consistency (almost surely and in quadratic mean) and asymptotic normality of $\{Z_n\}$. An asymptotic bias term (depending on the smoothness of m) shows up if the bandwidth sequence tends to zero at a specific rate. Fixed width confidence intervals are constructed using a suitable stopping rule based on estimates of the variance of the asymptotic normal distribution.

Our arguments can be extended to the problem of estimating extremal values of the regression function m . Note that $m = \tilde{m}/f_X$ and therefore $m' = \tilde{r}/f_X^2$, where

$$\tilde{r}(x) = f_X(x) \int y \frac{\partial}{\partial x} f(x, y) dy - \tilde{m}(x) f'_X(x).$$

Under a suitable assumption the problem of finding an extremum of m is equivalent to finding a (unique) zero of the function \tilde{r} . So it is reasonable to apply a procedure similar to (1). Additional difficulties turn up since f_X has to be estimated separately. We propose to perform the estimation by an additional i.i.d. sequence $\{\tilde{X}_i\}$ with the same distribution as X . Define

$$\begin{aligned} Z'_{n+1} = & Z'_n - a_n h_n^{-3} Y_n \{ K((Z'_n - \bar{X}_n)/h_n) \\ & \cdot K'((Z'_n - X_n)/h_n) \\ & K'((Z'_n - \bar{X}_n)/h_n) K((Z'_n - X_n)/h_n) \}, \\ & n \geq 1. \quad (2) \end{aligned}$$

We shall prove that $\{Z'_n\}$ is consistent and asymptotically normally distributed. Fixed width confidence intervals are computed by the same technique as for $\{Z_n\}$.

When we know f_X , the algorithm (2) can be simplified as follows, since the additional $\{\tilde{X}_i\}$ are obsolete in this case:

$$\begin{aligned} Z'_{n+1} = & Z'_n - a_n h_n^{-2} Y_n \{ K'((Z'_n - X_n)/h_n) f_X(Z'_n) \\ & K'((Z'_n - X_n)/h_n) f'_X(Z'_n) \}, \quad n \geq 1. \quad (3) \end{aligned}$$

The additional difficulty of estimating simultaneously f_X did not occur in the case of estimating zeros, since the problem for m could be transferred to the equivalent problem for \tilde{m} , which does not involve f_X . In practice the

Manuscript received April 12, 1985; revised May 12, 1986. This work was supported in part by the "Deutsche Forschungsgemeinschaft," SFB 123, "Stochastische Mathematische Modelle" and in part by AFOSR Grant 49620 82 C0009.

W. K. Härdle was with Johann Wolfgang Goethe Universität, Fachbereich Mathematik, D 6000 Frankfurt/Main. He is now with Institut für Wirtschaftstheorie II, Universität Bonn, Adenauerallee 24-26, D-5300 Bonn 1, Germany.

R. Nixdorf is with IBM, Entwicklung und Forschung, Schönaicher Str. 220, D-7030 Böblingen, Germany.

IEEE Log Number 8611424.

additional i.i.d. sequence $\{\bar{X}_i\}$ could be constructed by sampling in pairs and discarding the Y observations of one element. This results in some loss of efficiency but makes the practical application possible with the data at hand. Another proposal that we would like to make is related to the bootstrap. From the first N observations, a density estimate \hat{f}_X of f_X could be constructed and then the algorithm (2) could be started with $\{\bar{X}_i\}$ distributed with density \hat{f}_X . A third possibility would be to plug \hat{f}_X into the algorithm (3). We have not investigated the last two procedures, because they do not seem to be more efficient.

An alternative way of defining an estimator of the zero of the regression function m would be to construct an estimate of the whole function and then to use a zero of the function estimate as an estimator for the zero of the regression function (Müller [5]). This procedure would be time consuming in the case of sequential observation of the data, since for every new observation the whole function would have to be constructed, whereas our procedure just keeps one number in memory and updates that number via the formal prescription (1). Also in cases where an enormous amount of data has to be processed, an estimate of a zero based on the estimate of the whole regression function seems to be inadvisable since all the data has to be stored in the memory at any time.

Related work was done by Revesz [7] and Rutkowski [9], [10] who applied stochastic approximation methods to the estimation of m at a fixed point. Our derivation of fixed width confidence intervals was inspired by the papers of Chow and Robbins [2], McLeish [4], and Stute [12]. Stute used the kernel estimation technique that introduces a localizing effect and makes classical methods, such as Venter's [13], applicable.

The rest of the paper is organized as follows. Section II contains the results and gives the consistency proof for $\{Z_n\}$. In Section III we present the results of some simulations and an application of $\{Z_n\}$ to some real data. In the last section we give the rest of the proofs.

II. RESULTS

We first describe various conditions on the functions and parameters of the algorithms which are used in the sequel. We split up the assumptions into several parts since these will be used separately. A crucial assumption that makes the problem identifiable through \tilde{m} [resp. \tilde{f}] is the following.

A1) f_X is positive.

The speed of convergence of $\{a_n\}$ and $\{h_n\}$ is controlled by

- A2.1) $\sum_{n=1}^{\infty} a_n = \infty, \sum_{n=1}^{\infty} a_n h_n < \infty,$
- A2.2) $\sum_{n=1}^{\infty} a_n^2 h_n^{-2} < \infty,$ and
- A2.3) $\sum_{n=1}^{\infty} a_n^2 h_n^{-4} < \infty.$

The zero θ_0 of $m(x)$ (and of $\tilde{m}(x)$) is identified by

A3) $\inf_{\epsilon \leq |x - \theta_0| \leq 1/\epsilon} (x - \theta_0) \tilde{m}(x) > 0$ for all $\epsilon > 0$.

The smoothness of \tilde{m} is described by

- A4.1) \tilde{m} is Lipschitz continuous;
- A4.2) \tilde{m} is differentiable at θ_0 with $\tilde{m}'(\theta_0) > (1 - \gamma)/2, 1/5 \leq \gamma < 1/2;$
- A4.3) \tilde{m} is twice continuously differentiable, with bounded second derivative.

The kernel function K has to satisfy the following conditions.

A5.1) K is bounded and

$$\int K(u) du = 1 \quad \int uK(u) du = 0 \quad \int u^2 K(u) du < \infty.$$

A5.2) K is differentiable with bounded derivative K' and

$$\lim_{|u| \rightarrow \infty} |uK(u)| = 0 \quad \int |u|K'(u) du < \infty.$$

A5.3) K is twice differentiable and

$$\lim_{|u| \rightarrow \infty} |uK'(u)| = 0 \quad \int |u|K''(u) du < \infty.$$

The joint density $f(x, y)$ has to be smooth in its first argument.

A6.1) $|f(x, y) - f(z, y)| \leq |x - z|g_1(y)$ such that $\int (y^2 + 1)g_1(y) dy < \infty.$

A6.2) $(\partial^2/\partial x^2)f(x, y)$ is continuous and

$$\left| \frac{\partial}{\partial x} f(u, y) - \frac{\partial}{\partial x} f(v, y) \right| \leq |u - v|g_2(y)$$

with $\int (|y| + 1)g_2(y) dy < \infty.$

Moment assumptions are

- A7) $EY^2 < \infty,$
- A8) $EY^4 < \infty,$ and
- A9) $\sup_{x \in \mathbb{R}} E(Y^2|X = x) < \infty.$

The consistency of $\{Z_n\}$ is shown in the following theorem.

Theorem 1: Assume A1), A2.1), A2.2), A3), A4.1), A5.1), and A7). Then $\{Z_n\}$ converges to θ_0 almost surely (a.s.) and in quadratic mean.

Since the proof of this theorem is very simple and exemplifies the combination of the kernel method together with stochastic approximation arguments we would like to give it here. The proofs of the following results are delayed to Section IV.

Proof: Write

$$Z_{n+1} = Z_n - a_n \tilde{m}(Z_n) + a_n V_n$$

$$V_n = \tilde{m}(Z_n) - K_h(Z_n - X_n) Y_n$$

where

$$K_h(u) = h_n^{-1} K(u/h_n).$$

Let $\mathcal{G}_n = \sigma\{Z_1, Z_2, \dots, Z_n\}$. Condition A4.1) implies that $E(V_n|\mathcal{G}_n) = O(h_n)$ a.s.

$$E(V_n^2) = O(E(Z_n - \theta_0)^2) + O(h_n^{-2}).$$

Observe that with A3) and a Lipschitz constant $L_{\tilde{m}}$ we have

$$\begin{aligned} (Z_{n+1} - \theta_0)^2 &= (Z_n - \theta_0)^2 - 2a_n \tilde{m}(Z_n)(Z_n - \theta_0) \\ &\quad + a_n^2 \tilde{m}^2(Z_n) + 2a_n V_n (Z_n - \theta_0 - a_n \tilde{m}(Z_n)) \\ &\quad + a_n^2 V_n^2 \\ &\leq (1 + a_n^2 L_{\tilde{m}}^2)(Z_n - \theta_0)^2 + a_n^2 V_n^2 \\ &\quad + 2a_n V_n (Z_n - \theta_0 - a_n \tilde{m}(Z_n)). \end{aligned}$$

Hence by A7),

$$\begin{aligned} E(Z_{n+1} - \theta_0)^2 &\leq (1 + a_n^2 L_{\tilde{m}}^2)E(Z_n - \theta_0)^2 \\ &\quad + O(h_n) a_n (1 + a_n L_{\tilde{m}}) E|Z_n - \theta_0| \\ &\quad + a_n^2 E(V_n^2) \\ &\leq (1 + \beta_n)E(Z_n - \theta_0)^2 + \delta_n \end{aligned}$$

where

$$\begin{aligned} \beta_n &= O(h_n^{-2} a_n^2 + h_n a_n + a_n^2), \\ \delta_n &= O(h_n a_n + h_n^{-2} a_n^2). \end{aligned}$$

Note that by A2.1) and A2.2), $\sum \beta_n < \infty$ and $\sum \delta_n < \infty$.

This implies that the sequence $E(Z_n - \theta_0)^2$ is bounded so that with A2.2) and A2.1), $\sum a_n^2 E V_n^2 < \infty$ and $\sum a_n |E(V_n|\mathcal{G}_n)| < \infty$ a.s.

The assertion follows now from Venter [13, theorem 1]. Nixdorf [6, theorem 1.1.2] has given a corrected version of Venter's theorem. The asymptotic normality is shown in the following theorem.

Theorem 2: Assume A1), A3), A4.2), A4.3), A5.1), A6.1), and A8). Let $a_n = n^{-1}$, $h_n = n^{-\gamma}$, $1/5 \leq \gamma < 1/2$. Then

$$n^{(1-\gamma)/2} \{Z_n - \theta_0\} \rightarrow \mathcal{L} N(b(\gamma), \sigma^2(\gamma))$$

where

$$\begin{aligned} b(\gamma) &= 0, \quad \text{if } 1/5 < \gamma < 1/2 \\ &= \tilde{m}''(\theta_0) \int u^2 K(u) du / (2\tilde{m}'(\theta_0) - 1 + \gamma), \\ &\quad \text{if } \gamma = 1/5; \end{aligned}$$

$$\sigma^2(\gamma) = \int K^2 \int y^2 f(\theta_0, y) dy / (2\tilde{m}'(\theta_0) - 1 + \gamma).$$

Fixed width asymptotic confidence intervals for the unknown parameter θ_0 are constructed via estimators of the asymptotic bias $b(\gamma)$ and variance $\sigma^2(\gamma)$. Estimators of $\int y^2 f(\theta_0, y) dy$, $\tilde{m}'(\theta_0)$, $\tilde{m}''(\theta_0)$ are, respectively,

$$S_{1n} = n^{-1} \sum_{i=1}^n K_{h_i}(Z_i - X_i) Y_i^2 \quad (4)$$

$$S_{2n} = n^{-1} \sum_{i=1}^n K_{h_i}'(Z_i - X_i) Y_i \quad (5)$$

$$S_{3n} = n^{-1} \sum_{i=1}^n K_{h_i}''(Z_i - X_i) Y_i.$$

An estimator for the asymptotic variance $\sigma^2(\gamma)$ is therefore

$$\begin{aligned} s_n &= \int K^2 S_{1n} / (2S_{2n} - 1 + \gamma), \quad \text{if } 2S_{2n} - 1 + \gamma > 0, \\ &= 1, \quad \text{otherwise.} \end{aligned}$$

So the following stopping rule seems reasonable.

$$N(d) = \inf \{n \in \mathbb{N} | s_n + n^{-1} \leq n^{1-\gamma} d^2 / z_{\alpha/2}^2\} \quad (6)$$

where $z_{\alpha/2}$ is the $(1 - \alpha/2)$ -quantile of the standard normal distribution. The fixed width confidence intervals are constructed via the following theorem.

Theorem 3: Let $a_n = n^{-1}$, $h_n = n^{-\gamma}$, $1/5 \leq \gamma < 1/3$ and assume A1), A3), A4.2), A4.3), A5.1), A5.2), A5.3), A6.1), and A8). Then if $N(d)$ is defined as in (6) for some $0 < \alpha < 1$, as $d \rightarrow 0$,

$$N(d)^{(1-\gamma)/2} \{Z_{N(d)} - \theta_0\} \rightarrow \mathcal{L} N(b(\gamma), \sigma^2(\gamma)).$$

When $1/5 < \gamma < 1/3$ an asymptotic confidence interval of fixed length $2d$ and asymptotic coverage probability $1 - \alpha$ is given by

$$[Z_{N(d)} - d, Z_{N(d)} + d].$$

For $\gamma = 1/5$ the bias can be estimated by

$$b_n = \int u^2 K(u) du S_{3n} / (2S_{2n} - 1 + \gamma).$$

Then with $H_n = Z_n - n^{-(1+\gamma)/2} b_n$ an asymptotic confidence interval is given by $\{H_{N(d)} - d, H_{N(d)} + d\}$.

In Theorem 3 the range of γ had to be reduced to $1/5 \leq \gamma < 1/3$ from that of Theorem 2 since otherwise S_{2n} would no longer be a consistent estimator of $\tilde{m}'(\theta_0)$.

It will be seen in the proof of Theorem 3 that, as $d \rightarrow 0$, $N(d)/l(d) \rightarrow 1$ almost surely where $l(d) = \inf \{n \in \mathbb{N} | \sigma^2(\gamma) \leq n^{1-\gamma} d^2 / z_{\alpha/2}^2\}$. Therefore, $N(d)$ exhibits the following limit behavior, as $d \rightarrow 0$,

$$d^{2/(1-\gamma)} N(d) \rightarrow (\sigma^2(\gamma))^{1/(1-\gamma)} z_{\alpha/2}^{2/(1-\gamma)}.$$

The analysis of the sequential procedure $\{Z_n\}$ is quite analogous to that of $\{Z'_n\}$. We define the (unique) zero of \tilde{r} as θ_M .

Theorem 4: Assume A1), A2.1), A2.3), A5.1), A5.2), A6.1), A8), and A9), and let A3) and A4.1) be fulfilled with \tilde{r} in the place of \tilde{m} . Then $\{Z'_n\}$ converges to θ_M almost surely and in the quadratic mean.

Theorem 5: Let $a_n = n^{-1}$ and $h_n = n^{-\gamma}$, $1/6 < \gamma < 1/4$ and assume A1), A5.1), A5.2), A5.3), A6.1), A6.2), A8), A9), and A3), $\tilde{r}'(\theta_M) > (1 - 4\gamma)/2$ with \tilde{r} in the place of \tilde{m} . Then

$$n^{(1-4\gamma)/2} \{Z'_n - \theta_M\} \rightarrow \mathcal{L} N(0, \sigma_M^2(\gamma))$$

where

$$\begin{aligned} \sigma_M^2(\gamma) &= f_X(\theta_M) \int y^2 f(\theta_M, y) dy \\ &\quad \int K^2 \int K'^2 / (2\tilde{r}'(\theta_M) - 1 + 4\gamma). \end{aligned}$$

For simplicity of presentation we did not allow for a wider range of γ under which an asymptotic bias term would occur. If \tilde{F} is twice continuously differentiable then the range of allowable exponents can be extended to $1/8 \leq \gamma < 1/4$. The discussion would be in analogy to Theorem 2 with \tilde{F} in the place of \tilde{m} . The rate of convergence of $\{Z_n\}$ is for $\gamma = 1/5$ equal to $n^{-2/5}$. This rate is typical for nonparametric smoothing problems, as Stone [11] has shown. Under stronger assumptions Müller [5] also achieved this rate. Major and Revesz [3] considered the classical Robbins-Monro algorithm in the situation when the derivative of the regression function gets close to zero. They showed that in this case a different rate of convergence is obtained. We believe that similar arguments should be applicable in our setting.

Estimators for the numerator and denominator of $\sigma_M^2(\gamma)$ are constructed in the following way:

$$S'_{1n} = n^{-1} \sum_{i=1}^n K_{h_i}(Z'_i - \bar{X}_i) n^{-1} \sum_{k=1}^n K_{h_k}(Z'_k - X_k) Y_k^2$$

is an estimator for $\int f_X(\theta_M) f_Y^2(\theta_M, y) dy$, whereas

$$S'_{2n} = n^{-1} \sum_{i=1}^n K_{h_i}(Z'_i - \bar{X}_i) n^{-1} \sum_{k=1}^n K'_{h_k}(Z'_k - X_k) Y_k - n^{-1} \sum_{i=1}^n K_{h_i}(Z'_i - X_i) Y_i n^{-1} \sum_{k=1}^n K'_{h_k}(Z'_k - \bar{X}_k)$$

converges under our assumptions to $\tilde{F}'(\theta_M)$, almost surely. Define

$$s_{n,M} = \int K^2 \int K'^2 S'_{1n} / (2S'_{2n} - 1 + 4\gamma)$$

$$I(d) = \inf \left\{ n \in \mathbb{N} \mid s_{n,M} + n^{-1} \leq n^{1-2\gamma} d^2 / z_{\alpha/2}^2 \right\}.$$

Then parallel to Theorem 3 we have the following.

Theorem 6: Let $a_n = n^{-1}$ and $h_n = n^{-\gamma}$, $1/6 < \gamma < 1/5$, and let the conditions of Theorem 5 be fulfilled. Then, as $d \rightarrow 0$,

$$I(d)^{1/(2-2\gamma)} \{ Z'_{I(d)} - \theta_M \} \rightarrow \mathcal{L} N(0, \sigma_M^2(\gamma)).$$

III. MONTE CARLO STUDY AND AN APPLICATION

In this section we report the results of a Monte Carlo experiment comparing the performance of our sequential procedure when some of the involved parameters are tuned at different levels. We also report an application of the algorithm (1) to some real data.

The basic experiment to assess the accuracy of Theorem 3 consisted of 200 Monte Carlo replications with the numbers $N(d)$, $Z_{N(d)}$ and $S_{N(d)}$ to be reported. The joint probability density function $f(x, y)$ that we used was $f(x, y) = I_{[0,1]}(x) \sigma^{-1} \varphi((y - m(x))/\sigma)$, φ was the probability density function of a standard normal distribution, and $m(x) = -a\{(1-x)^2 - 1/4\}$ for $a = 4, 8$ was the regression curve. We report the result for $Z_1 = 0.45$ (Table I) and for $Z_1 = 0.2$ (Table II). The parameter α was set to $\alpha = 0.05$. The zero that was to be estimated was $\theta_0 = 1/2$ and two different values of d and σ were fixed, namely $d = 0.05, 0.1$ and $\sigma = 0.1, 1.0$. As the kernel K we have chosen the Epanechnikov kernel $K(u) = (3/4)(1 - u^2)$ for $|u| \leq 1$ and $K(u) = 0$ for $|u| > 1$. The sequence of bandwidths was set to $h = h_n = n^{-\gamma}$, $\gamma = 0.21$. In Table I the results for the starting point $Z_1 = 0.45$ are shown. The numerical values of Table I indicate that the fixed accuracy result given in Theorem 3 yields a good approximation of θ_0 even for $d = 0.1$. This is seen from the counts in the $Z_{N(d)}$ column. It is indicated there how many times (from 200 Monte Carlo trials) the true parameter $\theta_0 = 1/2$ was in the confidence interval $[Z_{N(d)} - d, Z_{N(d)} + d]$. As

TABLE I

σ	d	$N(d)$				$Z_{N(d)}$				$D_{N(d)}$				$m'(\theta)$	
		Mean	Standard	Q_5	Q_{95}	Mean	Standard	Q_5	Q_{95}	Counts	Mean	Standard	Q_5		Q_{95}
0.1	0.05	137	10	120	156	0.518	0.021	0.483	0.553	188	0.024	0.002	0.02	0.028	4
0.1	0.05	229	17	200	259	0.515	0.019	0.483	0.547	194	0.043	0.003	0.037	0.048	8
0.1	0.1	43	3.5	38	50	0.519	0.03	0.469	0.574	199	0.027	0.005	0.018	0.035	4
0.1	0.1	62	6.7	51	74	0.517	0.039	0.449	0.582	196	0.05	0.007	0.038	0.062	8
1	0.05	642	97	496	806	0.51	0.025	0.467	0.548	188	0.105	0.013	0.127	0.084	4
1	0.05	469	46	394	551	0.515	0.023	0.475	0.559	181	0.081	0.006	0.07	0.093	8
1	0.1	129	37	76	207	0.52	0.054	0.429	0.61	184	0.11	0.028	0.065	0.168	4
1	0.1	103	18	72	133	0.525	0.043	0.461	0.596	192	0.09	0.015	0.062	0.114	8

$Z_0 = 0.45, \gamma = 0.21, \alpha = 0.05$.

TABLE II

σ	d	$N(d)$				$Z_{N(d)}$				$D_{N(d)}$				$m'(\theta)$	
		Mean	Standard	Q_5	Q_{95}	Mean	Standard	Q_5	Q_{95}	Counts	Mean	Standard	Q_5		Q_{95}
0.1	0.05	163	12	141	183	0.517	0.018	0.485	0.547	192	0.03	0.002	0.025	0.034	4
0.1	0.05	255	17	227	283	0.518	0.019	0.484	0.552	189	0.04	0.003	0.042	0.052	8
0.1	0.1	56	7	46	70	0.513	0.027	0.464	0.561	199	0.04	0.007	0.031	0.059	4
0.1	0.1	74	8	61	90	0.515	0.036	0.456	0.593	196	0.064	0.007	0.05	0.077	8
1	0.05	646	83	518	802	0.516	0.026	0.471	0.562	174	0.105	0.011	0.088	0.126	4
1	0.05	479	44	417	550	0.512	0.023	0.475	0.55	188	0.082	0.006	0.073	0.093	8
1	0.1	139	31	89	192	0.515	0.044	0.437	0.595	192	0.118	0.024	0.077	0.158	4
1	0.1	118	19	86	146	0.522	0.042	0.454	0.59	193	0.102	0.015	0.074	0.125	8

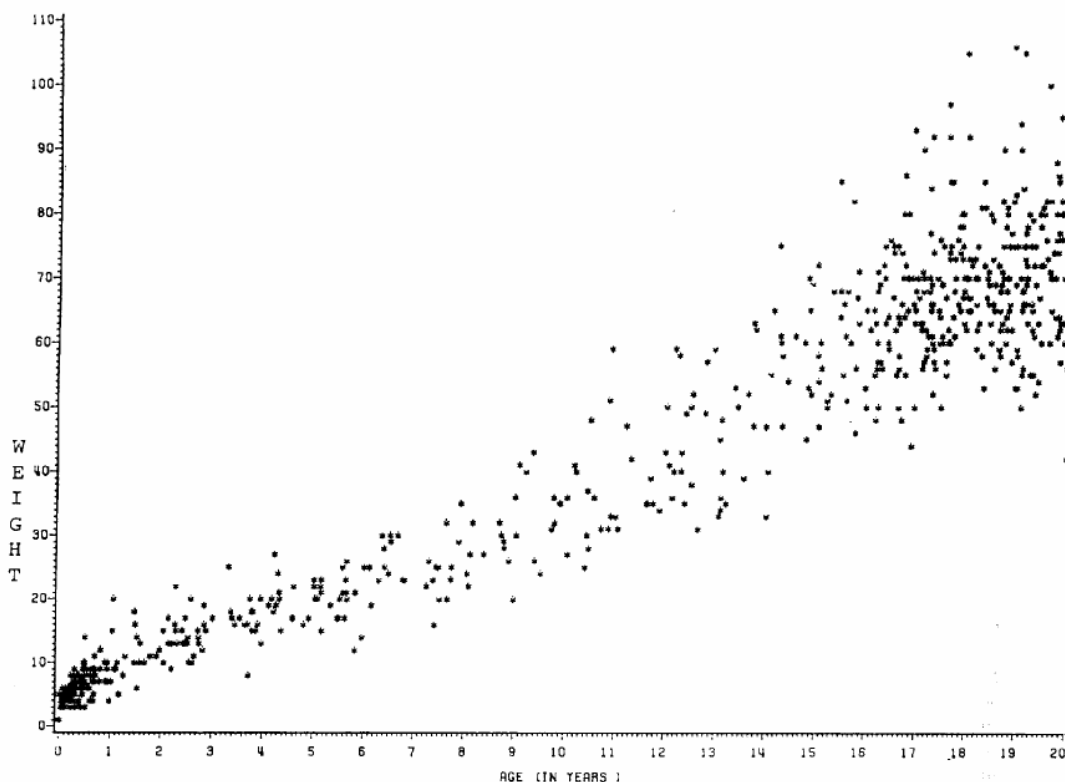


Fig. 1. Bodyweight (in kilograms) versus age (in years) of 732 female persons.

a measure of spread we added the quantiles Q_{95} and Q_5 in the third and fourth column of each entry. A small paradox occurs when we compare the values for different values of a . It is expected that the procedure (1) stops earlier with $a = 8$ than with $a = 4$, since the higher derivative in the zero should speed up the convergence of $\{Z_n\}$ to θ_0 . In both Tables I and II it is seen that the average of the stopping times (over 200 Monte Carlo runs) is considerably higher for $a = 8$ and $\sigma = 0.1$ than for $a = 4$ and $\sigma = 0.1$. This effect is due to the crude approximation $\text{var}(Y|X = x) \approx \sigma^2, x \approx \theta_0$, as can be seen from the values for $S_{N(d)}$. In the case of $a = 8$ the statistic $S_{N(d)}$ considerably overestimates the true asymptotic variance $\sigma(\gamma)$. For comparison we list some correct $\sigma(\gamma) = \sigma(\sigma, a, \gamma)$. For instance, $\sigma(0.1, 4, 0.21) = 0.00083$ whereas $\sigma(0.1, 8, 0.21) = 0.00039$.

In an application we took the sequence of random variables $X_i = \text{age}$, and $Y_i = \text{weight}$ of female corpses which was gathered from 1969 to 1981 by the Institute of Forensic Medicine of Heidelberg. It is an interesting question in forensic medicine to estimate the mean age from the weight of unknown corpses. We restricted our attention to the ages between 0 and 20 years to fulfill assumption A3).

We put $m_0 = 40$ kg, and we applied the procedure (1) and ended with different starting values Z_1 at $Z_{N(d)} = 11.6$ years and $N(d) = 563$, for $d = 0.1$ and $N(d) = 224$ for $d = 0.2 (Z_1 = 0.4)$. A plot of the first 732 data pairs, restricted to ages between 0 and 20 years, should illustrate the accuracy of $Z_{N(d)}$ (Fig. 1).

IV. PROOFS

The Theorems are proved by a functional central limit theorem given by Berger [1], who extended a result of Walk [14], that made it applicable in our setting. Lemma 1 describes the asymptotic behavior of

$$W_n(t) = n^{-1/2}R_{[nt]} + n^{-1/2}(nt - [nt]) \cdot \{R_{[nt]+1} - R_{[nt]}\}, \quad 0 \leq t \leq 1, \quad (7)$$

where

$$R_k = k^{1/2} [k^{(1-\gamma)/2} (Z_{k+1} - \theta_0) - b(\gamma)], \quad k \in \mathbb{N}.$$

Lemma 1: Let the conditions of Theorem 3 be satisfied. Then $W_n(t)$ as defined in (7) converges weakly in $C[0, 1]$ to

the Gaussian process

$$G_1(t) = \left(\int K^2 \int y^2 f(\theta_0, y) dy \right)^{1/2} \cdot \int_{(0,1)} (u/t)^{\tilde{m}'(\theta_0) - (2-\gamma)/2} dW(u), \quad (8)$$

$0 \leq t \leq 1$, where W is the standard Wiener process starting at 0.

Proof: Define $\mathcal{B}_n = \sigma\{Z_1, \dots, Z_n\}$ and write

$$Z_{n+1} - \theta_0 = (1 - B_n/n)(Z_n - \theta_0) + n^{-(2-\gamma)/2} \tilde{V}_n + n^{-(2-\gamma)/2} T_n$$

where

$$\begin{aligned} \tilde{V}_n &= h_n^{-1/2} E \left\{ K \left(\frac{Z_n - X_n}{h_n} \right) Y_n | \mathcal{B}_n \right\} \\ &\quad - h_n^{-1/2} K \left(\frac{Z_n - X_n}{h_n} \right) Y_n \\ T_n &= n^{(1-\gamma)/2} \{ \tilde{m}(Z_n) - E [K_h(Z_n - X_n) Y_n | \mathcal{B}_n] \} \end{aligned}$$

and $\{B_n\}$ is a sequence of random variables converging almost surely to $\tilde{m}'(\theta_0)$ such that $B_n(Z_n - \theta_0) = \tilde{m}(Z_n)$. Such a sequence exists because \tilde{m} is differentiable in θ_0 and $Z_n \rightarrow \theta_0$ almost surely by Theorem 1. The assumption on a_n and h_n imply that $T_n \rightarrow (1/2) \int u^2 K(u) du \tilde{m}''(\theta_0)$. Note that $E(\tilde{V}_n | \mathcal{B}_n) = 0$ and that by (A7) and (A6.1),

$$\begin{aligned} E(\tilde{V}_n^2 | \mathcal{B}_n) &\rightarrow \int K^2 \int y^2 f(\theta_0, y) dy \text{ a.s.} \\ E(\tilde{V}_n^2) &= O(1). \end{aligned}$$

Furthermore we have for all $s > 0$,

$$\begin{aligned} E(\tilde{V}_n^2 I(\tilde{V}_n^2 \geq sn) | \mathcal{B}_n) &\leq O(h^{-2}) P(\tilde{V}_n^2 \geq sn | \mathcal{B}_n) \\ &\leq O(h^{-2} n^{-1}) = o(1) \text{ a.s.} \end{aligned}$$

The lemma follows now from the generalization of a theorem of Walk [14], given by Berger [2]; see also Nixdorf [6].

The following lemma gives an analogous result for the Kiefer-Wolfowitz type sequence $\{Z'_n\}$ defined in (2).

Lemma 2: Let the conditions of Theorem 5 be satisfied. Define $W_n(t)$ as in Lemma 1 and

$$R_k = k^{1/2} k^{(1-4\gamma)/2} (Z'_{k+1} - \theta_M).$$

Then $W_n(t)$ converges weakly in $C[0,1]$ to the Gaussian process

$$G_2(t) = \left\{ f_X(\theta_M) \int y^2 f(\theta_M, y) dy \int K^2 \int K'^2 \right\}^{1/2} \cdot \int_{(0,1)} (u/t)^{\tilde{f}'(\theta_M) - (1-\gamma)/2} dW(u), \quad 0 \leq t \leq 1.$$

Proof of Theorem 2: Use Lemma 1 and evaluate $G_1(t)$ at $t = 1$.

Proof of Theorem 3: The estimators S_{1n}, S_{2n} defined in (4), (5) converge to $\int y^2 f(\theta_0, y) dy, \tilde{m}'(\theta_0)$, respectively. This entails that, as $d \rightarrow 0$,

$$N(d)/l(d) \rightarrow 1 \text{ a.s.}$$

Now apply Lemma 1.

Proof of Theorem 4: Proceed as with the proof of Theorem 1.

Proof of Theorem 5: Use Lemma 2 and evaluate $G_2(t)$ at $t = 1$.

Proof of Theorem 6: Proceed as with the proof of Theorem 3.

ACKNOWLEDGMENT

We would like to thank Laszlo Györfi for helpful suggestions which led to a substantial improvement of the paper.

REFERENCES

- [1] Berger, "Asymptotic behavior of a class of stochastic approximation procedures," *Probability Theory and Related Fields*, 1986, to appear.
- [2] Y. S. Chow and H. Robbins, "On the asymptotic theory of fixed-width sequential confidence intervals for the mean," *Ann. Math. Statist.*, vol. 36, pp. 457-462, 1965.
- [3] P. Major and P. Révész, "A limit theorem for the Robbins-Monro approximation," *Z. Wahrscheinlichkeitstheorie*, u. V. G. 27, pp. 79-86, 1973.
- [4] R. McLeish, "Functional and random central limit theorems for the Robbins-Monro process," *J. Appl. Prob.*, vol. 13, pp. 148-154, 1976.
- [5] H. G. Müller, "Kernel estimator of zeros and of location and size of extrema of regression functions," *Scand. J. Statist.*, vol. 12, pp. 221-232, 1985.
- [6] R. Nixdorf, "Stochastische Approximation in Hilberträumen durch endlichdimensionale Verfahren," *Mitt. Math. Sem. Gießen*, Heft 154, 1982.
- [7] P. Révész, "How to apply the method of stochastic approximation in the non-parametric estimation of a regression function," *Math. Oper. Statist.*, vol. 8, pp. 119-126, 1977.
- [8] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, pp. 400-407, 1951.
- [9] L. Rutkowski, "Sequential estimates of a regression function by orthogonal series with applications in discrimination," in *Lecture Notes in Statistics 8*. New York: Springer Verlag, 1981, pp. 236-244.
- [10] —, "On-line identification of time-varying systems by nonparametric techniques," *IEEE Trans. Automat. Contr.*, vol. 27, pp. 228-230, 1982.
- [11] C. I. Stone, "Optimal rates of convergence for non-parametric estimators," *Ann. Statist.*, vol. 8, pp. 1348-1360, 1980.
- [12] W. Stute, "Sequential fixed-width confidence interval for a non-parametric density function," *Z. Wahrsch.*, vol. 62, pp. 113-123, 1983.
- [13] J. H. Venter, "On Dvoretzky stochastic approximation theorems," *Ann. Math. Statist.*, vol. 37, pp. 1534-1544, 1966.
- [14] H. Walk, "An invariance principle for the Robbins-Monro process in a Hilbert space," *Z. Wahrsch.*, vol. 39, pp. 135-150, 1977.

AUTOMATIC CURVE SMOOTHING

Wolfgang Härdle
Institut Wirtschaftstheorie II
Universität Bonn
Adenauerallee 24-26
D-5300 Bonn,
Federal Republic of Germany

1. INTRODUCTION

Regression smoothing is a method for estimating the mean function from observations $(x_1, Y_1), \dots, (x_n, Y_n)$ of the form

$$Y_i = m(x_i) + \epsilon_i, \quad i=1, \dots, n,$$

where the observation errors are independent, identically distributed, mean zero random variables. There are a number of approaches for estimating the regression function m . Here we discuss nonparametric smoothing procedures, which are closely related to local averaging, i.e. to estimate $m(x)$, average the Y_i 's which are in some neighborhood of x . The width of this neighborhood, commonly called bandwidth or smoothing parameter, controls the smoothness of the curve estimate. Under weak conditions (bandwidth shrinks to zero not too rapidly as n increases) the curve smoothers consistently estimate the regression function m . In practice, however, one has to select a smoothing parameter in some way. A too small bandwidth, resulting in high variance, is not acceptable and so is *oversmoothing* which creates a large bias. It is therefore highly desirable to have some *automatic curve smoothing* procedure.

W. Härdle

Proposed methods for choosing the window size automatically are based on estimates of the prediction error or adjustments of the residual sum of squares. It has been shown by Härdle, Hall and Marron (1986) (HHM) that all these proposals are asymptotically equivalent but can be quite different in a practical situation. In this paper we highlight these difficulties with automatic curve smoothing and construct situations where some of the proposals seem to be preferable.

2. AUTOMATIC CURVE SMOOTHING

To simplify the presentation, assume the design points are equally spaced, i.e. $x_i = i/n$, and assume that the errors have equal variance, $E\epsilon^2 = \sigma^2$. We study *kernel smoothers*

$$\hat{m}_h(x) = n^{-1} h^{-1} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) Y_i$$

where h is the bandwidth and K is a symmetric kernel function. It is certainly desirable to tailor the automatic curve smoothing so that the resulting regression estimate is close to the true curve. Most automatic bandwidth procedures are designed to optimize the averaged squared error (ASE)

$$d_A(h) = n^{-1} \sum_{i=1}^n [\hat{m}_h(x_i) - m(x_i)]^2 w(x_i),$$

where w is some weight function. These automatic bandwidth selectors are defined by multiplying $p(h) = n^{-1} \sum_{i=1}^n (Y_i - \hat{m}_h(x_i))^2 w(x_i)$ by a correction factor $E(n^{-1} h^{-1})$. The examples we treat here are

General Cross-Validation (Craven and Wahba 1979),

$$E_{GCV}(n^{-1} h^{-1}) = (1 - n^{-1} h^{-1} K(0))^{-2}.$$

Akaike's Information Criterion (Akaike 1970),

$$E_{AIC}(n^{-1} h^{-1}) = \exp(2n^{-1} h^{-1} K(0)).$$

Finite Prediction Error (Akaike 1974),

$$E_{FPE}(n^{-1} h^{-1}) = (1 + n^{-1} h^{-1} K(0)) / (1 - n^{-1} h^{-1} K(0)).$$

Automatic Curve Smoothing

A model selector of Shibata (1981),

$$\mathbb{E}_S(n^{-1}h^{-1}) = 1 + 2n^{-1}h^{-1}K(0).$$

The bandwidth selector T of Rice (1984),

$$\mathbb{E}_T(n^{-1}h^{-1}) = (1 - 2n^{-1}h^{-1}K(0))^{-1}.$$

Let \hat{h} denote the bandwidth that minimizes $(p \cdot E)(h)$. The automatic curve smoother is defined as $\hat{m}_{\hat{h}}(x)$. This automatic curve smoothing procedure is asymptotically optimal for the above \mathbb{E} in the sense that

$$\frac{d_A(\hat{h})}{d_A(\hat{h}_0)} \xrightarrow{P} 1,$$

where \hat{h}_0 denotes the minimizer of d_A . The relative differences are quantified in the

Theorem. Let $\hat{h}_0 \sim n^{-1/5}$ then

$$n^{3/10}(\hat{h} - \hat{h}_0) \rightarrow N(0, \sigma^2)$$

$$n[d_A(\hat{h}) - d_A(\hat{h}_0)] \rightarrow C \cdot \chi_1^2.$$

in distribution, where σ^2 and C are defined in HHM.

A very remarkable feature of this result is that the constants σ^2 and C are independent of \mathbb{E} . In a simulated example we generated 100 samples of size $n=75$ with $\sigma=0.05$ and $m(x) = \sin(\lambda 2\pi x)$. The kernel function was taken to be $K(x) = (15/8)(1-4x^2)^2 I(|x| \leq 1/2)$. Table 1 shows the number exceedances by ratios of error criteria d_A and $E\{d_A\} = d_M$, for 100 data sets of size $n=100$.

PARAMETER LAMBDA = 1		1.05	1.1	1.2	1.4	1.6	1.8	2	4	6	8
RICE											
DA	50	34	15	4	2	1	1	0	0	0	0
DM	18	8	0	0	0	0	0	0	0	0	0
GCV											
DA	51	34	18	12	9	7	7	1	0	0	0
DM	30	17	11	7	3	2	2	0	0	0	0
FPE											
DA	75	63	52	49	47	40	37	10	2	0	0
DM	65	56	52	47	40	40	40	0	0	0	0
AIC											
DA	91	86	85	83	81	69	63	13	3	0	0
DM	86	84	84	82	79	79	79	0	0	0	0
SHIBATA											
DA	100	100	100	100	99	90	81	15	3	0	0
DM	100	100	100	100	100	100	100	0	0	0	0

Table 1

W. Härdle

Rice's proposal T shows a quite good performance. Note that T has a slight bias towards oversmoothing, since this \hat{E} has a pole at $2n^{-1}K(0)$ whereas all the other selectors have no pole or a pole at $n^{-1}K(0)$, the "no smoothing point". By increasing λ to 2 (Table 2) the T-selector loses its good performance, and is clearly outperformed by Generalized Cross-Validation (GCV).

PARAMETER		LAMBDA = 2									
		1.05	1.1	1.2	1.4	1.6	1.8	2	4	6	8
GCV											
DA		37	24	15	4	3	3	2	0	0	0
DM		31	10	10	2	0	0	0	0	0	0
RICE											
DA		53	26	6	0	0	0	0	0	0	0
DM		25	5	0	0	0	0	0	0	0	0
FPE											
DA		92	85	75	57	36	19	9	0	0	0
DM		84	80	80	80	0	0	0	0	0	0
AIC											
DA		99	96	90	69	43	23	11	0	0	0
DM		97	95	95	95	0	0	0	0	0	0
SHIBATA											
DA		100	100	95	74	46	23	11	0	0	0
DM		100	100	100	100	0	0	0	0	0	0

Table 2

The reason for that is the tendency of T to oversmooth; this behavior is penalized in a situation where the reduction of bias becomes more important than the reduction of variance. This becomes apparent in Table 3 where λ was 3.

PARAMETER		LAMBDA = 3									
		1.05	1.1	1.2	1.4	1.6	1.8	2	4	6	8
GCV											
DA		51	28	8	1	0	0	0	0	0	0
DM		69	4	0	0	0	0	0	0	0	0
RICE											
DA		69	36	15	1	0	0	0	0	0	0
DM		99	17	1	0	0	0	0	0	0	0
SHIBATA											
DA		89	70	33	4	2	1	1	0	0	0
DM		100	100	0	0	0	0	0	0	0	0
FPE											
DA		89	70	33	4	2	1	1	0	0	0
DM		100	100	0	0	0	0	0	0	0	0
AIC											
DA		100	100	95	51	19	4	3	0	0	0
DM		100	100	100	100	0	0	0	0	0	0

Table 3

Automatic curve smoothing

References

- Akaike, H. (1970). Statistical predictor information. Annals of the Institute of Statistical Mathematics, 22, 203-217.
- Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, AC19, 719-723.
- Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. Numerische Mathematik, 31, 377-403.
- Härdle, W., Hall, P. and Marron, J.S. (1986). How far are automatically chosen regression smoothing parameters from their optimum? (with discussion). J. Amer. Stat. Assoc., to appear.
- Rice, J. (1984). Bandwidth choice for nonparametric regression. Annals of Statistics, 12, 1215-1230.
- Shibata, R. (1981). An optimal selection of regression variables. Biometrika, 68, 45-54.

WHAT REGRESSION MODEL SHOULD BE CHOSEN WHEN THE
STATISTICIAN MISSPECIFIES THE ERROR DISTRIBUTION ?

Wolfgang Härdle¹

ABSTRACT. We consider the situation where the statistician fails to choose the correct Likelihood function in a regression model. We propose a model selection rule and show its asymptotic optimality. Relationships to C_p and extensions of AIC are discussed.

1. INTRODUCTION. Let $\underline{y}_n = (Y_1, \dots, Y_n)'$ be a random vector of n independent observations with mean vector $\underline{\mu}_n = (\mu_1, \dots, \mu_n)'$. Assume that the i^{th} mean μ_i is associated with an infinite covariate x_i in a linear way, i.e.

$$\mu_i = \langle x_i, \beta \rangle$$

where the parameter β and the covariate are in ℓ_2 , the space of square summable sequences equipped with the canonical inner product. Suppose that the observation error $e_i = Y_i - \mu_i$ has distribution F with density f . In general the statistician does not know f , so he might fix a different error density h and use the model

$$H_p = \left\{ \prod_{i=1}^n h(Y_i - \langle x_i, \beta(p) \rangle) : \beta'(p) = (0, \dots, \beta_{p_1}, 0, \dots, \beta_{p_2}, \dots, \beta_{p_{k(p)}}, 0, \dots) \right\}$$

where $p_1 < p_2 < \dots < p_{k(p)}$ and $k(p) \geq 1$ is the dimension. Such a mismatch of the chosen model and the true error distribution can happen in a variety of cases. For instance, the statistician could apply a robust regression procedure (Huber, 1973) but the data is in fact Gaussian. We may also imagine the reverse situation. A natural way of evaluating goodness of the regression model H_p is to introduce some kind of distance between the predicted re-

¹Partially supported by Deutsche Forschungsgemeinschaft SFB 123 "Stochastische Mathematische Modelle".

ression surfaces and the true model regression μ_n . We consider here the Euklidean distance

$$L_n(p) = \|\mu_n - \hat{\mu}_n(p)\|^2,$$

where $\hat{\mu}_n(p) = \langle x_1, \hat{\beta}(p) \rangle$ denotes the predicted regression surface based on the maximum likelihood estimate $\hat{\beta}(p)$ in model H_p . The assumed error density h is thought of being fixed, so the above loss $L_n(p)$ depends only on

$$p = (p_1, p_2, \dots, p_k(p))$$

which we call from now on *the model* p .

Which model p should be selected if a variety P_n of models is possible? In this paper we derive an efficient model selection procedure and prove that it asymptotically minimizes $L_n(p)$ over a set of models P_n . Results of this type have been obtained by Shibata (1981), Breiman and Friedman (1983) in the case of $h \equiv f \equiv \varphi$, the density of the normal distribution function. Recently Li (1984) gave conditions for asymptotic efficiency of least squares estimators using model choice procedures based on cross validation, FPE among others. In these special cases our procedure is equivalent. It is also related to an extension of AIC given by Takeuchi (1976). This connection is investigated in Section 3, the main result in the next section.

2. ASYMPTOTIC EFFICIENCY WHEN THE REGRESSION MODEL IS MISSPECIFIED. Define the linear operator $X_n: \ell_2 \rightarrow \mathbb{R}^n$ by

$$X_n' = (x_1' x_2' \dots x_n'), \text{ the vector of observation errors } e_n = (e_1, e_2, \dots, e_n)' \text{ and } \psi(u) = -\frac{d}{du} \log h(u), \gamma = E_F \psi^2(e) / (E_F \psi'(e))^2.$$

An expansion of $L_n(p)$ motivates the score

$$W_n'(p) = -\|\hat{\mu}_n(p)\|^2 + 2\gamma k(p) + \|\mu_n\|^2.$$

To see this, observe that

$$\begin{aligned} W_n'(p) - L_n(p) &= -\{\|\hat{\mu}_n(p) - \mu_n\|^2 + 2\langle \hat{\beta}(p) - \beta, \beta \rangle_{B_n} + \|\mu_n\|^2\} \\ &\quad + 2\gamma k(p) + \|\mu_n\|^2 - \|\hat{\mu}_n(p) - \mu_n\|^2 \\ (1) \qquad &= -2\|\hat{\mu}_n(p) - \mu_n\|^2 + 2\gamma k(p) \\ &\quad - 2\{-\|\hat{\mu}_n(p) - \mu_n\|^2 + \langle \hat{\beta}(p) - \beta, \hat{\beta}(p) \rangle_{B_n}\} \\ &= 2\{\gamma k(p) - \langle \hat{\beta}(p) - \beta, \hat{\beta}(p) \rangle_{B_n}\}, \end{aligned}$$

where $\langle u, v \rangle_{B_n} = u' B_n v$, $B_n = X_n' X_n$.

Suppose that the last term in (1) is tending to a constant uniformly over models $p \in P_n$. Then minimizing $W_n'(p)$ over P_n will be the same task, at least asymptotically, as minimizing $L_n(p)$. Two unknowns are still involved in $W_n'(p)$. The constants $\|\mu_n\|^2$ and γ depend on the unknown regression function and the unknown true error distribution F . The constant $\|\mu_n\|^2$ does not cause difficulties since it is independent of the model: So minimizing $W_n'(p) - \|\mu_n\|^2$ is the same as minimizing $W_n'(p)$. The scaling factor γ can be estimated by a consistent sequence of estimators

$$\hat{\gamma}_n = \frac{n^{-1} \sum_{i=1}^n \psi^2(Y_i - \langle x_i, \hat{\beta}(p_n) \rangle)}{[n^{-1} \sum_{i=1}^n \psi'(Y_i - \langle x_i, \hat{\beta}(p_n) \rangle)]^2}$$

where $\{p_n\}$ is a model sequence of increasing dimension.

We will therefore define

$$W_n(p) = - \|\hat{\mu}_n\|^2 + 2\hat{\gamma}_n k(p)$$

as the score function that is to be minimized over P_n . The concept of *asymptotic efficiency* is defined as follows.

DEFINITION. A selected \hat{p} is called asymptotically efficient if, as $n \rightarrow \infty$,

$$\frac{L_n(\hat{p})}{\inf_{p \in P_n} L_n(p)} \xrightarrow{P} 1$$

The following conditions are needed.

CONDITION 1. The function ψ is twice differentiable and fulfills

- (i) $E_F \psi(e) = 0$
- (ii) $q = E_F \psi'(e) > 0$
- (iii) $\exists N > 0: E_F [q^{-1}(\psi'(e) - q)]^{2N} < \infty$.

CONDITION 2. The matrix $B_n(p) = X_n'(p)X_n(p)$ has full rank $k(p)$, where $X_n(p)$ is the (n, p) matrix containing only the non-zero control variables of model p . There exists a $N > 0$ such that

$$\sum_{p \in P_n} \tilde{R}_n(p)^{-N} \rightarrow 0, \text{ as } n \rightarrow \infty,$$

where $\tilde{R}_n(p) = E_F \tilde{L}_n(p)$, $\tilde{L}_n(p) = \|\mu_n - \tilde{\mu}_n(p)\|^2$ and $\tilde{\mu}_n(p)$ denotes the Gauß-Markov estimator $X_n(p)B_n^{-1}(p)X_n'(p)\tilde{Y}_n$. applied to the pseudodata $Y_i = \mu_i + \tilde{e}_i$, $\tilde{e}_i = \psi(e_i)/q$.

CONDITION 3. Let $h(p)$ be the largest diagonalelement of the hat matrix $M_n(p) = X_n(p)B_n^{-1}(p)X_n'(p)$.

Assume $\sup_{p \in P_n} h(p)\tilde{R}_n(p) \rightarrow 0$, as $n \rightarrow \infty$.

THEOREM. Choose \hat{p} such that it minimizes $W_n(p)$ over P_n . Under conditions 1-3, \hat{p} is asymptotically efficient.

REMARK 1. The estimates $\hat{\beta}(p)$ will be compared with the Gauß-Markov estimates in model p , that are based on the (non observable) pseudodata $\tilde{Y}_i = \mu_i + \tilde{e}_i$. It will then be seen that the problem of asymptotic efficiency can be solved by an analogous problem formulated for the linear estimates $\tilde{\mu}_n(p)$. Details of the proof of the theorem are contained in Härdle (1985).

REMARK 2. It follows from condition 2 that $k^2(p)/n \rightarrow 0$, as $n \rightarrow \infty$. This is an analogue of the (necessary) condition " $p^2/n \rightarrow 0$ ", that can be found in Huber (1981, p.166).

REMARK 3. If ψ is a bounded function, as is assumed in robust regression analysis, condition 2 can be weakened. It is seen from the proof in Härdle (1985) that $\sum_{p \in P_n} \exp(-c\tilde{R}_n(p)) \rightarrow 0$, $c > 0$ is sufficient.

3. CONNECTION TO OTHER METHODS. In the case of least squares estimators there are a variety of model selection procedures, such as generalized cross validation, FPE, AIC or Mallows' (1973) C_p . Shibata (1981) and Li (1984) have shown the equivalence of these procedures. The linearization of $W_n(p)$ i.e. the score function W_n based on the nonobservable pseudodata \tilde{Y}_n has a similar structure as C_p . The latter score for \tilde{Y}_n reads

$$\begin{aligned}
 C_p &= \|\tilde{y}_n - \tilde{\mu}_n\|^2 + 2\gamma k(p) \\
 &= \|\tilde{\epsilon}_n\|^2 + \tilde{L}_n(p) + 2\tilde{\epsilon}_n'(I_n - M_n(p))\mu_n \\
 &\quad + 2\{\gamma k(p) - \tilde{\epsilon}_n' M_n(p)\tilde{\epsilon}_n\}.
 \end{aligned}$$

Expand $W_n(p)$ with $\hat{\gamma}_n$ replaced by γ

$$\begin{aligned}
 W_n(p) &= -\|\tilde{\mu}_n(p)\|^2 + 2\gamma k(p) \\
 &= -\|\tilde{\mu}_n(p) - \mu_n\|^2 - \|\mu_n\|^2 - 2(\tilde{\mu}_n(p) - \mu_n)'(\mu_n - \tilde{\mu}_n(p)) \\
 &\quad - 2(\tilde{\mu}_n(p) - \mu_n)'\tilde{\mu}_n(p) + 2\gamma k(p) \\
 &= \tilde{L}_n(p) + 2\gamma k(p) - 2\tilde{\epsilon}_n' M_n(p)\tilde{\epsilon}_n + 2\tilde{\epsilon}_n'(I_n - M_n(p))\mu_n \\
 &\quad + 2\mu_n'\tilde{\epsilon}_n - \|\mu_n\|^2.
 \end{aligned}$$

The last two terms are independent of p . The remaining terms are identical to those in Mallows' C_p .

There is a different way of deriving $W_n(p)$. Our way was to argue that $\hat{\mu}_n(p)$ is asymptotically like the least square estimator $\tilde{\mu}_n(p)$ based on the unobservable pseudodata \tilde{y}_n . This made it possible to argue as in the linear case. Takeuchi (1976) argued in a different way, when he heuristically extended Akaike's (1970) AIC to the case of mismatching the true likelihoodfunction. Takeuchi gave no proof, but his derivation is interesting, we therefore want to present it here again.

Denote by $I(f, h_{\beta}(p))$ the Kullback-Leibler information number between f and $h_{\beta}(p) \in H_p$. Consider the prediction error $E^Y(I(f, h_{\hat{\beta}}(p)))$, where $\hat{\beta}(p)$ is the maximum likelihood estimator, which is based on $\{(x_i, y_i)\}_{i=1}^n$ a data set with $\{y_i\}$ distributed as $\{Y_i\}$. Write

$$\begin{aligned}
 E^Y(I(f, h_{\hat{\beta}}(p))) &= E^Y \int f(u) \log \frac{f(u)}{h(u; \hat{\beta}(p))} du \\
 &= \int f(u) \log f(u) du - E^Y \int f(u) \log h(u; \hat{\beta}(p)) du
 \end{aligned}$$

and observe that the first term is independent of the model p .

Expand

$$\begin{aligned}
 \log h(u, \hat{\beta}(p)) &= \log h(u; \beta^*(p)) + (\hat{\beta}(p) - \beta^*(p)) \frac{\partial}{\partial \beta} \log h(u; \beta^*(p)) \\
 &\quad + \frac{1}{2} (\hat{\beta}(p) - \beta^*(p))' \left(\frac{\partial^2}{\partial \beta \partial \beta'} \log h(u; \beta^*(p)) \right) (\hat{\beta}(p) - \beta^*(p)) \\
 &\quad + \dots,
 \end{aligned}$$

where $\beta^*(p)$ minimizes $I(f, h_{\beta^*(p)})$.

Then

$$\int f(u) \log f(u; \hat{\beta}(p)) du \\ \doteq \int f(u) \log f(u; \beta^*(p)) du - \frac{1}{2} (\hat{\beta}(p) - \beta^*(p))' J (\hat{\beta}(p) - \beta^*(p))$$

where

$$J = E \left(- \frac{\partial^2}{\partial \beta \partial \beta'} \log h(u; \beta^*) \right).$$

Since the maximum likelihood estimator is asymptotically normally distributed

$$\sqrt{n} (\hat{\beta}(p) - \beta^*(p)) \xrightarrow{L} N(0, J^{-1} I J^{-1})$$

with $I = E \left(\frac{\partial}{\partial \beta} \log h(u; \beta^*(p)) \frac{\partial}{\partial \beta'} \log h(u; \beta^*(p)) \right)$

it follows that

$$E^Y \int f(u) \log h(u; \hat{\beta}(p)) du \doteq \int f(u) \log h(u; \beta^*(p)) du - \text{tr}(I J^{-1}) / 2n.$$

On the other hand for the data set $\{(X_i, Y_i)\}_{i=1}^n$

$$\log h(\underline{y}_n; \beta^*(p)) = \log h(\underline{y}_n; \hat{\beta}(p)) \\ + (\beta^*(p) - \hat{\beta}(p))' \frac{\partial}{\partial \beta} \log h(\underline{y}_n; \hat{\beta}(p)) \\ + \frac{1}{2} (\beta^*(p) - \hat{\beta}(p))' \frac{\partial^2}{\partial \beta \partial \beta'} \log h(\underline{y}_n; \hat{\beta}(p)) (\beta^*(p) - \hat{\beta}(p))$$

where $\hat{\beta}(p)$ is the maximum likelihood estimator based on $\{(X_i, Y_i)\}_{i=1}^n$.

Then

$$\int f(u) \log h(u; \beta^*(p)) du = n^{-1} \int f(\underline{y}_n - X_n \beta) \log h(\underline{y}_n; \beta^*(p)) d\underline{y}_n \\ \doteq n^{-1} \int f(\underline{y}_n - X_n \beta) \log h(\underline{y}_n; \hat{\beta}(p)) d\underline{y}_n - \text{tr}(I J^{-1}) / 2n,$$

and

$$E^Y \int f(u) \log h(u; \hat{\beta}(p)) du \\ \doteq n^{-1} \int f(\underline{y}_n - X_n \beta) \log h(\underline{y}_n; \hat{\beta}(p)) d\underline{y}_n - \frac{1}{n} \text{tr}(I J^{-1}).$$

Consequently the task to minimize $E^Y (I(f; h_{\hat{\beta}(p)}))$ is the same as to minimize the approximate quantity

$$-E(\log h(\underline{y}_n; \hat{\beta}(p))) + \text{tr}(I J^{-1})$$

and if we replace the expectation by the observation that we have at hand we obtain

$$\text{GAIC}(p) = -2 \log h(\underline{y}_n; \hat{\beta}(p)) + 2 \text{tr}(I J^{-1}).$$

which we call *Generalized AIC*.

In the regression setting we have

$$\log h(\underline{y}_n; \beta(p)) = \sum_i \log h(Y_i - \langle x_i, \beta(p) \rangle)$$

$$IJ^{-1} = \begin{matrix} & V_1 & \dots & 0 & & W_1 & \dots & 0 \\ X' & : & \dots & X & X' & : & \dots & X^{-1} \\ & 0 & \dots & V_n & & 0 & \dots & W_n \end{matrix}$$

where $V_i = \text{var}(\psi(Y_i - \langle x_i, \beta^*(p) \rangle))$

$W_i = E\psi'(Y_i - \langle x_i, \beta^*(p) \rangle)$.

A Taylor expansion shows that

$V_i \doteq E\psi^2(e_i)$ if $\langle x_i, \beta - \beta^*(p) \rangle \rightarrow 0$

$W_i \doteq E\psi'(e_i)$

Therefore $\text{tr}(IJ^{-1}) \doteq p \cdot \frac{E\psi^2}{E\psi'}$.

Now

$$\begin{aligned} \log h(\underline{y}_n; \beta(p)) &= \log h(\underline{y}_n; \hat{\beta}(p)) + (\beta - \hat{\beta}(p))' \frac{\partial}{\partial \beta} \log h(\underline{y}_n; \hat{\beta}(p)) \\ &\quad + \frac{1}{2} (\beta - \hat{\beta}(p))' \frac{\partial^2}{\partial \beta \partial \beta} \log h(\hat{\beta}(p)) + \dots \\ &\doteq \log h(\underline{y}_n; \hat{\beta}(p)) - \frac{1}{2} (\beta - \hat{\beta}(p))' J (\beta - \hat{\beta}(p)). \end{aligned}$$

Therefore

$$\begin{aligned} \log h(\underline{y}_n; \hat{\beta}(p)) &= \log h(\underline{y}_n; \beta) + \frac{1}{2} \|\beta - \hat{\beta}(p)\|_J^2 \\ &= \log h(\underline{y}_n; \beta) + \frac{1}{2} \|\beta\|_J^2 - \langle \beta, \hat{\beta}(p) \rangle_J + \frac{1}{2} \|\hat{\beta}(p)\|_J^2 \end{aligned}$$

and

$$\begin{aligned} -2 \log h(\underline{y}_n; \hat{\beta}(p)) &\doteq -2 \log h(\underline{y}_n; \beta) - \|\beta\|_J^2 + 2 \langle \beta, \hat{\beta}(p) \rangle_J - \|\hat{\beta}(p)\|_J^2 \\ &= -2 \log h(\underline{y}_n; \beta) - \|\beta\|_J^2 + 2 \langle \beta, \hat{\beta}(p) - \beta \rangle_J + 2 \|\beta\|_J^2 \\ &\quad - \|\hat{\beta}(p)\|_J^2. \end{aligned}$$

The crossterm tends to zero as $\langle x_i, \beta - \beta^*(p) \rangle \rightarrow 0$.

So we have that minimizing

$$\text{GAIC}(p) = -2 \log h(\underline{y}_n; \hat{\beta}(p)) + 2 \text{tr}(IJ^{-1})$$

is asymptotically the same as minimizing

$$-\|\hat{\beta}(p)\|_J + 2 \gamma q p$$

since $J \doteq X'X/q$, this is approximately equal to $W_n(p)$.

BIBLIOGRAPHY

1. Akaike, H., "Statistical Predictor Identification", *Ann. Inst. Math. Stat.*, 22 (1970), 203-217.
2. Breiman, L. and Freedman, D., "How many variables should be entered in a regression equation", *J. Amer. Stat. Assoc.*, 78 (1983), 131-136.
3. Härdle, W., "An effective selection of regressive variables when the error distribution is correctly specified", submitted for publication, (1985)
4. Huber, P., "Robust regression: Asymptotics, conjectures, and Monte Carlo", *Ann. Statist.*, 1 (1973), 799-821.
5. Huber, P., *Robust Statistics*, Wiley, New York, (1981).
6. Mallows, C., "Some comments on C_p ", *Technometrics*, 15 (1973), 661-675.
7. Li, K.C., Asymptotic optimality for C_p , C_c , cross-validation and generalized cross-validation: Discrete index set., Manuscript, (1984).
8. Shibata, R., "An optimal selection of regression variables", *Biometrika*, 68 (1981), 45-54.
9. Takeuchi, K., "Distribution of information statistics and a criterion of model fitting", *Suri Kagaku*, 153 (1976), 12-18, (in Japanese).

INSTITUT FÜR GESELLSCHAFTS- UND
WIRTSCHAFTSWISSENSCHAFTEN
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN
FEDERAL REPUBLIC OF GERMANY

SEQUENTIAL KERNEL SMOOTHING FOR ESTIMATION OF ZEROS AND
LOCATION OF EXTREMA OF REGRESSION FUNCTIONS

Wolfgang K. Härdle
Wirtschaftstheorie II
Adenauerallee 24-26
Universität Bonn
D-5300 Bonn, FRG

1. Introduction

An enumeration of zeros and of locations of extrema often suffices to describe the approximate shape of regression functions. The estimation of these quantities is not only quantifying the shape of the functions but also offers the possibility of comparing shapes in a group of similar shaped functions. An illustrating example is the human height growth curve, where location and size of a peak in the second derivative of this regression curve serve to describe the so-called midgrowth spurt, see Müller (1985). Such "longitudinal parameters", as they have been called by Gasser et al. (1984), can be used for group comparisons.

The nonparametric regression model that we consider is

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, 2, \dots$$

with a sequence of independent identically distributed bivariate random variables $(X, Y), (X_1, Y_1), (X_2, Y_2), \dots$ and $m(x) = E(Y|X=x)$, the unknown nonparametric regression curve. Note that the structure of this sampling scheme is different from the so-called fixed design model, where the predictor variables $\{X_i\}$ are fixed in advance or can be tuned by the experimenter. In this paper we consider the sequential estimation of zeros and location of extrema of $m(x)$ by combining nonparametric kernel smoothing with stochastic approximation methods. The proposed sequential scheme is based on weighted averaging of the response variables $\{Y_i\}$.

For the estimation of a zero θ_0 of m , for instance, we

propose the recursive procedure

$$Z_{n+1} = Z_n - a_n h_n^{-1} K((Z_n - X_n)/h_n) Y_n, \quad n \geq 1. \quad (1).$$

Here Z_1 denotes an arbitrary random variable, $\{a_n\}$ and $\{h_n\}$ are sequences of positive real numbers and $K: \mathbb{R} \rightarrow \mathbb{R}$ is a kernel function. The kernel K parametrizes the shape of the weight sequence, whereas the bandwidths $\{h_n\}$ regulate its size. The recursion (1) is constructed in analogy to the classical Robbins and Monro (1951) procedure but differs with respect to the weights $h_n^{-1} K((Z_n - X_n)/h_n)$. Note that for kernel functions with bounded support this weight can be zero, so the estimation process $\{Z_n\}$ may stay at the same value for a while. We show that $\{Z_n\}$ converges to a zero of $\tilde{m}(x) = m(x)f_X(x)$, (f_X the marginal density of X) and derive asymptotic normality of $\{Z_n\}$. The latter result serves in constructing fixed width confidence intervals for the zero θ_0 of the regression curve.

The location of extrema can be identified by observing that $m' = \tilde{r}/f_X^2$, with

$$\tilde{r}(x) = f_X(x) \int y \frac{\partial}{\partial x} f(x,y) dy - \tilde{m}(x) f_X'(x),$$

where $f(x,y)$ denotes the joint density of (X,Y) . Under suitable assumptions (e.g. f_X strictly positive) the problem of finding the location of an extremum of m is equivalent to finding a zero of \tilde{r} . We therefore propose to perform the estimation of this location by

$$Z'_{n+1} = Z'_n - a_n h_n^{-3} Y_n \{K((Z'_n - \bar{X}_n)/h_n) K'((Z'_n - X_n)/h_n) - K'((Z'_n - \bar{X}_n)/h_n) K((Z'_n - X_n)/h_n)\}, \quad n \geq 1. \quad (2)$$

Here $\{\bar{X}_n\}$ denotes an additional i.i.d. sequence with the

same distribution as X . We shall show that $\{Z'_n\}$ is consistently estimating the location of an extremum of m .

An alternative way of defining an estimator of the zero of the regression function m could be to construct an estimate of the whole function and then to use a zero of the function estimate as an estimator for the zero of the regression function, see Müller (1985). This procedure can be extremely time and space consuming in the case of sequential observation of the data: For every new observation the whole function would have to be constructed, whereas our procedure just keeps one number in memory and updates this number recursively.

Related work on the sequential estimation of the regression function itself can be found in Revesz (1977) and Rutkowski (1981, 1982). The idea of deriving fixed width confidence intervals was inspired by the papers of Chow and Robbins (1965), McLeish (1976), and Stute (1983).

2. Results

We only describe the important conditions on the functions and parameters of the recursive algorithm. Assumptions of more technical character can be found in Härdle and Nixdorf (1987) where also proofs are given. The speed of convergence of $\{a_n\}$ and $\{h_n\}$ is controlled by

$$\sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} a_n h_n < \infty, \quad (3)$$

$$\sum_{n=1}^{\infty} a_n^2 h_n^{-2} < \infty, \quad (4)$$

$$\sum_{n=1}^{\infty} a_n^2 h_n^{-4} < \infty. \quad (5)$$

The zero θ_0 of $m(x)$ (and of $\tilde{m}(x)$) is identified by

$$\inf_{\epsilon \leq |x - \theta_0| \leq 1/\epsilon} (x - \theta_0) \tilde{m}(x) > 0 \text{ for all } \epsilon > 0. \quad (6)$$

The kernel function K has to satisfy one of the following conditions.

$$\begin{aligned} &K \text{ is bounded and } \int K(u) du = 1, \\ &\int u K(u) du = 0, \quad \int u^2 K(u) du < \infty. \end{aligned} \quad (7)$$

K is differentiable with bounded derivative K' and

$$\lim_{|u| \rightarrow \infty} |u K(u)| = 0 \quad \int |u| K'^2(u) du < \infty. \quad (8)$$

K is twice differentiable and

$$\lim_{|u| \rightarrow \infty} |u K'(u)| = 0 \quad \int |u| K''(u) du < \infty. \quad (9)$$

The consistency of $\{Z_n\}$ is shown in

Theorem 1. Assume (3), (4), (6), (7) and $EY^2 < \infty$. Then $\{Z_n\}$ converges to θ_0 almost surely and in quadratic mean.

Asymptotic normality follows from

Theorem 2. Assume (6-7). Let $EY^4 < \infty$ and suppose that the joint density $f(x, y)$ is twice differentiable, $a_n = n^{-1}$, $h_n = n^{-\gamma}$, $1/5 \leq \gamma < 1/2$. Then

$$n^{(1-\gamma)/2} (Z_n - \theta_0) \xrightarrow{D} N(b(\gamma), \sigma^2(\gamma))$$

where

$$b(\gamma) = 0 \quad \text{if } 1/5 < \gamma < 1/2,$$

$$\tilde{m}'(\theta_0) \int u^2 K(u) du / (2\tilde{m}'(\theta_0) - 1 + \gamma) \quad \text{if } \gamma = 1/5,$$

$$\sigma^2(\gamma) = \int K^2 \int y^2 f(\theta_0, y) dy / (2m'(\theta_0) - 1 + \gamma).$$

Fixed width asymptotic confidence intervals for the unknown parameter θ_0 are constructed via estimators of the asymptotic bias $b(\gamma)$ and variance $\sigma^2(\gamma)$. Estimators of $\int y^2 f(\theta_0, y) dy$, $\tilde{m}'(\theta_0)$, $\tilde{m}''(\theta_0)$ are respectively,

$$S_{1n} = n^{-1} \sum_{i=1}^n K_{h_i} (Z_i - X_i) Y_i^2$$

$$S_{2n} = n^{-1} \sum_{i=1}^n K'_{h_i} (Z_i - X_i) Y_i$$

$$S_{3n} = n^{-1} \sum_{i=1}^n K''_{h_i} (Z_i - X_i) Y_i,$$

where $K_h(u) = h^{-1} K(u/h)$, $h = h_n$.

An estimator for the asymptotic variance $\sigma^2(\gamma)$ is therefore

$$s_n = \int K^2 S_{1n} / (2S_{2n}^{-1+\gamma}), \text{ if } 2S_{2n}^{-1+\gamma} > 0 \\ = 1, \text{ otherwise.}$$

On the basis of this estimator the following stopping rule seems reasonable:

$$N(d) = \inf \{n \in \mathbb{N} : s_n + n^{-1} \{n^{1-\gamma} d^2 / z_{\alpha/2}^2\}$$

here $z_{\alpha/2}$ denotes the $(1-\alpha/2)$ -quantile of the standard normal distribution. The fixed width confidence intervals can be constructed via

Theorem 3. Let $a_n = n^{-1}$, $h_n = n^{-\gamma}$, $1/5 \leq \gamma < 1/3$ and assume (6-9) and $EY^4 < \infty$. Then if $N(d)$ is defined as above for some $0 < \alpha < 1$, as $d \rightarrow 0$,

$$N(d)^{(1-\gamma)/2} \{Z_{N(d)} - \theta_0\} \xrightarrow{d} N(b(\gamma), \sigma^2(\gamma)).$$

In the case $1/5 < \gamma < 1/3$ an asymptotic confidence interval of fixed length $2d$ and asymptotic coverage probability $1-\alpha$ is given by

$$\left[Z_{N(d)} - d, Z_{N(d)} + d \right].$$

In the case $\gamma = 1/5$ the bias has to be estimated by

$$b_n = \int u^2 K(u) du S_{3n} / (2S_{2n}^{-1+\gamma}).$$

Then with $H_n = Z_n - n^{(-1+\gamma)/2} b_n$ an asymptotic confidence interval is given by

$$\left[H_{N(d)} - d, H_{N(d)} + d \right].$$

The analysis of the sequential procedure $\{Z'_n\}$ is quite similar to that of $\{Z_n\}$.

Theorem 4. Define the zero of \tilde{r} as θ_M and assume (7-8),

$EY^4 < \infty$ and (6) fulfilled with \tilde{r} in the place of m . Then $\{Z'_n\}$ converges to θ_m almost surely and in the quadratic mean.

Theorem 5. Let $a_n = n^{-1}$, $h_n = n^{-\gamma}$, $1/6 \leq \gamma < 1/4$ and suppose that the assumptions of Theorem 4 hold. Then

$$n^{(-1+\gamma)/2} (Z'_n - \theta_M) \xrightarrow{D} N(0, \sigma_M^2(\gamma)).$$

where

$$\sigma_M^2(\gamma) = f_X(\theta_M) \int K^2 \int K \cdot 2 \int y^2 f(\theta_M, y) dy / (2\tilde{r}'(\theta_M) - 1 + 4\gamma).$$

Note that the rate of convergence of $\{Z'_n\}$ is for $\gamma=1/5$ equal to $\{n^{-2/5}\}$. This rate is optimal under our assumptions on the nonparametric regression function, as Stone(1980) has shown. Under stronger assumptions on the bandwidth sequence Müller(1985) achieved a slightly slower rate.

3. Simulation study

The basic experiment consisted of 200 Monte Carlo replications with the number $N(d)$, $Z_{N(d)}$ and $S_{N(d)}$ to be reported. The joint probability density function was $f(x,y) = I_{(0,1)}(x) \sigma^{-1} \psi((y-m(x))/\sigma)$, ψ the standard normal probability density and $m(x) = -a((1-x^2)-1/4)$ for $a=4,8$ was the regression curve. Table 1 below shows the results for $Z_1=0.45$. For Table 2 the starting point Z_1 was set to 0.2. The parameter α was set to $\alpha=0.05$. The zero is $\theta_0=1/2$ and d was set at 0.05, 0.1 and σ was 0.1, 1.0. The kernel $K(u) = 0.75(1-u^2)I_{(-1,1)}(u)$ was used and the bandwidths were set at $h=h_n = n^{-\gamma}$, $\gamma=0.21$.

TABLE 1

N(d)						
a	d	Mean	STD	Q ₅	Q ₉₅	
0.1	0.05	137	10.0	120	156	
0.1	0.05	229	17.0	200	259	
0.1	0.10	43	3.5	38	50	
0.1	0.10	62	6.7	51	74	
1.0	0.05	642	97.0	496	806	
1.0	0.05	469	46.0	394	551	
1.0	0.10	129	37.0	76	207	
1.0	0.10	103	18.0	72	133	

Z _{N(d)}						
a	d	Mean	STD	Q ₅	Q ₉₅	Counts
0.1	0.05	0.518	0.021	0.483	0.553	188
0.1	0.05	0.515	0.019	0.483	0.547	194
0.1	0.10	0.519	0.030	0.469	0.574	199
0.1	0.10	0.517	0.039	0.449	0.582	196
1.0	0.05	0.510	0.025	0.467	0.548	188
1.0	0.05	0.515	0.023	0.475	0.559	181
1.0	0.10	0.520	0.054	0.429	0.610	184
1.0	0.10	0.525	0.043	0.461	0.596	192

s _{N(d)}						
a	d	Mean	STD	Q ₅	Q ₉₅	m' (θ ₅)
0.1	0.05	0.024	0.002	0.020	0.028	4
0.1	0.05	0.043	0.003	0.037	0.048	8
0.1	0.10	0.027	0.005	0.018	0.035	4
0.1	0.10	0.050	0.007	0.038	0.062	8
1.0	0.05	0.105	0.013	0.127	0.084	4
1.0	0.05	0.081	0.006	0.070	0.093	8
1.0	0.10	0.110	0.028	0.065	0.168	4
1.0	0.10	0.090	0.015	0.062	0.114	8

TABLE 2

N(d)						
a	d	Mean	STD	Q ₅	Q ₉₅	
0.1	0.05	163	12	141	183	
0.1	0.05	255	17	227	283	
0.1	0.10	56	7	46	70	
0.1	0.10	74	8	61	90	
1.0	0.05	646	83	518	602	
1.0	0.05	479	44	417	550	
1.0	0.10	139	31	89	192	
1.0	0.10	118	19	86	146	

Z _{N(d)}						
a	d	Mean	STD	Q ₅	Q ₉₅	Counts
0.1	0.05	0.517	0.018	0.485	0.547	192
0.1	0.05	0.518	0.019	0.484	0.552	189
0.1	0.10	0.513	0.027	0.464	0.561	199
0.1	0.10	0.515	0.036	0.456	0.593	196
1.0	0.05	0.516	0.026	0.471	0.562	174
1.0	0.05	0.512	0.023	0.475	0.550	188
1.0	0.10	0.515	0.044	0.437	0.595	192
1.0	0.10	0.522	0.042	0.454	0.590	193

s _{N(d)}						
a	d	Mean	STD	Q ₅	Q ₉₅	m'(θ_0)
0.1	0.05	0.030	0.002	0.025	0.034	4
0.1	0.05	0.040	0.003	0.042	0.052	8
0.1	0.10	0.040	0.007	0.031	0.059	4
0.1	0.10	0.064	0.007	0.050	0.077	8
1.0	0.05	0.105	0.011	0.088	0.126	4
1.0	0.05	0.082	0.006	0.073	0.093	8
1.0	0.10	0.118	0.024	0.077	0.158	4
1.0	0.10	0.102	0.015	0.074	0.125	8

The numerical values of Table 1 indicate that the fixed accuracy result of Theorem 3 yields a good approximation of θ_0 even for $d=0.1$. This can be read from the counts in the $Z_{N(d)}$ column. It is indicated there how many times, out of the 200 Monte Carlo runs, the true parameter was in the confidence interval.

$$\left[Z_{N(d)}^{-d}, Z_{N(d)}^{+d} \right].$$

As a measure of dispersion we added the empirical quantiles Q_{95} and Q_5 in the third and fourth column of each entry.

Note that the values for $N(d)$ for $a=8$ are actually bigger than the values for $a=4$. This seems to contradict the intuition since it is expected that the procedure stops earlier for larger derivatives at the zero. This effect can be explained from the crude approximation of the conditional standard deviation $(\text{var}(Y|X=x))^{1/2}$ by s_n in a neighborhood of θ_0 . The tables show that the statistic $s_{N(d)}$ considerably overestimates the true asymptotic scale $\sigma(\gamma)$. For comparison we list some correct $\sigma(\gamma) = \sigma(a, \gamma)$. For instance $\sigma(0.1, 4, 0.21) = 0.00083$ whereas $\sigma(0.1, 8, 0.21) = 0.00039$

BIBLIOGRAPHY

Chow, Y.S. and Robbins, H. (1965) On the asymptotic theory of fixed width sequential confidence intervals for the mean. Ann.Math.Statist., 36, 457-462

Gasser, T., Müller, H.G., Köhler, W., Molinari, L. and Prader, A. (1984) Nonparametric regression analysis of growth curves. Ann.Statist., 12, 210-229

Härdle, W.K. and Nixdorf, R. (1987) Nonparametric Sequential estimation of Zeros and extrema of regression Functions IEEE Trans. Inf. Theory, 32, in print

McLeish, R. (1976) Functional and random central limit theorems for the Robbins-Monro Process. J. Appl. Prob., 13, 148-154

Müller, H.G. (1985) Kernel estimators of zeros and of location and size of extrema of regression functions. Scand.J.Statist.,12,221-232

Revesz,P.(1977) How to apply the method of stochastic approximation in the nonparametric estimation of a regression function.Math.Oper.Series Statistics,8,119-126

Robbins, H. and Monro, S.(1951) A stochastic approximation method. Ann.Math.Statist.,22,400-407

Rutkowski, L. (1981) Sequential estimates of a regression function by orthogonal series with application in discrimination. in Lecture Notes in Statistics Springer Verlag, New York

Rutkowski, L. (1982) On-line identification of time varying systems by nonparametric techniques. IEEE Trans. Autom.Control,27,228-232

Stone, C.J. (1980) Optimal rates of convergence for nonparametric estimators. Ann.Statist.,8,1348-1360

Stute, W. (1983) Sequential fixed width confidence intervals for a nonparametric density function. Z.Wahrscheinlichkeitstheorie 62,113-123

SUMMARY

Let $(X,Y),(X_1,Y_1),(X_2,Y_2),\dots$ be independent identically distributed pairs of random variables and let $m(x)=E(Y|X=x)$ be the regression function of Y on X . The estimation of zeros and of location of extrema of this regression curve is considered by combining the nonparametric kernel method with stochastic approximation techniques. Consistency and asymptotic normality of the proposed procedures is shown, providing fixed width confidence intervals. The proposed algorithms are investigated by simulations.

RESUME

Soit $(X, Y), (X_1, Y_1), (X_2, Y_2) \dots$ des couples aléatoires indépendants et identiquement distribué. On note par $m(x) = E(Y/X=x)$ la fonction de regression de Y ou X. L'estimation des zero et des lieux des extrema de cette fonction de regression est abordée en combinant les techniques nonparametrique utilisant la methode du noyau et celles d'approximation stochastique. La convergence et la normalité asymptotique de la procedure proposée ist établie et fournit des intervalles de confiance. Les resultats sont illustrés au moins d'experiance par simulations.

Nonparametric Kernel Regression Estimation— Optimal Choice of Bandwidth

WOLFGANG HÄRDLE¹ and GABRIELLE KELLY¹

Universität Bonn and University College, Cork

Summary. The use of kernel regression estimators is well known in the estimation of regression surfaces. The estimators involve a kernel with bandwidth $h (> 0)$. The choice of h is important since a small h gives an estimator with a large variance, but if a large h is used then the bias is large. The bias is under specific smoothness assumptions, a functional of higher derivatives of the regression curve. From a nonparametric viewpoint it is therefore desirable to choose the bandwidth in such a way that the variance and the bias are balanced independently of the smoothness of the curve. In this paper it is shown how such an asymptotically optimal h can be found. The construction of such an optimal bandwidth independent of the smoothness of the regression curve gives a positive answer to Question 3 of STONE's (1982) paper. The proof only requires mild assumptions on the underlying density and the moments of the dependent variable y . An interesting relationship is discovered between the moments of y and the smoothness of the kernel. The results of the present work extend that of STONE (1984) on kernel density estimation.

AMS 1980 subject classifications: Primary: 62 G 05; secondary: 62 G 20.

Key words: Kernel regression estimation; automatic smoothing; choice of smoothing parameter.

1. Discussion

Given vectors $(X_1, Y_1), \dots, (X_n, Y_n)$, where X is $d \times 1$, from a density $f_{X,Y}(x, y)$ in $(d+1)$ -dimensional Euclidean space, the problem is to estimate the regression curve of Y on X given by

$$m(x) = E[Y | X=x] = \int y f_{X,Y}(x, y) dy / f(x),$$

where $f(x)$ is the marginal density of X . The NADARAYA-WATSON estimators (NADARAYA, 1964; WATSON, 1964) have the form

$$m_{nh}^*(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i) Y_i / f_{nh}(x), \tag{1.1}$$

where $K_h(x) = h^{-d} K(x/h) = h^{-d} K(x_1/h, \dots, x_d/h)$ is a delta function sequence involving a kernel K and $f_{nh}(x)$ is the familiar ROSENBLATT-PARZEN kernel estimator

¹ Research partially supported by Deutsche Forschungsgemeinschaft, SFB 123 „Stochastische Mathematische Modelle“ and by Public Health Service Grant 5R01 GM21 215-10.

of the marginal density $f(x)$ given by

$$f_{nh}(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i). \tag{1.2}$$

The parameter $h = h(n)$ is called bandwidth and regulates the speed of the delta function sequence.

Here we consider slightly different estimators of the form

$$m_{nh}(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i) Y_i / f(x)$$

with the nonrandom denominator $f(x)$. These estimators, considered by JOHNSTON (1982), are also interesting in their own right in cases where the marginal density $f(x)$ is known. There are situations where the experimenter can choose the design variable X and to reduce sampling bias he may wish to randomize and then use the estimator m_{nh} . In a later paper the results here will be extended to the NADARAYA-WATSON estimators.

How do we measure the appropriateness of the bandwidth h ? In a first attempt one could try to optimize some functional, such as integrated squared error loss, of the difference $m_{nh}(x) - m(x)$. However, it is well known that the variance of $m_{nh}(x)$ is proportional to $f^{-1}(x)$ so a reasonable measure of the performance of the estimator is $(m_{nh}(x) - m(x))^2 f(x)$. Integrating this with respect to $f(x)$ gives us the integrated weighted squared error loss function

$$\text{ISE}(n, h) = \int (m_{nh}(x) - m(x))^2 f^2(x) dx, \tag{1.3}$$

that was also considered in a similar setting by NADARAYA (1982).

Our goal is to choose $h = h(n)$ to minimize (1.3). Observe that

$$\text{ISE}(n, h) = \int m_{nh}^2(x) f^2(x) dx - 2 \int m_{nh}(x) m(x) f^2(x) + \int m^2(x) f^2(x) dx.$$

Therefore minimizing $\text{ISE}(n, h)$ is equivalent to minimizing $\int m_{nh}^2(x) f^2(x) dx - 2 \int m_{nh}(x) m(x) f^2(x) dx$. The cross-term $\int m_{nh}(x) m(x) f^2(x) dx$ is unknown because $m(x)$ is unknown. Note however

$$\begin{aligned} \mathbf{E}[\int m_{nh} m f^2] &= \mathbf{E}[\int K_h(x - X_i) Y_i m(x) f(x) dx] \\ &= \int \int K_h(x - u) m(u) m(x) f(u) f(x) du dx \\ &= \mathbf{E}[\sum_{i \neq j} \sum K_h(X_i - X_j) Y_i Y_j / n(n-1)]. \end{aligned} \tag{1.4}$$

Also $\int m_{nh}^2 f^2 = \sum_i \sum_j K_h^{(2)}(X_i - X_j) Y_i Y_j / n^2$ where $K_h^{(2)}$ denotes the convolution of K_h with itself. Thus an estimate of $\text{ISE}(n, h) - \int m^2 f^2$ is

$$\hat{\text{ISE}}(n, h) = \sum_i \sum_j K_h^{(2)}(X_i - X_j) Y_i Y_j / n^2 - 2 \sum_{i \neq j} \sum K_h(X_i - X_j) Y_i Y_j / n^2$$

and we choose $\hat{h} = \hat{h}(n)$ to minimize $\hat{\text{ISE}}(n, h)$. The main result of this paper is that \hat{h} is asymptotically optimal in the sense that

$$\lim_n \frac{\text{ISE}(n, \hat{h})}{\min_h \text{ISE}(n, h)} = 1, \tag{1.5}$$

with probability one, subject to some mild assumptions on $K, f(x)$ and the moments of Y . This result extends a Theorem by STONE (1984) for kernel density estimators to the case of regression curves.

There are other techniques to select reasonable bandwidths. An expansion using "vanishing moment" conditions on K shows that if m is twice continuously differentiable

$$E\{\text{ISE}\} = An^{-1}h^{-d} + Bh^4 + o(n^{-1}h^{-d} + h^4)$$

where A and B are constants depending on m, f . Ignoring now lower order terms this shows that the bandwidth sequence minimizing $E\{\text{ISE}\}$ is proportional to $n^{-1/5}$. Therefore only the proportionality factor, involving A, B has to be found. This approach was taken by HALL (1984) who considered the optimization of h in a range $[an^{-1/5}, bn^{-1/5}]$, $a, b > 0$. In the so-called fixed design setting RICE (1984) showed asymptotic optimality of different bandwidth selectors in the same range $[an^{-1/5}, bn^{-1/5}]$. He considered selectors derived from AIC (AKAIKE, 1974), FPE (AKAIKE, 1970) and cross-validation among others and showed their asymptotic equivalence if the design variables are equispaced. However it should be noted that if $f(x)$ is not uniform these bandwidth selectors are not asymptotically equivalent (HÄRDLE and MARRON, 1985b).

Our approach here is related to cross-validation. To see this observe that the cross-validation function

$$CV(h) = n^{-1} \sum_j (Y_j - n^{-1} \sum_{i \neq j} K_h(X_j - X_i) Y_i / f(X_j))^2 f(X_j),$$

setting $w \equiv f$ in HÄRDLE and MARRON (1985a) equals

$$n^{-1} \sum_j Y_j^2 f(X_j) + n^{-1} \sum_j [n^{-1} \sum_{i \neq j} K_h(X_j - X_i) Y_i / f(X_j)]^2 f(X_j) - 2n^{-2} \sum_j \sum_{i \neq j} K_h(X_j - X_i) Y_i Y_j.$$

The first term in this sum is independent of h , the second term appears to be a discrete approximation of $\int m_{nh}^2 f^2$ and the third term exactly equals the second sum in $\text{ISE}(h)$.

The above-mentioned asymptotic optimality result (1.5) gives a positive answer to Question 3 in STONE (1982). To see this, note that from (1.5) we know that there exists a constant $c > 1$ such that independent of the smoothness of m

$$\lim_{n \rightarrow \infty} P \{ \text{ISE}(h) \geq c \text{ISE}(h^*) \} = 0, \tag{1.6}$$

where h^* is the minimizer of $\text{ISE}(h)$. Now under STONE'S (1982) definition of smoothness classes $\text{ISE}(h^*) \sim n^{-2p/(2p+d)}$ (MARRON and HÄRDLE, 1986), where p denotes the number of existing derivatives of m . So (1.6) gives the answer to his question. Other approaches to derive a result similar to (1.6) but with different estimators and slightly different forms of loss functions were taken by CHEN (1984) and HÄRDLE and MARRON (1984a). In the last mentioned paper, the range of bandwidths over which is optimized is $[n^{-1+\delta}, n^{-\delta}]$ for some small positive δ .

This somewhat restricted range does not make it possible to achieve the optimal rate of convergence (STONE, 1982) for regression functions with $p > \frac{1}{2} \left(\frac{1}{\delta} - d \right)$. Our present paper improves upon that restriction since the bandwidth is optimized over the positive real numbers.

This paper is organized as follows: In section 2 we state some assumptions and the main result together with two lemmas that prove the theorem. In section 3 auxiliary lemmas are proved.

As in STONE (1984) a Poissonization argument (section 4) is used to compute higher moments. If h is restricted to be chosen only from a finite set H_n the number of moments of Y that are required for (1.6) to hold can be explicitly computed. An interesting relationship between the cardinality of this set H_n , the smoothness of the kernel, and the number of required moments of Y is discussed in section 5.

2. Main result

We make the following assumptions:

- (i) K has compact support and $\int K(u) du = 1$;
- (ii) K is symmetric about 0;
- (iii) There are constants $M > 0$, $\zeta > 0$ such that $|K(y) - K(u)| \leq M |y - u|^\zeta$, for $x, y \in \mathbb{R}^d$;
- (iv) $m(x)$ is bounded and $E[Y^2 | X = x]$ is bounded in x ;
- (v) All moments of Y exist and $EY \neq 0$;
- (vi) f is bounded.

Theorem. *Let assumptions (i) to (vi) be satisfied. Then the selector $\hat{h} > 0$ which minimizes*

$$n^{-2} \sum_i \sum_j K_{\hat{h}}^{(2)}(X_i - X_j) Y_i Y_j - 2n^{-2} \sum_{i \neq j} K_{\hat{h}}(X_i - X_j) Y_i Y_j$$

is asymptotically optimal, in the sense that,

$$\lim_{n \rightarrow \infty} \frac{\text{ISE}(n; \hat{h})}{\inf_{h > 0} \text{ISE}(n; h)} = 1 \quad \text{w.p.1.}$$

The following lemma shows that the bias of m_{nh} vanishes asymptotically if and only if h tends to zero. Define

$$m_h(x) = E m_{nh}(x) = \int K_h(x - u) m(u) f(u) du / f(x), \tag{2.1}$$

and

$$\|m_h - m\|^2 = \int (m_h - m)^2 f^2(x) dx. \tag{2.2}$$

Lemma 1. *There are positive constants b, γ such that*

$$\|m_h - m\|^2 \geq \gamma (h^b \wedge 1) \quad \text{for } h \in \mathbb{R}^+. \tag{2.3}$$

Proof. We first show that $\|m_h - m\|^2$ is bounded away from zero for h outside any neighborhood of the origin, i.e. given $\delta > 0$ we show $\inf_{h > \delta} \|m_h - m\|^2 > 0$. Suppose this is false. Then $\exists \delta > 0$ and an $h > \delta$ for which

$$\|m_h - m\|^2 = 0,$$

hence, $\int [\int K_h(x-u) m(u) f(u) du - m(x) f(x)]^2 dx = 0$. Writing $g(u) = m(u) f(u)$ we have

$$\int K_h(x-u) g(u) du - g(x) = 0 \quad \forall x.$$

Letting F denote FOURIER transform (which exists for the following quantities since we assume $E|Y| < \infty$) we have

$$F(K_h) F(g) = F(g)$$

and therefore

$$F(K_h) \equiv 1$$

which is impossible by the RIEMANN-LEBESGUE theorem.

It remains to show $\|m_h - m\|^2 \geq \gamma h^b$, for $h \leq \delta$ for some $0 < \delta < 1$. Now let φ, φ_h and g be the FOURIER transforms of K, K_h and g respectively. Then $\varphi(0) = E(Y)$ and

$$(2\pi)^d \|m_h - m\|^2 = \int (1 - \varphi_h)^2 |\varphi|^2.$$

Since $\varphi(0) \neq 0$ by assumption, there is a compact set C centered at the origin for which $|\varphi|^2 \geq K$ on C . Also $\varphi_h(t) = \varphi(ht)$. Then

$$\begin{aligned} \|m_h - m\|^2 &> K \int_C (1 - \varphi_h(t))^2 dt \\ &= K \int_C \left(1 - \int_{-\infty}^{\infty} K(u) \cos(tuh) du \right)^2 dt \\ &= K \int_C \left(\sum_{n=1}^{\infty} \int \frac{(tuh)^{2n}}{(2n)!} K(u) du \right)^2 dt. \end{aligned}$$

Again by the RIEMANN-LEBESGUE lemma there exists an integer p with

$$\begin{aligned} \int u^k K(u) du &= 0 \quad \text{for } k \leq 2p, \quad \text{and} \\ \int u^k K(u) du &\neq 0 \quad \text{for } k > 2p. \end{aligned}$$

Then

$$\begin{aligned} \|m_h - m\|^2 &\geq K \int_C \left(\sum_{n=p+1}^{\infty} \int \frac{(tuh)^{2n}}{(2n)!} K(u) du \right)^2 dt \\ &= Ch^{4(p+1)}, \end{aligned}$$

where $C > 0$. Now choosing $b > 4(p+1)$ establishes the result.

To verify that h is asymptotically optimal it suffices to show

Lemma 2.

$$\limsup_{h, h'} \frac{|\text{ISE}(n, h') - \text{ISE}(n, h) - (\widehat{\text{ISE}}(n, h') - \widehat{\text{ISE}}(n, h))|}{\text{ISE}(n, h) + \text{ISE}(n, h')} = 0, \tag{2.4}$$

with probability one.

The theorem is now easily shown.

Proof of Theorem. Let

$$\widehat{\text{ISE}}(n, \hat{h}) = \inf_h \widehat{\text{ISE}}(n, h)$$

and let

$$\text{ISE}(n, h^*) = \inf_h \text{ISE}(n, h).$$

Now let $\varepsilon > 0$ be given. Then by Lemma 2 we have with probability 1,

$$\frac{\text{ISE}(n, \hat{h}) - \text{ISE}(n, h^*) - (\widehat{\text{ISE}}(n, \hat{h}) - \widehat{\text{ISE}}(n, h^*))}{\text{ISE}(n, \hat{h}) + \text{ISE}(n, h^*)} \leq \varepsilon.$$

This implies

$$0 \cong \widehat{\text{ISE}}(n, \hat{h}) - \widehat{\text{ISE}}(n, h^*) \cong (1 - \varepsilon) \text{ISE}(n, \hat{h}) - (1 + \varepsilon) \text{ISE}(n, h^*)$$

which entails

$$(1 + \varepsilon) \text{ISE}(n, h^*) \cong (1 - \varepsilon) \text{ISE}(n, \hat{h}),$$

or

$$1 \leq \frac{\text{ISE}(n, \hat{h})}{\text{ISE}(n, h^*)} \leq \frac{1 + \varepsilon}{1 - \varepsilon}.$$

Since $\varepsilon > 0$ was arbitrary, so

$$\mathbf{P} \left\{ \lim_{n \rightarrow \infty} \left| \frac{\text{ISE}(n, \hat{h})}{\text{ISE}(n, h^*)} - 1 \right| < \delta \right\} = 1 \quad \forall \delta > 0.$$

We will now show Lemma 2.

For this define $J_{nh} = \|m_h - m\|^2 + \frac{1}{nh^2}$. The idea is to replace the random denominator in Lemma 2 by J_{nh} .

Lemma 3. *If the following conditions hold*

$$\underline{\lim}_h \sup \frac{\text{ISE}(n, h)}{J_{nh}} > 0 \quad \text{with probability one} \tag{2.5}$$

and

$$\limsup_n \sup_{h, h'} \frac{|\text{ISE}(n, h') - \text{ISE}(n, h) - (\widehat{\text{ISE}}(n, h') - \widehat{\text{ISE}}(n, h))|}{J_{nh} + J_{nh'}} = 0 \tag{2.6}$$

with probability one,

then (2.4) holds.

Proof. We show conditions (2.5) and (2.6) imply condition (2.4). The first condition says that there exists a constant $c > 0$ such that

$$\mathbf{P} \left\{ \underline{\lim}_h \sup \text{ISE}(n, h) > cJ_{nh} \right\} = 1.$$

Let $\varepsilon > 0$ given, then $\exists n_0$ and $\forall n \geq n_0$

$$\mathbf{P} \left\{ \text{ISE}(n, \hat{h}) > cJ_{nh} \right\} \cong 1 - \varepsilon.$$

Let

$$Y_n = \sup_{h, h'} \frac{|\text{ISE}(n, h') - \text{ISE}(n, h) - (\widehat{\text{ISE}}(n, h') - \widehat{\text{ISE}}(n, h))|}{\text{ISE}(n, h) + \text{ISE}(n, h')}.$$

Then $\forall n \geq n_0$

$$Y_n < \frac{1}{c} \sup_{h, h'} \frac{|\text{ISE}(n, h') - \text{ISE}(n, h) - (\widehat{\text{ISE}}(n, h') - \widehat{\text{ISE}}(n, h))|}{J_{nh} + J_{nh'}}.$$

By assumption the left-hand side converges to zero with probability one, so Y_n does also. This establishes the statement (2.4). It remains to prove equations (2.5) and (2.6) which is done in the following section.

3. Auxiliary results

We now define

$$G_{nh} = \sum_{i=1}^n m_h(X_i) Y_i f(X_i) / n - \mathbb{E}[m_h(X) Y f(X)]$$

and

$$G_n = \sum_{i=1}^n m(X_i) f(X_i) Y_i - \mathbb{E}[m(X) f(X) Y].$$

Then

$$\begin{aligned} & \text{ISE}(n, h) - \widehat{\text{ISE}}(n, h) - \int m^2 f^2 - 2G_n \tag{3.1} \\ &= 2(G_{nh} - G_n) + 2 \iint_{x+u} K_h(x-u) yv [P_n(dx, dy) - P(dx, dy)] \\ & [P_n(du, dv) - P(du, dv)], \end{aligned}$$

where P_n is the empirical distribution of (X_i, Y_i) and P is the joint distribution of (X, Y) . To prove (2.6) we must show both terms on the r.h.s. of (3.1) divided by J_{nh} converge to zero with probability one. Since

$$\text{ISE}(n, h) = \int (m_{nh} - m_h)^2 f^2 + \|m_h - m\|^2 + 2 \int (m_{nh} - m_h) (m_h - m) f^2 \tag{3.2}$$

to prove (2.5) we will show

$$\limsup_n \sup_h \frac{\|m_h - m\|}{J_{nh}} > 0 \tag{3.3}$$

and

$$\limsup_n \sup_h \int (m_{nh} - m_h) (m_h - m) f^2 / J_{nh} = 0 \tag{3.4}$$

with probability one.

To establish (3.3) observe that

$$\begin{aligned} \sup_h \|m_h - m\|^2 / J_{nh} &= \sup_h 1 / [1 + (nh \|m_h - m\|^2)^{-1}] \\ &> 1 / [1 + (nh_0 \|m_{h_0} - m\|^2)^{-1}] \quad \text{for some } h_0 \in \mathbb{R}^+. \end{aligned}$$

Choose n large so that $1/[nh_0 \|m_{h_0} - m\|^2] < \varepsilon$. Then

$$\sup_h \|m_h - m\|^2 / J_{nh} > 1/(1 + \varepsilon) > 0 .$$

It remains to prove the following lemmas:

Lemma 4.

$$(a) \limsup_n \sup_h |G_{nh} - G_n| / J_{nh} = 0 \quad \text{with probability one.}$$

and

$$(b) \limsup_n \sup_h \left| \int (m_{nh} - m_h) (m_h - m) f^2 / J_{nh} \right| = 0 \quad \text{with probability one .}$$

Lemma 5.

$$\limsup_n \sup_h \int \int_{x \neq u} K_h(x-u) yv [F_n(dx, dy) - F(dx, dy)] [F_n(du, dv) - F(du, dv)] / J_{nh} = 0 \quad \text{with probability one.}$$

Lemma 5 will be proved by a Poissonization argument in section 4. We now prove Lemma 4.

Proof of Lemma 4 (a). Let

$$Z_{ih} = m_h(X_i) Y_i f(X_i) - m(X_i) f(X_i) Y_i \\ - [E(m_h(X) Y f(X)) - E(m(X) f(X) Y)] .$$

Then

$$n^{-1} \sum_{i=1}^n Z_{ih} = G_{nh} - G_n .$$

Clearly Z_{ih} are i.i.d. and $E(Z_{ih}) = 0, i \geq 1$. If we assume initially $|Y_i| \leq K, i \geq 1$, for some $K > 0$, it is easy to show using the assumption on the boundedness of f that there is a positive constant c such that $|Z_{ih}| \leq c$ and $\text{Var}(Z_{ih}) \leq c \|m_h - m\|^2$. Thus using BERNSTEIN'S inequality for bounded random variables (HOEFFDING 1963) we have

$$P \{ \bar{Z}_{nh} \geq t \} \leq \exp \{ -\tau \lambda / 2 (1 + \lambda / 3) \}$$

where $0 \leq \lambda \leq t / \|m_h - m\|^2$ and $\tau = nt/c$. Let $\varepsilon > 0$ be given. Suppose $\|m_h - m\| \geq n^{\varepsilon-1/2}$. Put $t = n^{\varepsilon-1/2} \|m_h - m\|$ and $\lambda = n^{\varepsilon-1/2} / \|m_h - m\|$. Then $\lambda \tau = n^{2\varepsilon}/c$ and

$$P \{ |\bar{Z}_{nh}| \geq n^{\varepsilon-1/2} \|m_h - m\| / \|m_h - m\| \geq n^{\varepsilon-1/2} \} \leq \exp \{ -n^{2\varepsilon}/3c \} .$$

Suppose $\|m_h - m\| < n^{\varepsilon-1/2}$. Put $t = n^{2\varepsilon-1}$ and $\lambda = 1$. Again $\lambda \tau = n^{2\varepsilon}/c$ and

$$P \{ |\bar{Z}_{nh}| \geq n^{2\varepsilon-1} / \|m_h - m\| < n^{\varepsilon-1/2} \} \leq \exp \{ -n^{2\varepsilon}/3c \} .$$

Therefore

$$P \{ |\bar{Z}_{nh}| \geq n^{\varepsilon-1/2} \|m_h - m\| + n^{2\varepsilon-1} \} \leq 2 \exp \{ -n^{2\varepsilon}/3c \} .$$

Thus for any finite set H_n where $\#H_n \leq An^a$, and A and a are positive constants,

$$\lim_n \mathbf{P} \{ |\bar{Z}_{nh}| \geq n^{\varepsilon-1/2} \|m_h - m\| + n^{2\varepsilon-1}, \text{ for some } h \in H_n \} = 0.$$

Now for $0 < \varepsilon < 1/2 (1 + b)$

$$\lim_n \sup_h \frac{n^{\varepsilon-1/2} \|m_h - m\| + n^{2\varepsilon-1}}{J_{nh}} \leq \lim_n \sup_{u>0} \frac{n^{\varepsilon-1/2}u + n^{2\varepsilon-1}}{u^2 + 1/nu^{2/b}} = 0, \tag{3.5}$$

where b is defined in Lemma 1. Therefore,

$$\mathbf{P} \left\{ \lim_n \max_{h \in H_n} \frac{|\bar{Z}_{nh}|}{J_{nh}} > 0 \right\} = 0.$$

Now consider Y not bounded and let

$$Z_{ih}^* = Z_{ih} I \{ |Y_i| \leq A_n \}, \quad 1 \leq i \leq n,$$

where $A_n, n \geq 1$ are positive constants. Again Z_{ih}^* are i.i.d., $\mathbf{E}(Z_{ih}^*) = 0$ and there are positive constants K_n such that $|Z_{ih}^*| \leq K_n$ and $\text{Var}(Z_{ih}^*) \leq K_n \|m_h - m\|^2$ for $1 \leq i \leq n$.

Let $\bar{Z}_{nh}^* = \sum_{i=1}^n Z_{ih}^*/n$. Then

$$\begin{aligned} \mathbf{P} \{ |\bar{Z}_{nh}| \geq t \} &= \mathbf{P} \{ |\bar{Z}_{nh}| \geq t, \text{ for all } 1 \leq i \leq n \mid Y_i \leq A_n \} \\ &\quad + \mathbf{P} \{ |\bar{Z}_{nh}| \geq t, \mid Y_i > A_n, \text{ for some } 1 \leq i \leq n \} \\ &\leq \mathbf{P} \{ |\bar{Z}_{nh}^*| \geq t \} + n\mathbf{P} \{ |Y| > A_n \} \\ &\leq \exp \{ -n^{2\varepsilon}/3A_n \} + n\mathbf{E}(|Y|^k)/(A_n)^{-k}. \end{aligned}$$

Therefore

$$\begin{aligned} \mathbf{P} \{ |\bar{Z}_{nh}| \geq n^{\varepsilon-1/2} \|m_h - m\| + n^{2\varepsilon-1} \text{ for some } h \in H_n \} \\ \leq n^a \exp \{ -n^{2\varepsilon}/3A_n \} + n^{a+1} \mathbf{E}(|Y|^k)/(A_n)^k. \end{aligned}$$

Choose $A_n < n^{2\varepsilon-\delta}$ where $0 < \delta < 2\varepsilon$, then the first term when summed over n , converges to zero and in order for $\sum_{n=1}^{\infty} A_n^{-k} n^{a+1}$ to be finite subject to $0 < \varepsilon < 1/2 (1 + b)$, we need $k > (2 + \alpha + a)(b + 1)$ where $\alpha > 0$. By assumption (v) we can choose k so that this condition is satisfied. Thus, using (3.5)

$$\mathbf{P} \left\{ \lim_n \sup_{h \in H_n} \frac{|\bar{Z}_{nh}|}{J_{nh}} > 0 \right\} = 0.$$

It remains to prove the result for $h \in \mathbf{R}^+$. The $|\bar{Z}_{nh}|/J_{nh}$ is a decreasing function of h^- , for h large, thus by appropriate choice of A , where $\#H_n = An^a$ we have

$$\sup_h \frac{|\bar{Z}_{nh}|}{J_{nh}} \leq \sup_{h' \in H_n} \frac{|\bar{Z}_{nh'}|}{J_{nh'}} + \sup_{|h-h'| \leq n^{-a}} \left| \frac{\bar{Z}_{nh}}{J_{nh}} - \frac{\bar{Z}_{nh'}}{J_{nh'}} \right|.$$

We need to show

$$\sup_{|h-h'| \leq n^{-a}} \sum_{i=1}^n n^{-1} \left| \frac{Z_{ih}}{J_{nh}} - \frac{Z_{ih'}}{J_{nh'}} \right| \rightarrow 0 \text{ with probability one.}$$

Now

$$\begin{aligned} \sum_{i=1}^n \left| \frac{Z_{ih}}{J_{nh}} - \frac{Z_{ih'}}{J_{nh'}} \right| / n &= \sum_{i=1}^n \frac{|J_{nh'} Z_{ih} - J_{nh} Z_{ih'}|}{J_{nh} J_{nh'}} / n \\ &= \sum_{i=1}^n n^{-1} |(Z_{ih} - Z_{ih'}) J_{nh'} + (J_{nh'} - J_{nh}) Z_{ih'}| / J_{nh} J_{nh'}. \end{aligned} \tag{3.6}$$

Considering the first term of (3.6) we have

$$\frac{|Z_{ih} - Z_{ih'}|}{J_{nh}} \leq \frac{C \int |K_h(X_i - u) - K_{h'}(X_i - u) f(u)| du}{J_{nh}} |Y_i f(X_i)|,$$

or some positive constant c . Now

$$\begin{aligned} &\int \frac{|K_h(X_i - u) - K_{h'}(X_i - u)|}{J_{nh}} f(u) du \\ &= \frac{1}{J_{nh}} \left\{ \int \left| \frac{(h' - h)}{hh'} K\left(\frac{X_i - u}{h}\right) f(u) + \frac{f(u)}{h'} K\left(\frac{X_i - u}{h}\right) - K\left(\frac{X_i - u}{h'}\right) \right| du \right\}. \end{aligned}$$

Using $|h' - h| \leq n^{-a}$ and HÖLDER continuity of K , it is easy to show there is a constant c such that this term is bounded by $c_1 n^{-l}$ for some $l > 0$, $c_1 > 0$. Considering the second term of (3.6) it is not difficult to show

$$\frac{|J_{nh} - J_{nh'}|}{J_{nh} J_{nh'}} \leq c_2 n^{-l}, \quad \text{where } c_2, l > 0.$$

Noting that $|Z_{ih}| \leq k |Y_i|$ where $k > 0$, we have

$$\sup_{|h' - h| \leq n^{-l}} \sum_{i=1}^n n^{-1} \left| \frac{Z_{ih}}{J_{nh}} - \frac{Z_{ih'}}{J_{nh'}} \right| \leq c_n \sum_{i=1}^n \frac{Y_i}{n},$$

where $c_n \rightarrow 0$ as $n \rightarrow \infty$. Since $\sum_{i=1}^n Y_i/n$ is bounded with probability one, this implies the result.

To prove part (b) of Lemma 4 set

$$Z_{ih} = \int (K_h(x - X_i) Y_i - m_h(x) f(x)) / (m_h(x) f(x) - m(x) f(x)) dx.$$

As before Z_{ih} are i.i.d., $E(Z_{ih}) = 0$ and there is a positive constant c such that $\text{Var}(Z_{ih}) \leq c \|m_h - m\|^2$. The conclusion follows by applying the argument of (a).

4. The Poissonization argument

Define $J_{nh^r} = h^r \wedge 1 + 1/nh^a$ for $r > 0$.

Proof of Lemma 5. Let $\lambda > 0$ be given and let $N(dx, dy)$ be a POISSON process on $\mathbb{R}^d \times \mathbb{R}^l$ with $EN(A) = \lambda P(A)$ where P is the joint probability measure (p.m.) of (X, Y) . Set $M(dx, dy) = N(dx, dy) - \lambda P(dx, dy)$. Given a positive integer l let P^l be the p.m. on $(\mathbb{R}^d \times \mathbb{R}^l)$ defined by

$$P^l(dZ_1, \dots, dZ_l) = P(dZ_1) \dots P(dZ_l) = P(dx_1, dy_1) \dots P(dx_l, dy_l).$$

Let k and l denote positive integers with $l \leq k$. Let Γ_{kl}^0 denote the collection of all k -tuples i_1, \dots, i_k of integers in $\{1, \dots, l\}$ such that:

- (a) each $i \in \{1, \dots, l\}$ appears one or more times among i_1, \dots, i_k ;
- (b) if $i, i' \in \{1, \dots, l\}$ and $i < i'$, then i appears before i' among i_1, \dots, i_k .

Given $Z = (Z_1, \dots, Z_k) \in (\mathbb{R}^d \times \mathbb{R})^k$ and $\gamma = (i_1, \dots, i_k) \in \Gamma_{kl}^0$, set $Z_\gamma = (Z_{i_1}, \dots, Z_{i_k})$.

Let Γ_{kl} denote the subcollection of all $\gamma = (i_1, \dots, i_k) \in \Gamma_{kl}^0$ such that each $i \in \{1, \dots, l\}$ appears two or more times among i_1, \dots, i_k .

Lemma 6. Let g be a BOREL function on $(\mathbb{R}^d \times \mathbb{R})^k$ such that

$$\sum_{l=1}^k \sum_{\gamma \in \Gamma_{kl}^0} |g(Z_\gamma)| P^l(dz) < \infty.$$

Then

$$\mathbf{E} \int \dots \int g(Z_1, \dots, Z_k) M(dZ_1) \dots M(dZ_k) = \sum_{l=1}^{\lfloor k/2 \rfloor} \lambda^l \sum_{\gamma \in \Gamma_{kl}} \int g(Z_\gamma) P^l(dz).$$

Proof. It suffices to consider functions g of the form $g(Z_1, \dots, Z_k) = \prod_1^k \varphi_j(Z_j)$, where $\varphi_j, 1 \leq j \leq k$ are bounded; the general result following by an L^1 approximation argument. For functions of the indicated form the result follows from the formula

$$\mathbf{E} e^{\int \sum_1^k t_j \varphi_j dM} = e^\varphi$$

where

$$\varphi = \lambda \int (e^{\sum t_j \varphi_j} - 1 - \sum t_j \varphi_j) dP.$$

Lemma 7. For each positive integer k there is a positive constant c_k such that

$$\mathbf{E} \left(\int \int_{u \neq t} v s K_h(u-t) M(du, dv) M(ds, dt) \right)^{2k} \leq c_k h^{-2k} \sum_{l=2}^{2k} \lambda^l h^{\lfloor (l+1)/2 \rfloor}, \quad (4.1)$$

for $\lambda > 0$, and $h \in \mathbb{R}$.

Proof. It follows from Lemma 6 that the indicated expectation is a finite linear combination of terms of the form

$$\lambda^l \int \dots \int \prod_m K_h^{v_m}(x_{i_m} - x_{j_m}) y_{i_m}^{v_m} y_{j_m}^{v_m} P(dZ_1) \dots P(dZ_l),$$

where $Z_i = (x_i, y_i), 1 \leq i_m \leq j_m \leq l, v_m > 0$ for all $m, 2 \leq l \leq 2k, \sum_m v_m = 2k$ and each $i \in \{1, \dots, l\}$ appears at least once in the sequence $i_1, j_1, i_2, j_2, \dots$. The above expression can be written as

$$\lambda^l h^{-2k} \int \dots \int \prod_m K^{r_m} \left(\frac{x_{i_m} - x_{j_m}}{h} \right) \mathbf{E} (y^{r_m} | X_{i_m} = x_{i_m}) f(x_1) \dots f(x_l) dx_1 \dots dx_l.$$

Since by assumption $\mathbf{E}(|Y|^{2k})$ is bounded for every k , and K is bounded we have that terms of this form are bounded in absolute value by a constant multiple of $\lambda^l h^{-2k} h^{\lfloor (l+1)/2 \rfloor}$. This implies the result.

Set $N = N(\mathbb{R}^d \times \mathbb{R})$.

Lemma 8. For each positive integer k there is a positive constant c_k such that

$$\begin{aligned} & \mathbf{E} \left[\left(\int_{u+t} \int vs K_h(u-t) (N(du, dv) - NP(du, dv)) (N(dt, ds) - NP(dt, ds)) \right) \right]^{2k} \\ & \leq c_k \left(\lambda + \lambda^{2k} + h^{-2k} \sum_{l=2}^{2k} \lambda^l h^{[l(l+1)/2]} \right), \text{ for } \lambda > 0 \text{ and } h > 0. \end{aligned}$$

Proof. Note that the integral can be expressed as

$$\begin{aligned} & \int_{u+t} \int vs K_h(u-t) M(du, dv) M(dt, ds) \tag{4.2} \\ & - 2(N-\lambda) \int_{u+t} \int vs K_h(u-t) P(du, dv) M(dt, ds) \\ & + (N-\lambda)^2 \int_{u+t} \int vs K_h(u-t) P(du, dv) P(dt, ds). \end{aligned}$$

The third term of (5.2) can be written as

$$\begin{aligned} & (N-\lambda)^2 \int_{u+t} \int K_h(u-t) m(u) m(t) f(u) f(t) dudt \\ & \leq (N-\lambda)^2 \int_{w \neq 0} \int_{-A}^A |K(w) m(wh+t) m(t) f(w+ht) f(t)| dw dt \\ & \leq K(N-\lambda)^2 (2A) \int |m(t) f(t)| dt = K' (N-\lambda)^2 \mathbf{E}|Y|, \end{aligned}$$

for positive constants K, K' . Thus taking expectation to the $2k$ th power, this term is bounded by

$$c_k \mathbf{E} (N-\lambda)^{4k}, \tag{4.3}$$

for some $c_k > 0$. The second term of (4.2) can be expressed as

$$-2(N-\lambda) \int \int sm_h(t) f(t) M(dt, ds)$$

and taking expectation to the $2k$ th power this is

$$\leq 2^{2k} (\mathbf{E} (N-\lambda)^{4k})^{1/2} \{ \mathbf{E} [\int \int sm_h(t) f(t) M(dt, ds)]^{4k} \}^{1/2}.$$

Since $|m_h(t) f(t)|$ is bounded, the above is

$$\leq c_k (\mathbf{E} (N-\lambda)^{4k})^{1/2} \{ \mathbf{E} [\int \int sM(dt, ds)]^{4k} \}^{1/2}.$$

We now use the approximation

$$\int \int sM(dt, ds) = \sum_{i=1}^k s_i (N(J_i) - \lambda P(J_i)) + \varepsilon,$$

where $J_i \subset \mathbb{R}^d \times \mathbb{R}$, $i = 1, \dots, k$, and note that all terms in the sum have mean zero.

Thus using Theorem 1 of WHITTLE (1960), we have the bound

$$c_k (\mathbf{E} (N-\lambda)^{4k})^{1/2} \left(\sum_i s_i^2 \psi_i^2(4k) \right)^{2k}$$

where

$$\begin{aligned} \psi_i^2(4k) &= (\mathbf{E} (N(J_i) - \lambda P(J_i))^{4k})^{2/4k} \\ &\leq (\lambda P(J_i) + \lambda^{2k} P(J_i)^{2k})^{1/2k} \end{aligned}$$

$$\begin{aligned} &\leq 2^{k/2} (\lambda^{2k} P(J_i)^{2k})^{1/2k} \\ &= c_k \lambda P(J_i). \end{aligned}$$

Thus the bound is

$$\begin{aligned} &c_k (\mathbb{E} (N - \lambda)^{4k})^{1/2} \left(\sum_i s_i^2 \lambda P(J_i) \right)^{2k} \tag{4.4} \\ &\leq \lambda^{2k} c_k (\mathbb{E} (N - \lambda)^{4k})^{1/2} \left(\iint s^2 P(dt, ds) \right)^{2k} \\ &= c_k \lambda^{2k} (\mathbb{E} (N - \lambda)^{4k})^{1/2} [\mathbb{E} (Y^2)]^{2k}. \end{aligned}$$

The bound for the first term of (4.2) is given by (4.1) and combining this with (4.3) and (4.4) implies the result.

Lemma 9. For each positive integer k there is a positive constant c_k such that

$$\begin{aligned} &\mathbb{E} \left[\left(\iint_{u \neq t} vs K_h(u-t) (P_n(du, dv) - P(du, dv)) (P_n(dt, ds) - P(dt, ds)) \right)^{2k} \right] \\ &\leq c_k n^{-2k}, \text{ for } n \geq 1 \text{ and } h > 0. \end{aligned}$$

Proof. Let $N_n(dx, dy) = nP_n(dx, dy)$. Then $NP_N(dx, dy)$ determines a Poisson process. Put

$$Z = \iint_{u \neq t} vs K_h(u-t) (N_n(du, dv) - nP(du, dv)) (N_n(dt, ds) - nP(dt, ds));$$

Let $\mathbb{E}(Z^{2k}) = \mu_n$, set $\mu_0 = 0$ and let $R(\lambda)$ be the expectation of Z^{2k} when n is replaced by a Poisson random variable N , having mean λ , N being independent of (X_i, Y_i) , $i \geq 1$. Then

$$R(\lambda) = \sum_n P(N=n) \mu_n = \sum_n \frac{\lambda^n e^{-\lambda}}{n!} \mu_n \tag{4.5}$$

determines a polynomial of degree $2k$ in λ with $R(0) = 0$ by Lemma 7. Also

$$\sum_{j=1}^{2k} \frac{|R^{(j)}(0)|}{j!} \lambda^j \leq c'_k \lambda^{2k}. \tag{4.6}$$

Since, if $\lambda > 1$, then $\lambda^j \leq \lambda^{2k}$, $j = 1, \dots, 2k$ so by choosing $c'_k > \sum_{j=1}^{2k} |R^{(j)}(0)|/j!$ we have (4.6). For $\lambda < 1$, $\sum_{j=1}^{2k} |R^{(j)}(0)|/j! \lambda^j < c'_k$ also, so (4.6) is true $\forall \lambda$. Consequently,

$$\mu_n = \sum_{j=1}^{2k} \frac{n! R^{(j)}(0)}{(n-j)! j!} \leq \sum_{j=1}^{2k} \frac{|R^{(j)}(0)|}{j!} n^j \leq c'_k n^{2k}.$$

This establishes the result. To prove Lemma 3 we first restrict h to belong to H_n . Let $\varepsilon > 0$ be given. Then

$$\begin{aligned} &P \left\{ \sup_{\substack{h \in H_n \\ u \neq t}} \iint vs K_h(u-t) (P(du, dv) - P_n(du, dv)) (P(dt, ds) - P_n(dt, ds)) > \varepsilon J_{nhr} \right\} \\ &\leq \sum_{h \in H_n} \frac{c_k n^{-2k}}{\varepsilon^{2k} (J_{nhr})^{2k}} \leq \frac{n^2 c_k n^{-2k}}{\varepsilon^{2k} (J_{nhr})^{2k}}. \end{aligned}$$

The result follows by considering four cases separately: $h \geq 1$, $n^{-1/(r+1)} \leq h < 1$, $n^{-2} \leq h < n^{-1/r+1}$, and $0 < h < n^{-2}$. To prove the result for $h \in \mathbb{R}^+$, we use an argu-

ment similar to that of Lemma 4 and note that

$$\begin{aligned} & \int \int uv (P_n(dx, du) - P(dx, du)) (P_n(ds, dv) - P(ds, dv)) \\ &= \left(\frac{1}{n} \sum_{i=1}^n Y_i - \mathbb{E}Y \right)^2 \end{aligned}$$

which converges to zero with probability one.

5. An interesting relationship

From Lemma 4 it can be seen that if we restrict our search for the optimal h to the set H_n (which is all that is feasible in practice), then we can relax our assumptions on the moments of Y and require that only moments up to order $(a + 2 + \alpha) \times (b + 1)$ exist where $\#H_n = An^a$ and b is a number related to the bias, and $\alpha > 0$ is an arbitrarily small constant.

It is interesting to note that if $K(x)$ is "smooth", then the constant b of Lemma 1 can be made small while if $K(x)$ is "rough", b is large (see the proof of Lemma 1). Thus we need to assume the existence of more moments for smooth than rough kernels. On the other hand if we increase the cardinality of H_n by taking a larger a , we have to ask for higher moments, i.e., for smoother tails of the distribution of Y , to uniformly, over H_n , approximate $\text{ISE}(\hat{h})$ and $\text{ISE}(h)$.

References

- AKAIKE, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203–217.
- AKAIKE, H. (1974). A new look at the statistical model identification. *I.E.E.E. Trans. Auto. Control.* **19**, 716–723.
- CHEN, K.-W. (1984). Asymptotically optimal selection of a piecewise polynomial estimator of a regression function. Ph. D. Dissertation, Department of Statistics, University of California, Berkeley.
- HALL, P. (1984). Asymptotic Properties of Integrated Square Error and Cross-Validation for Kernel Estimation of a Regression Function. *Zeitschrift für Wahrscheinlichkeitstheorie*, **67**, 175–196.
- HÄRDLE, W. and MARRON, J. (1985a). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.*, **13**, 1465–1481.
- HÄRDLE, W., and MARRON, J. (1985b). Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression. *Biometrika*, **72**, 481–484.
- JOHNSTON, G. (1982). Probabilities of maximal deviations for nonparametric regression function estimation. *J. Mult. Analysis* **12**, 402–414.
- HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13–30.
- MARRON, J. and HÄRDLE, W. (1986). Random Approximations to some measures of accuracy in nonparametric curve estimation, *J. Mult. Analysis*, to appear.
- NADARAYA, E. A. (1964). On estimating regression. *Theor. Prob. Appl.* **9**, 141–142.
- NADARAYA, E. A. (1983). A limit distribution of the square error distribution of non-

- parametric estimators of the regression function. *Zeitschrift für Wahrscheinlichkeitstheorie, U.V.G.* **64**, 37–48.
- RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.*, **12**, 1215–1230.
- STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040–1055.
- STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, **12**, 1285–1297.
- WATSON, G. S. (1964). Smooth regression analysis. *Sankhya, Ser. A.* **26**, 359–372.
- WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theor. Prob. Appl.* **5**, 302–305.

Received March 1985; revised February 1986.

WOLFGANG HÄRDLE
Stanford University and
Institut für Gesellschafts- und
Wirtschaftswissenschaften
Rheinische Friedrich-Wilhelms-Universität
Bonn
Adenauerallee 24–26
D - 5300 Bonn 1
BRD

GABRIELLE KELLY
University College, Cork, Ireland
and
Stanford University

Book Review

J. SPECHT: **APL-Praxis**. Reihe: Leitfäden der angewandten Informatik. B. G. Teubner, Stuttgart 1983, 192 S., DM 22,80.

Das Buch trägt den Untertitel „Demonstration von Sprach- und Stilelementen einer Programmiersprache“. Es wendet sich vor allem an Anwender der EDV mit guten Kenntnissen höherer Programmiersprachen und Programmiererfahrung. Dem Leser sollte nach Möglichkeit eine Anlage zur Verfügung stehen, auf der die Sprache APL implementiert ist. APL, eine Sprache, die erstmals 1962 von IVERSON vorgestellt wurde, zeichnet sich durch wenige vorgegebene Symbole und große Variabilität der möglichen Strukturen aus. Mit der Möglichkeit, gleiche Operationen auf unterschiedlichste Datenstrukturen anzuwenden, selbst Operationen und Zeichen zu definieren, besteht für den erfahrenen APL-Programmierer die Möglichkeit, äußerst kurze und effektive übersichtliche Programme zu erstellen. Die Sprache ist sowohl für die Textverarbeitung, als auch zur Lösung numerischer Aufgaben gut geeignet. Der Autor verbindet die Einführung der Sprachelemente und Vorstellung ihrer Anwendungsmöglichkeiten mit der Darstellung von über 30 Programmbeispielen, anhand derer die Wirkung der Sprachelemente illustriert wird. Die Beispiele umfassen sowohl die Textverarbeitung, Datentransformationen und Spielprogramme, als auch interessante Programme zur Lösung von Gleichungssystemen (APL-Funktion), Regressionsanalyse und linearen Optimierung. Im Anhang des als Lehrbuch konzipierten Buches sind die wichtigsten APL-Symbole, -Befehle und Systemfunktionen zusammengefaßt. Das Buch selbst ist mit Hilfe eines APL-Systems geschrieben.

J. POLZEHL

A NOTE ON PREDICTION VIA ESTIMATION OF THE CONDITIONAL MODE FUNCTION

G. COLLOMB

*Université Paul Sabatier, Laboratoire de Statistique et Probabilités, 118, route de Narbonne,
31062 Toulouse, France*

W. HÄRDLE*

*Johann Wolfgang Goethe-Universität, Fachbereich Mathematik, 6000 Frankfurt, Fed. Rep.
Germany*

S. HASSANI

*Université Paul Sabatier, Laboratoire de Statistique et Probabilités, 118, route de Narbonne,
31062 Toulouse, France*

Received 2 April 1985

Recommended by R.J. Serfling

Abstract: Let $\{(X_i, Y_i)\}_{i \in N} \subset E \times \mathbb{R}$, $E \subset \mathbb{R}^d$ be a strictly stationary process. The conditional density of Y given X is estimated by the kernel method. It is shown that the (empirically determined) mode of the kernel estimate is uniformly (in a compact) convergent to the conditional mode function when the process is Φ -mixing. This result is applied to a strictly stationary time series $\{Z_k\}_{k \in N}$ which is markovian of order q . It is seen that the so-called model predictor of Z_{N+1} from the observed data is converging to the predictor that is based on the full knowledge of the conditional density of Z_{N+1} given $\{Z_1, \dots, Z_N\}$.

AMS Subject Classification: 62G05, 62G20.

Key words: Kernel density estimate; Conditional mode; Φ -mixing process; Modal prediction.

1. Introduction

Let $\{(X_i, Y_i)\}_{i \in N}$ be a stationary Φ -mixing process which is valued in $E \times \mathbb{R}$ with $E \subset \mathbb{R}^d$. Suppose that a stretch of data $\{(X_i, Y_i)\}_{i=1}^n$ has been observed. We are interested in predicting Y from the data for a fixed value of X . In this paper we investigate the prediction of Y by the mode function (assuming that it is uniquely defined)

* Presently at Universität Bonn, 5300 Bonn, Fed. Rep. Germany. Work supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 123 'Stochastische Mathematische Modelle'.

0378-3758/87/\$3.50 © 1987, Elsevier Science Publishers B.V. (North-Holland)

$$\theta(x) = \operatorname{argmax}_{y \in R} f(y|x), \quad x \in E, \tag{1.1}$$

where $f(y|x)$ denotes the conditional density of Y given X . The conditional density is estimated by a kernel estimate $f_n(y|x)$ and the so-called *empirical mode predictor* is defined as the maximum of $f_n(y|x)$ over $y \in \mathbb{R}$.

The kernel method has a long tradition in nonparametric density and regression estimation. Watson (1964), for instance, considered the estimation of the conditional expectation as a predictor for Y and applied this method to some climatological time series data. Following work on nonparametric regression was mainly devoted to the estimation of $E(Y|X=x)$; see Collomb (1981) for a bibliography. However, if the conditional distribution of Y given X has a dominant center peak and a smaller peak far from the center the consideration of the conditional mode function $\theta(x)$ seems to be desirable. Future observations (X, Y) with X around x may tend to scatter around $\theta(x)$ whereas the conditional expectation $E(Y|X=x)$ may fall between these peaks. Is the empirical mode function $\theta_n(x)$ of $f_n(y|x)$ a reasonable nonparametric estimator of $\theta(x)$? Let $\{Z_k\}_{k \in N}$ denote a one-dimensional time series and define X , as the vector of the lag values $(Z_{i-1}, \dots, Z_{i-d})$ and Y_i as Z_i . Is in this framework $\theta_n(Z_n)$, $n = N-d$, a valuable predictor of Z_{N+1} if the data $\{Z_1, \dots, Z_N\}$ have been observed?

The object of this paper is to investigate asymptotic consistency properties of θ_n and to give some insight into situations where θ_n seems to be a useful predictor of Y . We show that the random function

$$\theta_n(x) = \operatorname{argmax}_{y \in R} f_n(y|x) \tag{1.2}$$

converges uniformly over a compact set $\mathcal{C} \subset E$ to the mode function $\theta(x)$. As a consequence to this result we obtain the uniform consistency of a mode based predictor of a strictly stationary time series $\{Z_k\}_{k \in N}$. An analogous result has been shown for M-type predictors and estimators by Robinson (1984), Collomb and Härdle (1986).

Before we proceed to state the result we give an example.

Example. Suppose that an MA(1) process $X_i = \alpha \varepsilon_{i-1} + \varepsilon_i$ has been transmitted and a receiver observes $Y_i = \beta X_i + \eta_i$ where α, β denote real constants and $\{\varepsilon_i\}, \{\eta_i\}$ are independent white noise processes. Suppose that $\eta_i = B_i N_{1i} + (1 - B_i) N_{2i}$, where $\{N_{1i}\}, \{N_{2i}\}$ are independent $N(0, 1), N(c, \sigma^2)$ distributed and $\{B_i\}$ is an independent Bernoulli sequence, i.e. $P(B_i = 0) = 1 - P(B_i = 1) = p < \frac{1}{2}$. This means that the receiver observes the rescaled X process plus with probability $1 - p$ a standard and with probability p a shifted normal random variable. The random variable η has density $(1 - p)\varphi(u) + \sigma^{-1} p\varphi(\sigma^{-1}(u - c))$. Clearly $E(Y|X=x) = \beta x + pc$ which can be made arbitrarily large by choosing c big enough. The conditional mode is a solution (w.r.t. y) to the nonlinear equation

$$\frac{\sigma^2(p-1)(y-\beta x)}{p(y-\beta x-c)} = e^{-(y-\beta x-c)^2/2\sigma+(y-\beta x)^2/2}$$

There is a solution to this equation since the function on the left-hand side has a pole at $\beta x + c$ and approaches $\sigma^2(p-1)/p$ as $y \rightarrow -\infty$. Numerical computations show that $|\beta x - \theta(x)| < 10^{-10}$ for $\sigma = p = c^{-1} = 0.1$ whereas $E(Y|X=x) = \beta x + 1$ in this concrete example.

Quite analogous examples can be constructed for a one-dimensional time series that follows a so-called *noise-replaces-signal* model where we observe $Z_k = B_k S_k + (1 - B_k)N_k$, with $\{S_k\}$ the signal, $\{N_k\}$ the noise and $\{B_k\}$ an independent Bernoulli sequence as above (Martin, 1981).

2. Results

We suppose that the mode function θ satisfies the following uniqueness condition on a compact set $\mathbb{C} \subset E$:

$$\begin{aligned} \forall \varepsilon > 0 \exists \alpha > 0 : (\forall t : \mathbb{C} \rightarrow \mathbb{R}) \sup_{x \in \mathbb{C}} |\theta(x) - t(x)| \geq \varepsilon \\ \Rightarrow \sup_{x \in \mathbb{C}} |f(\theta(x)|x) - f(t(x)|x)| \geq \alpha. \end{aligned} \tag{2.1}$$

Assume also that for the mixing coefficients $\{\Phi_m\}_{m \in \mathbb{N}}$ associated with $\{(X_i, Y_i)\}_{i \in \mathbb{N}}$ (Billingsley (1968), p. 190 ff.) the following holds for an increasing sequence $(m_n)_{n \in \mathbb{N}}$ to be specified below:

$$\exists A < \infty : n\Phi_{m_n}/m_n \leq A, \quad 1 \leq m_n \leq n, \quad n \in \mathbb{N}. \tag{2.2}$$

The kernel estimates of $f(y|x)$ are defined by

$$f_n(y|x) = \frac{\sum_{i=1}^n h^{-1} K_1((y - Y_i)/h) K_0((x - X_i)/h)}{\sum_{j=1}^n K_0((x - X_j)/h)} \tag{2.3}$$

where $h = h_n$ is a positive sequence of bandwidths tending to zero, as n goes to infinity, and $K_0(K_1)$ are kernel functions on $\mathbb{R}^d(\mathbb{R})$. More precisely,

$$\begin{aligned} K_j \text{ are bounded, integrating to one, } \quad j = 0, 1, \\ |z|^d K_0(z) \rightarrow 0 \quad \text{as } |z| \rightarrow \infty. \end{aligned} \tag{2.4}$$

The first result can now be stated.

Theorem. *Suppose that (2.1)–(2.2) hold. Assume that*

(a) *$f(x, y)$ is uniformly continuous on $\mathbb{C} \times \mathbb{R}$, where \mathbb{C} is an ε -neighborhood of \mathbb{C} in E and*

$$\exists \delta > 0 : \int f(x, y) dy \geq \delta \quad \forall x \in \mathbb{C},$$

f being the density of the distribution of (X_1, Y_1) ;

(b) K_j are Hölder-continuous, i.e.

$$\exists \gamma > 0 \exists L < \infty: |K_j(u) - K_j(v)| \leq L|u - v|^\gamma, \quad j = 0, 1,$$

and K_1 has bounded support,

(c) the sequence $\{m_n\}$ from (2.2) satisfies with h_n ,

$$\frac{nh_n^{d+1}}{m_n \log n} \rightarrow \infty, \quad n \rightarrow \infty; \tag{2.5}$$

(d) either

(d1) $\exists s > 0: E|Y|^s < \infty$ and $\exists \lambda > 0: \sum_{n=1}^{\infty} h_n^\lambda < \infty$, or

(d2) Y is bounded.

Then, as $n \rightarrow \infty$,

$$\sup_{x \in \mathbb{G}} |\theta_n(x) - \theta(x)| \rightarrow 0 \quad \text{almost completely.}$$

Remark. The condition (2.5) can be replaced by

$$\frac{nh_n^{d+1}}{(\log n)^2} \rightarrow \infty, \quad n \rightarrow \infty,$$

when $\{(X_i, Y_i)\}_{i \in \mathbb{N}}$ is geometrically Φ -mixing, for instance when this process is Markovian and Doeblin's condition (see Doob (1953), p. 256) is satisfied (see also Collomb (1984)).

This theorem applies to the above mentioned prediction problem of a real strictly stationary and Φ -mixing time series $\{Z_n\}_{k \in \mathbb{N}}$ according to the following argument.

If the law of the process would be known we could certainly construct a mode based predictor of Z_{n+1} from the observed data $\{Z_1, \dots, Z_N\}$. This predictor depends only on (Z_{N-d+1}, \dots, Z_N) when the process $\{Z_k\}_{k \in \mathbb{N}}$ is Markovian of order d and if in addition e.g. the Doeblin condition (Doob (1953), p. 256) is satisfied, the process is Φ -mixing so that $\theta(Z_{n-d+1}, \dots, Z_n)$ can be estimated by the empirical mode predictor $\theta_n(Z_{N-d+1}, \dots, Z_N)$, $n = N - d$, defined from (1.2) by

$$X_i = (Z_i, \dots, Z_{i+d-1}), \quad Y_i = Z_{i+d}, \quad i = 1, \dots, n. \tag{2.6}$$

This process is also Φ -mixing and we have therefore the following corollary.

Corollary. *If the assumptions of the theorem are satisfied by (2.6), then, as $N \rightarrow \infty$, $|\theta_n(Z_{N-d+1}, \dots, Z_N) - \theta(Z_{N-d+1}, \dots, Z_N)| \mathbf{1}((Z_{N-d+1}, \dots, Z_N) \in \mathbb{G}) \rightarrow 0$ almost completely.*

The empirical mode predictor based on a kernel estimate $f_n((y|x)$ of $f(y|x)$ is a reasonable predictor in nonparametric time-series analysis for two reasons. For the

kernel estimate there exists fast algorithms (Silverman (1982)) that allow several exploratory evaluations of $f_n(y|x)$ before a final decision is made. Second, the bandwidth h is an interpretable tuning parameter, that allows easy understanding of what the estimate is actually doing to the data. We leave it open here how the bandwidth should be selected in a practical situation. This will be considered in a forthcoming paper.

The rest of this section is devoted to the proof of the theorem. The lemmata needed for the proof are proven in Section 3 and Section 4.

Lemma 1. *The uniform convergence of $f_n(\cdot | \cdot)$ over $\mathbb{R} \times \mathbb{G}$ implies the uniform convergence of $\theta_n(x)$ over \mathbb{G} .*

The proof of Lemma 1 is in Section 3. In the following two lemmata the numerator and the denominator of $f_n(y|x)$ are considered separately.

Decompose

$$f_n(y|x) = \frac{f_{1n}(x, y)}{f_{0n}(x)}$$

where

$$f_{1n}(x, y) = n^{-1} h^{-d-1} \sum_{i=1}^n K_1\left(\frac{(y - Y_i)}{h}\right) K_0\left(\frac{(x - X_i)}{h}\right)$$

and

$$f_{0n}(x) = n^{-1} h^{-d} \sum_{i=1}^n K_0\left(\frac{(x - X_i)}{h}\right)$$

is the well-known Rosenblatt-Parzen density estimator. The basic idea is to show that f_{1n} and f_{0n} are separately uniformly consistent.

Lemma 2. *Under the assumptions of the theorem,*

$$\sup_{x \in \mathbb{G}} \sup_{y \in \mathbb{R}} |f_{1n}(x, y) - E f_{1n}(x, y)| \rightarrow 0 \quad a.c., \quad n \rightarrow \infty, \tag{2.7}$$

$$\sup_{x \in \mathbb{G}} |f_{0n}(x) - E f_{0n}(x)| \rightarrow 0 \quad a.c., \quad n \rightarrow \infty. \tag{2.8}$$

The proof of Lemma 2 is in Section 3.

Let $f_0(x)$ be the marginal density of X . The bias term is controlled by:

Lemma 3. *Under the assumptions of the theorem,*

$$\sup_{x \in \mathbb{G}} \sup_{y \in \mathbb{R}} |E f_{1n}(x, y) - f(x|y) E f_{0n}(x)| \rightarrow 0, \quad n \rightarrow \infty, \tag{2.9}$$

$$\sup_{x \in \mathbb{G}} |E f_{0n}(x) - f_0(x)| \rightarrow 0, \quad n \rightarrow \infty. \tag{2.10}$$

The proof of Lemma 3 is in Section 4.

The theorem follows now from Lemma 1 and the following inequality:

$$\begin{aligned} & \sup_{x \in \mathbb{G}} \sup_{y \in \mathbb{R}} |f_n(y|x) - f(y|x)| \\ & \leq \left\{ \sup_{x \in \mathbb{G}} \sup_{y \in \mathbb{R}} |f_{1n}(x, y) - E f_{1n}(x, y)| + M \delta^{-1} \sup_{x \in \mathbb{G}} |f_{0n}(x) - E f_{0n}(x)| \right. \\ & \quad \left. + \sup_{x \in \mathbb{G}} \sup_{y \in \mathbb{R}} |E f_{1n}(x, y) - f(y|x) E f_{0n}(x)| \right\} / \sup_{x \in \mathbb{G}} f_{0n}(x) \end{aligned}$$

where

$$M = \max \left\{ \sup_{x \in \mathbb{G}} f_0(x), \sup_{x \in \mathbb{G}} \sup_{y \in \mathbb{R}} f(x, y) \right\}$$

and δ is a lower bound for f_0 on \mathbb{G} . By Lemma 2 the first two terms in the inequality above tend to zero almost completely. If

$$\exists \bar{\delta} > 0: \sum_n P \left\{ \inf_{x \in \mathbb{G}} f_{0n}(x) \leq \bar{\delta} \right\} < \infty, \tag{2.11}$$

then by Lemma 3 the proof of the theorem will be complete. Claim (2.11) follows from

$$\inf_{x \in \mathbb{G}} f_{0n}(x) \geq \inf_{x \in \mathbb{G}} E f_{0n}(x) - \sup_{x \in \mathbb{G}} |f_{0n}(x) - E f_{0n}(x)|$$

and statement (2.10) of Lemma 3, since with δ as above

$$E f_{0n}(x) = h^{-d} \int K_0 \left(\frac{x-u}{h} \right) f_0(u) du \geq \frac{1}{2} \delta$$

for n large enough. This completes the proof of the theorem. The corollary follows immediately.

3. Proof of Lemma 1 and Lemma 2

For the rest of this paper we will write \sup_y instead of $\sup_{y \in \mathbb{R}}$ and \sup_x or \inf_x instead of $\sup_{x \in \mathbb{G}}$ or $\inf_{x \in \mathbb{G}}$.

By definition of $\theta_n(x)$ and $\theta(x)$ we have

$$\begin{aligned} & |f(\theta_n(x)|x) - f(\theta(x)|x)| \\ & \leq |f_n(\theta_n(x)|x) - f(\theta_n(x)|x)| + |f_n(\theta_n(x)|x) - f(\theta(x)|x)| \\ & \leq \sup_y |f_n(y|x) - f(y|x)| + \left| \sup_y f_n(y|x) - \sup_y f(y|x) \right| \\ & \leq 2 \sup_y |f_n(y|x) - f(y|x)|. \end{aligned}$$

The uniform uniqueness condition (2.1) yields that for all $\varepsilon > 0$ there exists a $\beta > 0$ such that

$$P\left(\sup_x |\theta_n(x) - \theta(x)| \geq \varepsilon\right) \leq P\left(\sup_x \sup_y |f_n(y|x) - f(y|x)| \geq \beta\right).$$

This proves Lemma 1.

We now come to the proof of Lemma 2. We only show (2.7); the statement (2.8) can be deduced in analogy.

Put, $f_{1n}(x, y) - E f_{1n}(x, y) = \sum_{i=1}^n \Delta_i$ where

$$\begin{aligned} \Delta_i = & n^{-1} h^{-d-1} K_1\left(\frac{(y - Y_i)}{h}\right) K_0\left(\frac{(x - X_i)}{h}\right) \\ & - E\left[n^{-1} h^{-d-1} K_1\left(\frac{(y - Y_i)}{h}\right) K_0\left(\frac{(x - X_i)}{h}\right)\right]. \end{aligned}$$

Define

$$\begin{aligned} \delta_n = & 2n^{-1} h^{-d-1} \bar{K}, & d_n = & 2n^{-1} \bar{K}M, & D_n = & n^{-2} h^{-d-1} \bar{K}^2 M, \\ \beta = & (8\bar{K})^{-1}, & B = & 6\beta \bar{K}^2 M, & \alpha_n = & \frac{\beta n h^{d+1}}{m_n}, \end{aligned}$$

with

$$\bar{K} = \max\left\{\sup_y |K_1(y)|, \sup_y |K_0(x)|, 1\right\}$$

and M as before an upper bound for $f(x, y) \vee f(x)$.

We show first that there are constants $a, b > 0$ such that

$$\sup_x \sup_x P(|f_{1n}(x, y) - E f_{1n}(x, y)| > \varepsilon) \leq a \exp\left\{-\frac{b n h^{d+1}}{m_n}\right\}. \tag{3.1}$$

An application of Collomb (1984), p. 449 yields that the left-hand side in (3.1) is bounded by

$$C_{m_n} \exp\left(-\frac{n h^{d+1} T(\varepsilon, m_n)}{m_n}\right)$$

where

$$T(\varepsilon, m) = \beta(\varepsilon - B(m^{-1} + 16\bar{\Phi}_m m^{-1})) \quad \text{with} \quad \bar{\Phi}_m = \sum_{i=1}^m \phi_i$$

and

$$C_m = 2 \exp(3\sqrt{\varepsilon} n \bar{\Phi}_m m^{-1}).$$

Condition (2.2) gives that

$$\exists n_1 \in \mathbb{N} : \forall n \in \mathbb{N}, n \geq n_1, \quad m_n \geq m \text{ with } m \geq \frac{4B}{\varepsilon} \text{ and } \bar{\Phi}_m m^{-1} \geq \frac{\varepsilon}{(64B)},$$

which shows that

$$T(\varepsilon, m_n) \geq \frac{1}{2}\beta\varepsilon.$$

Now (3.1) follows with $a = 2 \exp(3A \sqrt{\varepsilon})$, $b = \frac{1}{2}\beta\varepsilon$.

Next we show that

$$\sup_x \sup_y |f_{1n}(x, y) - E f_{1n}(x, y)| \rightarrow 0 \quad \text{a.c.}, \quad n \rightarrow \infty. \tag{3.2}$$

By assumption K_0, K_1 are assumed to be Hölder-continuous, so by (2.3),

$$|K_0(u_0)K_1(u_1) - K_0(v_0)K_1(v_1)| \leq \bar{K}L|u_1 - v_1|^\gamma + \bar{K}L|u_0 - v_0|^\gamma$$

and therefore

$$|f_{1n}(x, y) - f_{1n}(x^*, y^*)| \leq \bar{K}L h^{-(d+1)}|x - x^*|^\gamma + \bar{K}L h^{-(d+1)}|y - y^*|^\gamma. \tag{3.3}$$

Now put

$$a_n = h^{-\mu^{-1}}, \quad b_n = h^\alpha, \tag{3.4}$$

where μ, α are positive constants to be determined later on. The range of Y is then decomposed,

$$\sup_x \sup_y |f_{1n}(x, y) - E f_{1n}(x, y)| = U_n + V_n,$$

$$U_n = \sup_x \sup_{|y| \leq a_n} |f_{1n}(x, y) - E f_{1n}(x, y)|,$$

$$V_n = \sup_x \sup_{|y| \geq a_n} |f_{1n}(x, y) - E f_{1n}(x, y)|,$$

and the compact set $\mathbb{E} \times [-a_n, a_n]$ is covered by a finite net of balls with radius b_n :

$$\{B(x_j; b_n) \times B(y_j; b_n); j = 1, \dots, l_n\}.$$

By construction, $l_n = O(a_n b_n^{-d-1}) = O(h^{-(ad+d+\mu^{-1})})$ and

$$U_n \leq \max_{1 \leq j \leq l_n} \sup_{x \in B(x_j; b_n)} \sup_{y \in B(y_j; b_n)} \{|\psi_1(x, y) - E\psi_1(x, y)| + |f_{1n}(x_j; y_j) - E f_{1n}(x_j, y_j)|\}$$

with $\psi_1 = f_{1n} - E f_{1n}$. Now, by (2.3),

$$\max_{1 \leq j \leq l_n} \sup_{x \in B(x_j; b_n)} \sup_{y \in B(y_j; b_n)} |\psi_1(x, y)| \leq C_1 h^{-(d+1)-\gamma+\alpha\gamma}$$

which yields

$$U_n \leq 2C_1 h^{-(d+1+\gamma)+\alpha\gamma} + T_n, \quad T_n = \max_{1 \leq j \leq l_n} |f_{1n}(x_j, y_j) - E f_{1n}(x_j, y_j)|.$$

By (3.1) and the construction of the covering net,

$$P(T_n > \varepsilon) \leq C_2 h^{-\alpha(d+1)-\mu^{-1}} \exp\left(\frac{-bnh^{d+1}}{m_n}\right) \leq C_3 n^{-2}$$

by assumption of the theorem. Therefore $U_n \rightarrow 0$ a.c., $n \rightarrow \infty$. Note that (3.2) is shown in the case where Z is assumed to be bounded. The term V_n is now estimated,

$$V_n \leq W_n + E W_n$$

where $W_n = \sup_x \sup_y f_{1n}(x, y)$.

By the compactness of the support of K_1 we have

$$K_1\left(\frac{Y_1 - Y_i}{h}\right) \leq \tilde{K} \mathbf{1}(|Y_i| > \frac{1}{2}a_n)$$

and therefore

$$W_n \leq \tilde{K}^2 (nh^{d+1})^{-1} \sum_{i=1}^n \mathbf{1}(|Y_i| > \frac{1}{2}a_n),$$

which gives with $P(|Y| > \frac{1}{2}a_n) \leq (2a_n^{-1})^t E|Y|^t$ and Markov's inequality

$$P(|W_n| > \varepsilon) \leq \varepsilon^{-1} E W_n \leq C_3 h^{-(d+1)+t/\mu}.$$

The choice of $\mu = t/(\lambda + (d+1))$ shows that $V_n \rightarrow 0$ a.c., $n \rightarrow \infty$, and completes thus the proof of Lemma 2.

4. Proof of Lemma 3

We only show (2.9), the statement (2.10) follows by similar techniques. We have

$$|E f_{1n}(x, y) - f(y|x) E f_{0n}(x)| \leq |E f_{1n}(x, y) - f(x, y)| + f(y|x) |E f_{0n}(x) - f_0(x)|.$$

and prove

$$\sup_x \sup_y |E f_{1n}(x, y) - f(x, y)| \rightarrow 0, \quad n \rightarrow \infty. \tag{4.1}$$

Since $f(x, y)$ is assumed to be uniformly continuous on $\mathbb{C} \times \mathbb{R}$,

$$\forall \alpha > 0 \exists \delta_\alpha > 0: |z - y| \leq \delta_\alpha \text{ and } |x - u| \leq \delta_\alpha \Rightarrow |f(x, y) - f(u, z)| \leq \alpha. \tag{4.2}$$

Write

$$E f_{1n}(x, y) - f(x, y) = n^{-1} \sum_{i=1}^n S_{ni}(x, y)$$

with

$$S_{ni}(x, y) = E h^{-(d+1)} K_1\left(\frac{(y - Y_i)}{h}\right) K_0\left(\frac{(x - X_i)}{h}\right) - f(x, y).$$

Fix $\alpha > 0$. Then terms S_{ni} are estimated as follows:

$$\begin{aligned} |S_{ni}(x, y)| &\leq h^{-(d+1)} \int \left| K_1\left(\frac{(y-z)}{h}\right) K_0\left(\frac{(x-u)}{h}\right) \right| |f(x, y) - f(z-u)| dz du \\ &\leq \alpha + 2M \int_{\{t > \delta_\alpha/h; |v| > \delta_\alpha/h\}} |K_1(t)K_0(v)| dt dv \\ &\leq \alpha + 2M \beta_\alpha(h) \end{aligned}$$

where $\beta_\alpha(h)$ is a positive function independent of (x, y) by (4.2) tending to zero as $h \rightarrow 0$. This shows (4.1).

References

- Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.
- Collomb, G. (1981). Estimation non-parametrique de la regression: Revue bibliographique. *Internat. Statist. Rev.* **49**, 75-93.
- Collomb, G. (1984). Proprietés de convergence presque complète du predicteur a noyan. *Z. Wahrsch. Verw. Geb.* **66**, 441-460.
- Collomb, G. and W. Härdle (1986). Strong uniform convergence rates in robust nonparametric time series analysis and prediction: Kernel regression estimation from dependent observations. *Stochastic Process. Appl.* **23**, 77-89.
- Doob, J.L. (1953). *Stochastic processes*. Wiley, New York.
- Martin, R.D. (1981). Robust methods for time series. In: *Applied Time Series Analysis II*. Academic Press, New York.
- Robinson, P.M. (1984). Robust nonparametric autoregression. In: J. Franke, W. Härdle and D. Martin, Eds., *Robust and Nonlinear Time Series Analysis*. Springer, Berlin-New York.
- Silverman, B.W. (1982). Kernel density estimation using the fast Fourier transform. *Appl. Statist.* **31**, 93-97.
- Watson, G.S. (1964). Smooth regression analysis, *Sankhya A* **26**, 359-372.

AN EFFECTIVE SELECTION OF REGRESSION VARIABLES
WHEN THE ERROR DISTRIBUTION IS INCORRECTLY SPECIFIED*

WOLFGANG HÄRDLE

(Received Sept. 9, 1985; revised Apr. 24, 1986)

Summary

An asymptotically efficient selection of regression variables is considered in the situation where the statistician estimates regression parameters by the maximum likelihood method but fails to choose a likelihood function matching the true error distribution. The proposed procedure is useful when a robust regression technique is applied but the data in fact do not require that treatment. Examples and a Monte Carlo study are presented and relationships to other selectors such as Mallows' C_p are investigated.

1. Introduction and results

Suppose that $Y=(Y_1, \dots, Y_n)'$ is a random vector of n observations with mean $\mu=(\mu_1, \dots, \mu_n)'$ and assume that each component μ_i is associated with a covariate x_i , such that $\mu_i=\langle x_i, \beta \rangle$. Assume that the parameter vector is infinite dimensional; then at most n elements of β can be estimated on the basis of the observations. Suppose that a certain likelihood function, not necessarily matching the true error distribution, has been selected by the statistician, and that parameter estimates $\hat{\beta}(p)$ in a finite dimensional submodel p have been obtained by the maximum likelihood principle. The regression curve μ_i at x_i is then estimated by $\hat{\mu}_i(p)=\langle x_i, \hat{\beta}(p) \rangle$ and a loss $L_n(p)=\|\mu - \hat{\mu}(p)\|^2$ is suffered. We shall consider an efficient model selection procedure that asymptotically minimizes the loss $L_n(p)$ over a certain class of finite dimensional models of increasing dimension.

This paper completes earlier papers in various ways. Breiman and Freedman [3], Shibata [12] considered the problem of selecting regres-

* Research supported by Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 123 "Stochastische Mathematische Modelle" and AFOSR Contract No. F49620 82 C 0009.

Key words and phrases: Variable selection, regression analysis, robust regression, model choice.

sion variables when the true error distribution is known to be Gaussian and derived selectors that are equivalent to ours in this case. In the setting of least squares estimation Li ([8]) gave conditions for asymptotic efficiency of model choice procedures based on cross validation, FPE and other means.

Schrader and Hettmansperger ([11]) considered a robust analysis of variance based on Huber's M -estimates and propose a likelihood ratio type of test for testing between finite dimensional submodels. This viewpoint was also taken by Ronchetti ([10]) who derived a "robust model selection" procedure that is related to ours.

A mismatch of a chosen likelihood function and of a true error distribution can happen in the case when the statistician applies a robust regression estimation technique (Huber [6]) but the data is in fact Gaussian. One may also think of the reverse situation that a Gaussian maximum likelihood estimate (i.e., the least squares estimate) is computed but the true error distribution is different, possibly a long tailed outlier generating distribution.

The general idea of regression model selection procedures is to minimize a penalized form of the residual sum of squares. For instance Akaike's AIC ([1]) penalizes the dimensionality of the model with the penalty constant 2. The AIC-score is asymptotically optimal in the case of Gaussian errors and the least square estimation technique as was shown by Shibata ([12]). In the case of a mismatch between the true error distribution and the chosen likelihood function the proposed regression model procedure has a similar structure but the penalty constant is changed depending on the type of mismatch. This can be heuristically described as follows. If there are outliers in the data generated by a long tailed error distribution and AIC is applied based on a Gaussian maximum likelihood estimate the data will be overfitted since the model selection procedure will fit the outliers. The model selection procedure to be presented below penalizes more a high dimensional model since the penalty constant is bigger than 2. On the other hand if the data is indeed Gaussian and a robust regression technique is applied, the penalty constant will be less than 2. An example of this kind is considered in Section 5 where a simulation study is presented.

In the simple case that the data is Gaussian and the statistician chooses a Gaussian likelihood function, then our model selection procedure is equivalent to Mallows' C_p (see [9], Section 4). This entails equivalence to many other selectors such as FPE, AIC, GCV, as was shown by Li ([8]).

We will assume that the control variables $x_i = (x_{i1}, x_{i2}, \dots)'$, $i = 1, \dots, n$ and the parameter vector $\beta = (\beta_1, \beta_2, \dots)'$ are in l_2 . The model

can then be written as

$$Y = X\beta + e = \mu + e$$

where $e = (e_1, \dots, e_n)'$ is the vector of the independent observation errors having distribution F with density f and $X' = (x'_1 \ x'_2 \ \dots \ x'_n)$ is considered as a linear operator from l_2 to R^n . By $p = (p_1, p_2, \dots, p_{k(p)})$ we denote a finite dimensional submodel with parameter

$$\beta'(p) = (0, \dots, \beta_{p_1}, 0, \dots, \beta_{p_2}, 0, \dots, \beta_{p_{k(p)}}, 0, \dots).$$

The statistician chooses a likelihood function ρ of which he believes to represent the true error distribution, and estimates the parameters in a submodel p by maximizing the approximate likelihood function

$$\prod_{i=1}^n \rho(Y_i - x'_i(p)\beta(p))$$

where $x'_i(p) = (0, \dots, x_{ip_1}, 0, \dots, x_{ip_2}, 0, \dots, x_{ip_{k(p)}}, 0, \dots)$. Call this maximum likelihood estimate $\hat{\beta}(p)$, and define $\phi(u) = -(d/du) \log \rho(u)$, $\gamma = E_F \phi^2(e) / (E_F \phi'(e))^2$, $B_n = X'X$ and let P_n be a family of models p . A possible selection rule for choosing a model $p \in P_n$ could be defined by $W_n^{(1)}(p) = -\|\hat{\mu}(p)\|^2 + 2\gamma k(p) + \|\mu\|^2$, since

$$(1.1) \quad W_n^{(1)}(p) - L_n(p) = -\|\hat{\mu}(p)\|^2 + 2\gamma k(p) + \|\mu\|^2 - \|\hat{\mu}(p) - \mu\|^2 \\ = 2\{\gamma k(p) - \langle \hat{\beta}(p) - \beta, \hat{\beta}(p) \rangle_{B_n}\}$$

where $\langle u, v \rangle_{B_n}$ denotes the bilinear form $u'B_nv$ for vectors $u, v \in l_2$. It will be shown that the last term in (1.1) is tending to a constant uniformly over the model class P_n . Then minimizing $W_n^{(1)}(p)$ over P_n will be the same task, at least asymptotically, as minimizing $L_n(p)$. However, $W_n^{(1)}$ cannot be computed directly from the data since it depends on the unknown regression curve μ . But note that the last term in $W_n^{(1)}(p)$ is independent of the model p . We will therefore define

$$W_n(p) = -\|\hat{\mu}(p)\|^2 + 2\gamma k(p)$$

as the score function that is to be minimized over P_n . The problem of simultaneously estimating γ from the data, in order to make W_n completely data driven is considered in Section 4.

Remark 1. If the statistician is in the happy situation of knowing f , then he will choose $\rho \equiv f$. If f is symmetric, then by partial integration

$$I(F) = E_F \phi^2 = \int \phi^2 f = \int (f'/f)^2 f = \int f' \phi = \int \phi' f = E_F \phi'$$

and therefore the constant γ reduces to $(E_F \phi')^{-1} = I(F)^{-1}$, the Fisher-

information number in a location family with density f .

We will use the concept of asymptotic efficiency as in Li [8], Shibata [12] and Stone [13]: A selected \hat{p} is called *asymptotically optimal* if, as $n \rightarrow \infty$

$$(1.2) \quad \frac{L_n(\hat{p})}{\inf_{p \in P_n} L_n(p)} \xrightarrow{p} 1 .$$

The following condition on ϕ will be needed.

CONDITION 1. The function ϕ is centered i.e., $E_F \phi(e) = 0$ and twice differentiable with bounded second derivative. We furthermore assume that $E_F [q^{-1}(\phi'(e) - q)]^{2N} < \infty$ for some positive integer N and $q = E_F \phi'(e) > 0$.

The estimates $\hat{\beta}(p)$ will be compared with the Gauss-Markov estimates in the model p based on the (unobservable) pseudodata $\tilde{Y}_i = \mu_i + \tilde{e}_i$, $\tilde{e}_i = \phi(e_i)/q$. Define $X(p)$ as the (n, p) matrix containing the nonzero control variables in model p and assume that $B_n(p) = X'(p)X(p)$ has full rank $k(p)$. Then the Gauss-Markov estimate of μ based on the pseudodata $\tilde{Y} = (\tilde{Y}_1, \dots, \tilde{Y}_n)'$ is defined as $\tilde{\mu}(p) = H_n(p)\tilde{Y}$, where $H_n(p) = X(p) \cdot B_n^{-1}(p)X'(p)$ denotes the hat matrix in model p . The loss for the Gauss-Markov estimate is $\tilde{L}_n(p) = \|\tilde{\mu}(p) - \mu\|^2$ which will be approximately $L_n(p)$ as will be seen later on. The speed at which the cardinality of P_n is allowed to grow is controlled by

CONDITION 2. There exists a positive integer N such that with $\tilde{R}_n(p) = E_F \tilde{L}_n(p)$

$$\sum_{p \in P_n} \tilde{R}_n(p)^{-N} \rightarrow 0, \quad \text{as } n \rightarrow \infty .$$

Let $h(p)$ be the largest diagonal element of the hat matrix $H_n(p)$. The speed of $h(p)$ relative to $\tilde{R}_n(p)$ is controlled by

CONDITION 3.

$$\sup_{p \in P_n} h(p)\tilde{R}_n(p) \rightarrow 0, \quad \text{as } n \rightarrow \infty .$$

Remark 2. It follows from Condition 3 that

$$k^2(p)/n \rightarrow 0, \quad \text{as } n \rightarrow \infty ,$$

since $\tilde{R}_n(p) = \gamma k(p) + \|\mu - \mu(p)\|^2$, $\mu(p) = H_n(p)\mu$. This should be seen as an analogue of the necessary condition, $p^2/n \rightarrow 0$, that can be found in Huber ([7], p. 166). Conditions 2 and 3 imply also

$$(1.3) \quad \sum_{p \in P_n} h(p)^N \rightarrow 0 .$$

Remark 3. If ϕ is bounded, as is assumed in a robust regression analysis, Condition 2 can be weakened. It is seen from the proofs that in this case Bernstein's inequality could be used instead of Whittle's ([14], Theorem 2). Condition 2 could be weakened to $\sum_{p \in P_n} \exp(-C\tilde{R}_n(p)) \rightarrow 0$, for some $C > 0$. In the robust estimation of location so-called re-descending ϕ -functions have been introduced (see Andrews et al. [2]). A direct application of such a ϕ -function which is zero outside some interval is not possible, since points close to infinity also solve the likelihood equation. The usual approach is to couple such estimators to consistent estimators with monotone ϕ -functions as is described for instance in Härdle [5], p. 173. A similar procedure seems possible in the setting described here but we did not investigate it.

Condition 1 could be weakened to piecewise twice differentiable ϕ -functions, but as Huber [7] we decided to state a stronger condition in order to have a simpler outline of the proof.

Denote by \hat{p} a model $p \in P_n$ that minimizes $W_n(p)$ over P_n . The main result is as follows.

THEOREM. *Under Conditions 1-3, \hat{p} is asymptotically optimal.*

The rest of the paper is organized in five sections. In Section 2 the theorem above is shown, in Section 3 we give a variety of examples that satisfy our Conditions 1-3, and in Section 4 the estimation of γ and the relation to other model selection procedures is investigated. In Section 5 a Monte Carlo example of the lemmas that are needed in showing the asymptotic optimality.

2. Proof of Theorem

In the proof of Theorem, the following lemmas will be used.

LEMMA 2.1. *Under the conditions of Theorem, for all $\varepsilon > 0$*

$$P \left\{ \sup_{p \in P_n} \|\hat{\mu}(p) - \tilde{\mu}(p)\|^2 / \tilde{R}_n(p) > \varepsilon \right\} \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

LEMMA 2.2. *Under the conditions of Theorem, for all $\varepsilon > 0$*

$$P \left\{ \sup_{p \in P_n} |\tilde{L}_n(p) - \tilde{R}_n(p)| / \tilde{R}_n(p) > \varepsilon \right\} \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

LEMMA 2.3. *Under the conditions of Theorem, for all $\varepsilon > 0$*

$$P \left\{ \sup_{p \in P_n} |\gamma k(p) - (\tilde{\mu}(p) - \mu)' \tilde{\mu}(p) + \mu' \tilde{e}| / \tilde{R}_n(p) > \varepsilon \right\} \rightarrow 0, \\ \text{as } n \rightarrow \infty.$$

Recall that the Gauss-Markov estimate based on the pseudodata \tilde{Y}

is $\tilde{\beta}(p) = B_n^{-1}(p) X'(p) \tilde{Y}$. The crossterm in (1.1) will be approximated by a corresponding crossterm based on the linearized estimates $\tilde{\beta}(p)$.

$$(2.1) \quad \begin{aligned} & \langle \hat{\beta}(p) - \beta, \hat{\beta}(p) \rangle_{B_n} - \langle \tilde{\beta}(p) - \beta, \tilde{\beta}(p) \rangle_{B_n} \\ &= \|\tilde{\mu}(p) - \hat{\mu}(p)\|^2 + \langle \hat{\beta}(p) - \tilde{\beta}(p), \tilde{\beta}(p) - \beta(p) \rangle_{B_n} \\ & \quad + \langle \hat{\beta}(p) - \tilde{\beta}(p), \beta(p) \rangle_{B_n} + \langle \tilde{\beta}(p) - \beta, \hat{\beta}(p) - \tilde{\beta}(p) \rangle_{B_n}. \end{aligned}$$

By Lemma 2.1, the first term is of lower order than $\tilde{R}_n(p)$ uniformly over P_n , the second term is bounded by the Cauchy-Schwarz inequality and then Lemmas 2.1 and 2.2 are applied. The third term is handled by formula (6.2), given in the proof of Lemma 2.1, by setting $a = \beta(p)$, $\eta = \hat{\beta}(p)$. The fourth term is handled as the second term. Suppose that

$$(2.2) \quad \sup_{p, p' \in P_n} \left| \frac{(W_n(p) - W_n(p')) - (L_n(p) - L_n(p'))}{L_n(p) + L_n(p')} \right| = o_p(1),$$

and let p^* denote a minimizer of $L_n(p)$ over P_n . Then by (2.2) with probability greater than $1 - \epsilon$,

$$\frac{W_n(\hat{p}) - W_n(p^*) - (L_n(\hat{p}) - L_n(p^*))}{L_n(\hat{p}) + L_n(p^*)} \geq -\epsilon.$$

By the definition of \hat{p} , $W_n(\hat{p}) - W_n(p^*) \leq 0$, therefore,

$$-(L_n(\hat{p}) - L_n(p^*)) \geq -\epsilon(L_n(\hat{p}) + L_n(p^*))$$

$$L_n(p^*)(1 + \epsilon) \geq L_n(\hat{p})(1 - \epsilon)$$

$$1 \geq \frac{L_n(p^*)}{L_n(\hat{p})} \geq \frac{1 - \epsilon}{1 + \epsilon}$$

which shows that (1.2) holds, i.e., \hat{p} is asymptotically optimal. Formula (2.2) follows by observing that

$$\begin{aligned} & \frac{(W'_n(p) + \mu' \tilde{e} - L_n(p))}{L_n(p)} \\ &= \frac{2(\gamma k(p) - \langle \hat{\beta}(p) - \beta, \hat{\beta}(p) \rangle_{B_n} + \mu' \tilde{e})}{\tilde{R}_n(p)} \cdot \frac{\tilde{L}_n(p)}{L_n(p)} \cdot \frac{\tilde{R}_n(p)}{\tilde{L}_n(p)}. \end{aligned}$$

The first factor is tending to zero in probability, uniformly over P_n by Lemma 2.3 and formula (2.1). The two other factors tend to one in probability, uniformly over P_n , by Lemmas 2.1 and 2.2.

3. Examples

We start with a reformulation of Condition 2 in the case of hier-

archical model sequences, i.e., $P_n = \{(1), (1, 2), \dots, (1, 2, \dots, p_n)\}$ with p_n tending to infinity. In this case Condition 2 follows for $N=2$ from

CONDITION 2'.

$$\inf_{p \in P_n} \tilde{R}_n(p) \rightarrow \infty, \quad \text{as } n \rightarrow \infty.$$

We slightly abuse notation by writing j for $(1, \dots, j)$. Then Condition 2 follows from $\tilde{R}_n(j) = \gamma j + \|\mu(p) - \mu\|^2$ and

$$\begin{aligned} \sum_{p \in P_n} \tilde{R}_n(p)^{-2} &= \sum_{j=1}^{J_n} \tilde{R}_n(j)^{-2} + \sum_{j=J_n+1}^{P_n} \tilde{R}_n(j)^{-2} \\ &\leq J_n \{ \inf_{p \in P_n} \tilde{R}_n(p) \}^{-2} + \gamma^{-2} \sum_{j=J_n+1}^{\infty} j^{-2} \rightarrow 0, \end{aligned}$$

if J_n tends to infinity slowly enough. In the following examples we assume that P_n represents a hierarchical model sequence. The following lemma, which is due to Shibata [12], is useful in checking Condition 2'.

LEMMA 3.1. *Assume that with a positive divergent sequence $\{c_n\}$ the linear operator $c_n^{-1}B_n$ converges weakly to a nonsingular operator $B: l_2 \rightarrow l_2$, such that every $p \times p$ principal submatrix $B(p)$ has full rank p for all $p > 0$. If β has infinitely many nonzero coordinates, then Condition 2' holds and p^* diverges to infinity, as $n \rightarrow \infty$.*

Are the conditions of Theorem fulfilled for typical examples? We check conditions in examples given by Shibata [12].

Example 1. Consider the polynomial regression on the interval $[0, 1)$. Here

$$X_{ij} = \left(\frac{i-1}{n} \right)^{j-1}, \quad i=1, \dots, n, \quad j=1, 2, \dots$$

and

$$Y_i = \sum_{j=1}^{\infty} \left(\frac{i-1}{n} \right)^{j-1} \beta_j + e_i, \quad i=1, \dots, n$$

are observed.

Condition 1 is model independent and is an assumption about the error distribution. Condition 2' is satisfied via Lemma 3.1 (set $c_n = n$). It remains to check Condition 3. The symmetric matrix $B_n^{-1}(p)$ has a spectral decomposition

$$B_n^{-1}(p) = \Gamma_n \Lambda_n \Gamma_n',$$

where $A_n = \text{diag}(\lambda_1(p), \dots, \lambda_p(p))$ and $\Gamma_n = (\gamma_1, \dots, \gamma_p)$, $\gamma_j = (\gamma_{1j}, \dots, \gamma_{pj})'$ the i -th normalized eigenvector of $B_n^{-1}(p)$. Lemma 3.1 insures that $\lambda_{\min}(p)$ the smallest eigenvalue of $n^{-1}B_n(p)$ is bounded above zero by a constant C . Therefore each diagonal element h_i of $H_n(p)$ can be estimated by

$$h_i = \sum_{j=1}^p \sum_{k=1}^p x_{ij} x_{ik} \left(\sum_{l=1}^p \gamma_{lj} \gamma_{lk} \lambda_l(p) \right) \leq \sum_{l=1}^p \lambda_l(p) \left(\sum_{j=1}^p x_{ij}^2 \right) \left(\sum_{k=1}^p \gamma_{lk}^2 \right) \\ \leq p \lambda_{\max}(p) \sum_{j=1}^p x_{ij}^2 \leq p^2 \lambda_{\min}(p)^{-1} \leq C^{-1} p^2/n .$$

So Condition 3 is fulfilled if we ask for

$$(3.1) \quad \sup_{1 \leq p \leq p_n} p^2 \bar{R}_n(p)/n \rightarrow 0 .$$

A necessary condition is $p^3/n \rightarrow 0$ which is slightly stronger than Huber's conditions ([7]).

Example 2. Consider the following representation of the regression curve

$$\mu_i = \sum_{j=1}^{\infty} \beta_j \cos(\pi(j-1)(i-1)/nj) .$$

Here the observations are taken at $x=0, n^{-1}, \dots, ((n-1)/n)$. As in the example above Condition 2 is satisfied by Lemma 3.1, setting $c_n = n/2$. Condition 3 is satisfied by similar arguments as above if we assume that (3.1) holds.

Example 3. Consider the robust M -estimation of location at different units x_j . Observations are taken repeatedly at p_n different units and n/p_n observations are taken at the point x_j , $j=1, \dots, p_n$. Assume that $E_{\mathcal{F}} \phi(e) = 0$, then Condition 1 is satisfied if ϕ, ϕ' are bounded. Shibata ([12], p. 51) shows that Condition 2 is satisfied if the vectors of the control-variables (x_1, \dots, x_{p_n}) are linearly independent. Condition 3 can be checked as above.

4. Other methods and estimation of γ

There are a variety of other model selection methods, most of which were shown to be equivalent to Mallows' C_p . We therefore compare our method with C_p only. For simplicity, we work with the linearized estimate $\tilde{\mu}(p)$ based on the pseudodata \tilde{Y} . Mallows' score function ([9]) reads

$$C_p(p) = \|\tilde{Y} - \tilde{\mu}(p)\|^2 + 2\gamma k(p) \\ = \|\tilde{e}\|^2 + \tilde{L}_n(p) + 2\tilde{e}'(I_n - H_n(p))\mu + 2\{\gamma k(p) - \tilde{e}'H_n(p)\tilde{e}\} .$$

The first term is independent of p , the third and the last term vanish uniformly over model classes P_n , as can be seen in the next section. This shows that a model selected by C_p is asymptotically optimal.

It can now be seen that $W_n(p)$ has a similar structure.

$$\begin{aligned} W_n(p) &= -\|\tilde{\mu}(p)\|^2 + 2\gamma k(p) \\ &= -\|\tilde{\mu}(p) - \mu\|^2 - \|\mu\|^2 - 2(\tilde{\mu} - \mu)'(\mu - \tilde{\mu}) - 2(\tilde{\mu} - \mu)' \tilde{\mu} + 2\gamma k(p) \\ &= \tilde{L}_n(p) + 2\gamma k(p) - 2\tilde{e}' H_n(p) \tilde{e} + 2\tilde{e}'(I - H_n(p))\mu + 2\mu' \tilde{e} - \|\mu\|^2. \end{aligned}$$

Here the last two terms are independent of the model. The remaining terms are identical to those in Mallows' C_p , which shows that $W_n(p)$ is equivalent to C_p .

It could be argued that the score function that is proposed here is not so reasonable in a practical application since the constant γ is unknown to the statistician. However, if the constant γ can be consistently estimated (independent of p) then the score function based on an estimated γ is also asymptotically optimal. A consistent estimate $\hat{\gamma}_n$ of γ is provided, for instance, by

$$n^{-1} \sum_{i=1}^n \phi^2(\hat{e}_i(p_n)) / \left(n^{-1} \sum_{i=1}^n \phi'(\hat{e}_i(p_n)) \right)^2$$

where $\hat{e}_i(p_n)$ denote residuals from a fit with a deterministic model p_n , increasing in magnitude as $n \rightarrow \infty$. A Taylor expansion and the Cauchy-Schwarz inequality show that $\hat{\gamma}_n \xrightarrow{p} \gamma$, as $n \rightarrow \infty$.

5. A simulation study

A small Monte Carlo study was carried out to study the behavior of $W_n(p)$ when applied to some real data. The data were generated according to Example 1 (Section 3) with $\mu_i = \sin(z_i)$, $z_i = -\pi + 2((i-1)/n)\pi$, $n=100, 200$ and normal Gaussian error. The original data for $n=100$ is shown in Figure 1. The data do not directly suggest a certain type of model, to a model selection procedure seems to be appropriate. Some of the observations (around $x \approx 1$) look a little bit isolated so that an applied statistician might want to apply a robust regression technique. In this example we have chosen a ϕ -function that is linear in $[-2, 2]$ and a constant outside. Such a ϕ -function does not satisfy Condition 2 as it stands but as it was argued in Remark 3 the results also hold for this specific choice of a nondifferentiable ϕ -function. Straightforward calculations show that $\gamma = 1.274$. The values of $L_n(p)$ and $W_n(p) = \|\mu(p)\|^2 + 2 \cdot 1.274 \cdot p$ (in the hierarchical model case) are presented in Table 1 for $n=100$ and $n=200$. For both sample sizes the p that minimizes $L_n(p)$ is 4. The selected \hat{p} for $n=100$ was $\hat{p}=4$ and for $n=200$ it was

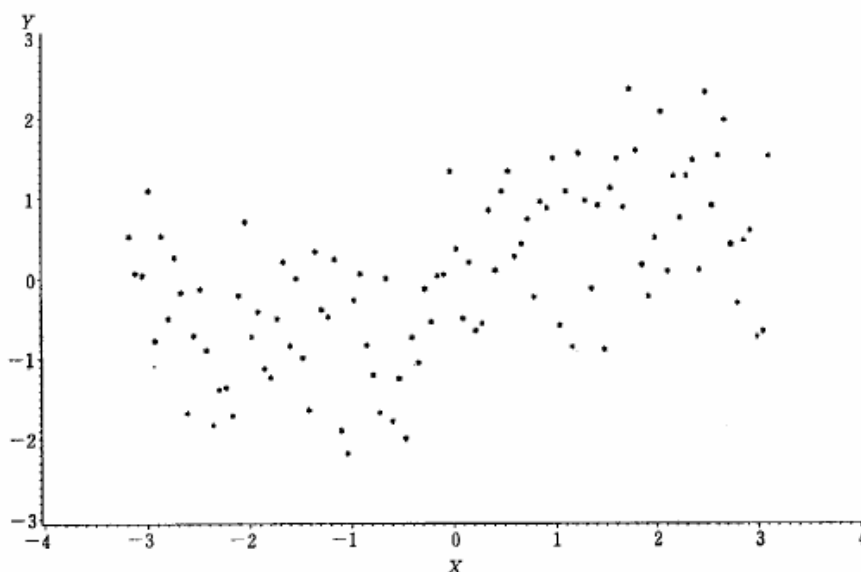


Fig. 1. Original data, $n=100$.

Table 1.

$n=100$			$n=200$		
p	$L_n(p)$	$W_n(p)$	p	$L_n(p)$	$W_n(p)$
2	20.5891	-24.715	2	43.0369	-71.85
3	22.6782	-24.977	3	43.0157	-69.34
4	3.0268	-37.416	4	5.3095	-122.71
5	3.1176	-35.006	5	6.1518	-120.93
6	3.8345	-32.620	6	6.9489	-123.51
7	3.9246	-30.165	7	7.4540	-121.48
8	5.0409	-28.835	8	8.6594	-120.01
9	5.0816	-26.327	9	12.2348	-121.03
10	11.3078	-30.014	10	12.7236	-118.97
11	11.3094	-27.467	11	12.7411	-116.44
12	11.4650	-25.075	12	12.8033	-113.95
13	14.2398	-25.302	13	12.9057	-111.51
14	14.9177	-23.432	14	12.9229	-108.98
15	15.4464	-21.412	15	18.3032	-111.81
16	15.4861	-18.904	16	18.3735	-109.33
17	15.5287	-16.399	17	20.0331	-108.44
18	15.5287	-13.851	18	20.0331	-105.89
19	15.5287	-11.303	19	20.0331	-103.35
20	15.5288	-8.755	20	24.0869	-104.85

$\hat{p}=6$. The shape of the functions $L_n(p)$ and $W_n(p)$ are given in Figures 2 and 3. The parameters for $n=100$ were $\hat{\beta}_1(\hat{p})=0.1017$, $\hat{\beta}_2(\hat{p})=0.9979$, $\hat{\beta}_3(\hat{p})=-0.0006$, $\hat{\beta}_4(\hat{p})=-0.1108$ and thus quite close to the true parameters $\beta_1=\beta_3=0$, $\beta_2=1$, $\beta_4=1/6$. Although for $n=200$ score $W_n(p)$ misses the order of the model that minimizes $L_n(p)$ the fit will not be too bad as the difference of $L_n(4)$ and $L_n(6)$ suggested. In Figure 4 the true curve μ_i and the fitted model curve $\hat{\beta}_i(\hat{p})$, $p=4$ are shown. The

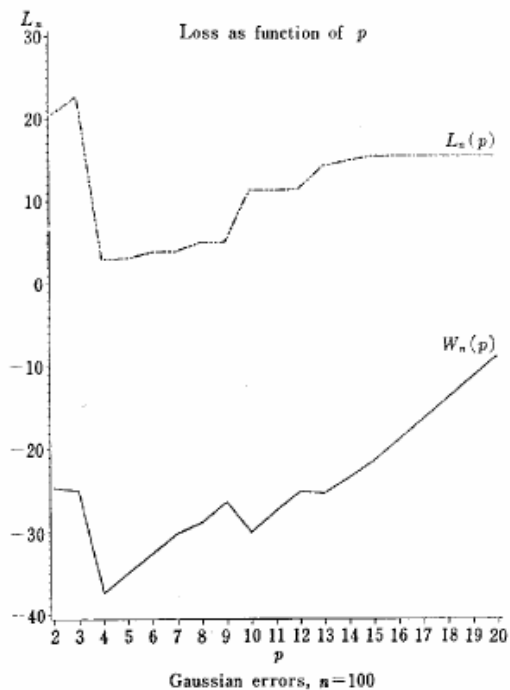


Fig. 2. L_n and W_n .

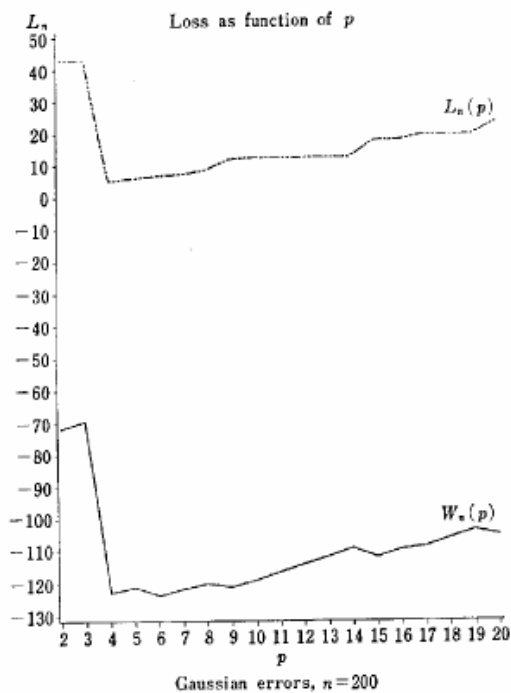


Fig. 3. L_n and W_n .

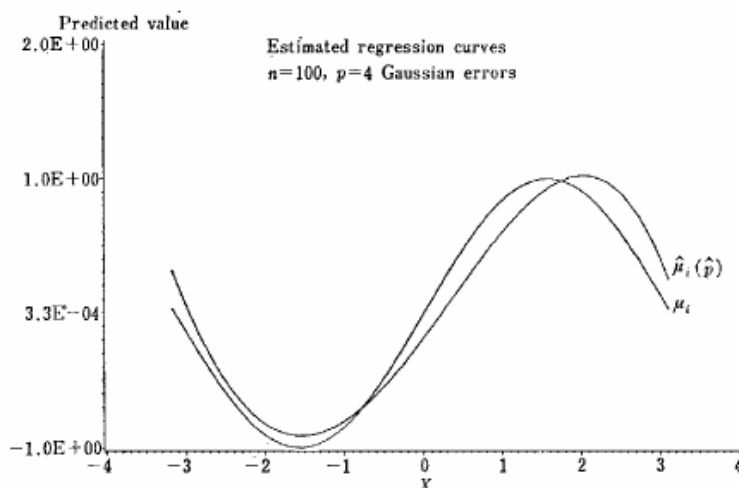


Fig. 4. μ_i and $\hat{\mu}_i(\hat{p})$.

fit was constructed for $n=100$ and with the parameters $\hat{\beta}_j(\hat{p})$, $j=1, \dots, 4$ as given above. We also studied the case $n=100$, the minimum p for L_n was 6 and this \hat{p} was also selected by W_n . This supports the theory that suggests an increasing \hat{p} as n tends to infinity.

6. Proofs

In this section we give the proofs of Lemmas 2.1-2.3. The proof of Lemma 2.1 follows a related proof in Huber [7], Section 7.4. Similar ideas were used by Cox ([4]) who considered M -type smoothing splines. In order to simplify notation we will consider the hierarchical case only, i.e., the model “ p ” is identified with “ $(1, 2, \dots, k(p)), k(p)=p$ ”. Furthermore we assume without loss of generality that the coordinate system in the p -dimensional subspace of the first p components has been chosen so that $X'(p)X(p)=I_p$. Consider the mapping $\Phi: R^p \rightarrow R^p$, $\Phi_k(\eta) = -q^{-1} \sum_{i=1}^n \phi\left(Y_i - \sum_{j=1}^p x_{ij} \eta_j\right) x_{ik}$, $k=1, \dots, p$ where $\eta = (\eta_1, \dots, \eta_p)' \in R^p$. A zero (with respect to η) of Φ will be compared with a zero of $\phi_k(\eta) = \eta_k - \sum_{i=1}^n (\mu_i + \tilde{e}_i) x_{ik}$, where $\tilde{e}_i = \phi(e_i)/q$, $q = E_F \phi'(e)$. The zero of $\phi_k(\eta)$ is the least squares estimate $\tilde{\beta}(p) = X(p) \tilde{Y}$ based on the pseudodata \tilde{Y} . Consider an arbitrary normalized vector $a \in R^p$, $\|a\|=1$. A Taylor expansion of Φ , using Condition 1, leads to

$$\begin{aligned} \sum_{k=1}^p a_k (\Phi_k(\eta) - \phi_k(\eta)) &= -q^{-1} \sum_{k=1}^p a_k \sum_{i=1}^n (\phi'(e_i) - q) \left(\sum_{j=p+1}^{\infty} x_{ij} \beta_j \right) x_{ik} \\ &\quad - q^{-1} \sum_{k=1}^p a_k \sum_{i=1}^n (\phi'(e_i) - q) \sum_{j=1}^p x_{ij} x_{ik} (\beta_j - \eta_j) \end{aligned}$$

$$\begin{aligned}
 & -\frac{1}{2}q^{-1} \sum_{k=1}^p a_k \sum_{i=1}^n \phi'' \left(e_i + \nu \left(\sum_{j=1}^{\infty} x_{ij} \beta_j - \sum_{j=1}^p x_{ij} \eta_j \right) \right) \\
 & \qquad \qquad \qquad \cdot \left(\sum_{j=1}^{\infty} x_{ij} \beta_j - \sum_{j=1}^p x_{ij} \eta_j \right)^2 x_{ik} \\
 & = T_{1,n}(p) + T_{2,n}(p) + T_{3,n}(p), \quad \nu \in (-1, 1).
 \end{aligned}$$

We will now show that each of these terms uniformly vanishes over P_n , in the sense that

$$(6.1) \quad \sup_{p \in P_n} T_{\alpha,n}(p) / \tilde{R}_n^{1/2}(p) \xrightarrow{p} 0, \quad \alpha=1, 2, 3$$

for all (η, a) in the set

$$\mathcal{F}_n = \bigcap_{p \in P_n} \left\{ (\eta, a) : \sum_{i=1}^n \left(\sum_{j=1}^p x_{ij} (\beta_j - \eta_j) \right)^2 \leq K \tilde{R}_n(p), \|a\|=1 \right\}.$$

Define for $i=1, \dots, n$

$$\begin{aligned}
 V_i &= q^{-1}(\phi'(e_i) - q), \\
 B_{i,n}(p) &= \sum_{j=p+1}^{\infty} x_{ij} \beta_j, \\
 s_i &= \sum_{k=1}^p a_k x_{ik}, \\
 A_{i,n}(p) &= \sum_{j=1}^p x_{ij} (\beta_j - \eta_j).
 \end{aligned}$$

Note that

$$\|s\|^2 = \sum_{i=1}^n s_i^2 = \sum_{i=1}^n \left(\sum_{k=1}^p a_k x_{ik} \right)^2 = \|X(p)a\|^2 = \|a\|^2 = 1.$$

The first term $T_{1,n}(p)$ is estimated as follows.

$$\begin{aligned}
 & P \left\{ \sup_{p \in P_n} |T_{1,n}(p)| / \tilde{R}_n^{1/2}(p) > \varepsilon \right\} \\
 & \leq \sum_{p \in P_n} \varepsilon^{-2N} E \left\{ |T_{1,n}(p)|^{2N} / \tilde{R}_n^{2N}(p) \right\} \\
 & = \sum_{p \in P_n} \varepsilon^{-2N} E \left\{ \left| \sum_{i=1}^n s_i B_{i,n}(p) V_i \right|^{2N} / \tilde{R}_n^{2N}(p) \right\}.
 \end{aligned}$$

Applying Condition 1 and Whittle's inequality ([14], Theorem 2), this term is bounded by

$$\begin{aligned}
 & \sum_{p \in P_n} C_1 \varepsilon^{-2N} \left(\sum_{i=1}^n s_i^2 B_{i,n}^2(p) \right)^N / \tilde{R}_n^{2N}(p) \\
 & \leq C_1 \varepsilon^{-2N} \sum_{p \in P_n} \left(\max_{i=1}^n S_i^2 \right)^N \left(\sum_{i=1}^n B_{i,n}^2(p) \right)^N / \tilde{R}_n^{2N}(p), \quad C_1 > 0.
 \end{aligned}$$

Recall that $\tilde{R}_n(p) = \gamma p + \|(I_n - H_n(p))\mu\|^2 \geq \sum_{i=1}^n B_{i,n}^2(p)$. Conditions 2, 3 (see

formula (1.3) in Remark 2) and the simple inequality $s_i^2 \leq \sum_{j=1}^p x_{ij}^2 \sum_{k=1}^p a_k^2 = h_i(p)$, where $h_i(p)$ denotes the i -th diagonal element of the hat matrix $H_n(p)$, imply that (6.1) holds for $\alpha=1$.

The second term is estimated similarly. We omit some details. If $(\eta, \alpha) \in \mathcal{F}_n$, then as above

$$\begin{aligned} & P \left\{ \sup_{p \in P_n} |T_{2,n}(p)| / \tilde{R}_n^{1/2}(p) > \varepsilon \right\} \\ & \leq C_1 \varepsilon^{-2N} \sum_{p \in P_n} h(p)^N \left(\sum_{i=1}^n \Delta_{i,n}^2(p) \right)^N / \tilde{R}_n(p)^N \\ & \leq C_1 \gamma^{-N} K^N \varepsilon^{-2N} \sum_{p \in P_n} h(p)^N . \end{aligned}$$

Now apply Conditions 2 and 3 as described in formula (1.3) in Remark 2.

The third term, involving the second derivative of ϕ is bounded by

$$\frac{1}{2} q^{-1} \sup \phi'' \max_{i=1}^n |s_i| \sum_{i=1}^n (B_{i,n}(p) + \Delta_{i,n}(p))^2 .$$

If $(\eta, \alpha) \in \mathcal{F}_n$ we obtain that with a constant C_2

$$|T_{3,n}(p)| / \tilde{R}_n^{1/2}(p) \leq C_2 h(p)^{1/2} \tilde{R}_n(p)^{1/2} ,$$

which tends to zero by Condition 3.

Altogether we have shown that

$$(6.2) \quad \sup_{p \in P_n} \left| \sum_{k=1}^p a_k (\Phi_k(\eta) - \Psi_k(\eta)) \right| / \tilde{R}_n^{1/2}(p) \xrightarrow{p} 0$$

for all $(\eta, \alpha) \in \mathcal{F}_n$. This entails that for all η in the set

$$(6.3) \quad \begin{aligned} \mathcal{Q}_n &= \left\{ \eta \in R^p : \sup_{p \in P_n} \sum_{i=1}^n \left(\sum_{k=1}^p (\eta_k - \beta_k) X_{ik} \right)^2 / \tilde{R}_n(p) \leq K \right\} , \\ & \sup_{p \in P_n} \|\Phi(\eta) - \Psi(\eta)\| / \tilde{R}_n^{1/2}(p) \xrightarrow{p} 0 . \end{aligned}$$

Condition 2 and bounds on higher moments as above imply that with probability greater than $1 - \delta$,

$$(6.4) \quad \sup_{p \in P_n} \|\tilde{\mu}(p) - \mu(p)\|^2 / \tilde{R}_n(p) < \gamma + \varepsilon .$$

This shows that $\tilde{\beta}(p) \in \mathcal{Q}_n$ with high probability. Note that

$$(6.5) \quad \begin{aligned} \|\Phi(\eta) - \eta\| &= \|\Phi(\eta) - \Psi(\eta) + (\beta(p) - \tilde{\beta}(p)) + \beta(p)\| \\ &\leq \|\Phi(\eta) - \Psi(\eta)\| + \|\beta(p) - \tilde{\beta}(p)\| + \|\beta(p)\| . \end{aligned}$$

From formula (6.3) we know that the first term vanishes asymptotically.

From (6.4) we conclude that for K big enough

$$\sup_{p \in P_n} \|\beta(p) - \tilde{\beta}(p)\| / \tilde{R}_n^{1/2}(p) \leq \frac{1}{2} K^{1/2}.$$

Certainly the third term can be made less than $K^{1/2}p^{1/2}/2$. Thus the function $\eta \rightarrow \eta - \Phi(\eta)$ has a fixed point η^* in the compact, convex set \mathcal{Q}_n . Since this fixed point is necessarily a zero of Φ , it is seen that $\hat{\beta}(p)$ is in \mathcal{Q}_n with probability greater than $1 - \delta$. Substituting $\hat{\beta}(p)$ into equation (6.3) shows that Lemma 2.1 holds.

Lemma 2.2 is seen by the following equation.

$$\tilde{L}_n(p) - \tilde{R}_n(p) = \tilde{e}' H_n(p) \tilde{e} - \gamma k(p).$$

Condition 2 implies that

$$\sup_{p \in P_n} \left| \|H_n(p) \tilde{e}\|^2 - \gamma k(p) \right| / \tilde{R}_n(p) \xrightarrow{p} 0$$

which shows Lemma 2.2.

Lemma 2.3 follows similarly observing that

$$\langle \beta - \tilde{\beta}(p), \tilde{\beta}(p) \rangle_{B_n} = \tilde{e}' (I_n - H_n(p)) \mu - \tilde{e}' H_n(p) \tilde{e} - \tilde{e}' \mu.$$

Acknowledgement

I would like to thank Charles J. Stone, Ritei Shibata and Johanna Behrens for helpful discussions. The suggestions of an anonymous referee helped to improve the presentation of the paper.

JOHANN-WOLFGANG-GOETHE-UNIVERSITÄT, WEST GERMANY AND
UNIVERSITY OF NORTH CAROLINA*

REFERENCES

- [1] Akaike, H. (1970). Statistical predictor identification, *Ann. Inst. Statist. Math.*, **22**, 203-217.
- [2] Andrews, D. F., Bickel, P. J., Hampel, F., Huber, P., Rogers, W. and Tukey, J. W. (1972). *Robust Estimation of Location*, Princeton University Press, Princeton.
- [3] Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation, *J. Amer. Statist. Ass.*, **78**, 131-136.
- [4] Cox, D. (1983). Asymptotics for M -type smoothing splines, *Ann. Statist.*, **11**, 530-551.
- [5] Härdle, W. (1984). Robust regression function estimation, *J. Multivariate Anal.*, **14**, 169-180.
- [6] Huber, P. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo, *Ann. Statist.*, **1**, 799-821.
- [7] Huber, P. (1981). *Robust Statistics*, Wiley, New York.
- [8] Li, K. C. (1984). Asymptotic optimality for C_p, C_1 , cross-validation and generalized cross-validation: Discrete index set, Manuscript.
- [9] Mallows, C. (1973). Some comments on C_p , *Technometrics*, **15**, 661-675.

* Now at Universität Bonn, West Germany.

- [10] Ronchetti, E. (1985). Robust model selection in regression, *Statist. Prob. Letters*, **3**, 21-23.
- [11] Schrader, R. M. and Hettmansperger, T. P. (1980). Robust analysis of variance based upon a likelihood ratio criterion, *Biometrika*, **67**, 93-101.
- [12] Shibata, R. (1981). An optimal selection of regression variables, *Biometrika*, **68**, 45-54.
- [13] Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates, *Ann. Statist.*, **12**, 1285-1297.
- [14] Whittle, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables, *Theor. Prob. Appl.*, **3**, 302-305.

XploRe

A COMPUTING ENVIRONMENT FOR EXPLORATORY REGRESSION AND DENSITY SMOOTHING

Wolfgang HÄRDLE

Rechts- und Sozialwissenschaftliche Fakultät, Wirtschaftstheoretische Abteilung II, Universität Bonn
Adenauerallee 40-42, D-5300 Bonn 1, FRG
Faculty of Science and Technology, Kelo University, Yokohama, Japan

Abstract

XploRe is a graphically oriented interactive system for exploratory regression and density smoothing. Various nonparametric smoothing techniques for low and high dimensions are implemented. Higher dimensional response surfaces can be approximated by means of additive models: Alternating Conditional Expectations (ACE); Projection Pursuit Regression (PPR); Recursive Partitioning Regression Trees (RPR). XploRe uses the object oriented approach and makes extensive use of the inheritance principle. It is written in TURBO PASCAL and runs on IBM PC/AT, XT or compatibles with MS-DOS.

Zusammenfassung

XploRe ist ein graphisch ausgerichtetes System für explorative Regression und Dichteschätzung. Verschiedene nichtparametrische Dichteschätzungen für niedrige und hohe Dimensionen sind implementiert. Höher dimensionale Regressionsoberflächen kann man mit Hilfe folgender Modelle approximieren: Alternating Conditional Expectations (ACE); Projection Pursuit Regression (PPR); Recursive Partitioning Regression Trees (RPR). XploRe bedient sich des objekt-orientierten Ansatzes und macht ausführlichen Gebrauch vom "inheritance principle". Geschrieben ist es in TURBO PASCAL und ist mit MS-DOS auf IBM PC/AT, XT oder kompatiblen Geräten zu benutzen.

*How we think about data analysis
is strongly influenced by the computing
environment in which the analysis is done.*

McDONALD and PEDERSON (1986):

I. WHY AN INTERACTIVE COMPUTING ENVIRONMENT?

XploRe is an interactive system for analyzing various kinds of data smoothing operations. More precisely, XploRe is a graphically oriented computing environment for exploratory regression and density smoothing techniques with sophisticated data management tools. Data can be rotated, brushed, masked, labeled, transformed and smoothed. Higher dimensional data clouds can be analyzed by means of additive models: Projection Pursuit Regression; Recursive Partitioning Regression Trees; Alternating Conditional Expectations or Average Derivative Estimation. A personal computer, like an IBM PC/AT, XT or compatibles (under MS-DOS) is sufficient for the use of XploRe.

A personal computer or a workstation provides the need of a statistical analysis to improvise alternative ways of interpretation on the spot. A typical scenario in nonparametric regression smoothing is the determination of the best fitting polynomial to a given two-dimensional data set. There are methods which determine the order of a polynomial in an asymptotic sense (SHIBATA (1981)) but it is interesting to see how the fit changes, when the order of the polynomial varies in a small neighborhood around the "best fit". In order to see qualitative changes even for "small variations" of the polynomial order it is necessary to have an interactive computing device.

MCDONALD and PEDERSON (1986) point out that the computing environment strongly influences the analysis: If a statistician performs an exploratory or experimental data mining in low or high dimensions, he does in fact a special kind of programming work. An interactive computing environment that is designed for the special needs of experimental programming of data smoothing is therefore most appropriate. To see why this experimental programming cannot be performed with batch oriented systems consider the following analysis cycle (Figure 1.1). A typical round through this cycle is the following. First, a smoothing operation (e.g. response surface estimation) is performed based on a specific method and smoothing parameter. Second, the fit and residuals are examined for certain features (e.g. remaining structure in the residual pattern). In a third step one evaluates the effect and impact of detected features on the fitted curve (e.g. how seriously an outlier influences the smooth). The last step in a round might be to compare the current smooth with other fits, possibly stemming from alternative, parametric models. Such a round through the analysis cycle may be repeated many more times. It seems to be impossible to perform effectively this analysis cycle in a batch oriented computing environment. Another szenario inside such an analysis cycle is the masking operation on some data points (e.g. outliers). We might want to put aside some of the points and run a certain manipulation with the remaining data in order to study the effect of the left-out points. Batch oriented systems most badly serve this need for interactive decision making since one would basically have to write an additional program for identifying the points which are to be left out. In an interactive computing environment one would mark those points by mouse clicks for instance.

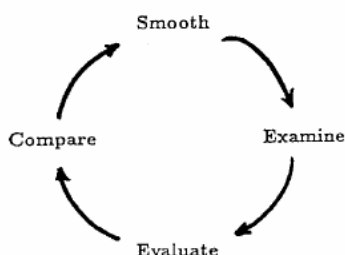


Figure 1.1: Typical analytic cycle

The design of XploRe meets the desiderata for *improvisational programming* by extensive use of interactive graphical methods (mouse oriented selection and identification; pull down menus). Moreover, it supports the user with a set of utilities for masking, brushing, labeling and even rotating of data. XploRe is an *open system* which is written in TURBO PASCAL. It is basically a framework awaiting more "soft work" that enhances the capabilities. Its construction has been influenced by similar systems like S (BECKER and CHAMBERS (1984)) or DINDE (OLDFORD and PETERS (1985)): XploRe uses the object oriented approach and makes extensive use of the inheritance principle to be described below. A detailed description of the functions and procedure to install user written code is given in AERTS and HOLTSBERG (1987).

This paper is organized as follows. Section 2 describes the objects, structure and the basic primitives of XploRe, in particular the *workunit* objects and the inheritance of attributes. Section 3 is devoted to the description of the display functions. In section 4 the user interface is explained via a construction of a *running median* primitive. Section 5 gives an overview over additive models for fitting high dimensional data. Section 6 gives details about the availability of the software.

Inheritance avoids redundant specification of information and simplifies modification, since information that is common is defined in, and need be changed in, only one place.

OLDFORD and PETERS (1985)

II. OBJECTS AND INHERITANCE

XploRe uses the *object oriented* approach, i.e. the basic elements that are dealt with are structures of simpler variable types and manipulations of data is made solely by reference to those structures (objects). For the purposes of data smoothing we found the following four objects sufficient: *vector*, *workunit*, *picture*, *text*. *Vectors* are the simplest objects, they contain a real data array of variable length. *Workunits* are collections of pointers to vectors and may include display and mask attributes. *Picture* objects are viewports, defining the location and tic marks of the axes in 2D or 3D views. *Texts* are sequences of text lines. The above objects can be *created/deleted, activated/deactivated, read/written, manipulated, displayed*.

Moreover, objects can *inherit* certain properties. Workunits can inherit display attributes, such as linestyle or symbols. They can also inherit a *mask*. A mask is a vector of integer classification numbers, including the option to

show points as "invisible". Picture objects inherit the location of the axes and the ticmarks on the screen. Suppose, for example, that a workunit is displayed in a certain picture object. The picture object may then be manipulated by rotation of the pointcloud or by clipping certain parts of the data. These viewport information is inherited by the picture object. If another projection of the same workunit or a different workunit is shown in the same picture object, we would obtain (even after clearing the screen) the same viewport aspect as for the first pointcloud. The inheritance principle thus simplifies overlaying and comparing several curves into the same viewport and hence the same scale. Since display attributes or masks are part of the workunit object, different objects can be distinguished quite easily without using an extra scrapbook aside the computer.

The notion of workunits seems to allow a flexible analysis of several data vectors at a time. Suppose that one wants to analyse a three dimensional data set consisting of vectors X, Y, Z . Workunit wu-one could consist of the vectors X, Y , another wu-two could point to all three vectors. When displaying wu-one one could have detected some interesting points, which one interactively has marked with the classification number "7". Other observations might have been given the mask "invisible". Earlier one might have decided to see then points as stars (except those that leave mask "7"). If wu-two wants to be shown with squares and needles pointing into the (X, Z) plane one can think of the following graphical presentation of the two workunits (Figure 2.1).

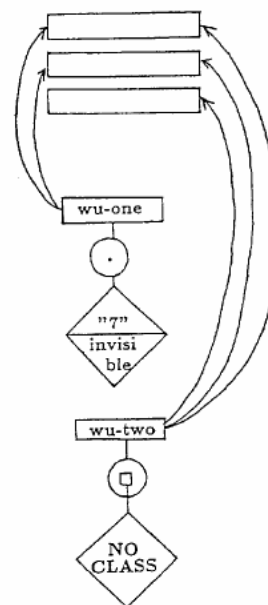


Figure 2.1: Graphical presentation of the two workunits

In a similar way a picture object can be represented as shown in Figure 2.2. The picture object inherited this specific constellation and viewpoint of the axis. It is also indicated above, that the ticmarks may be different along all axis.

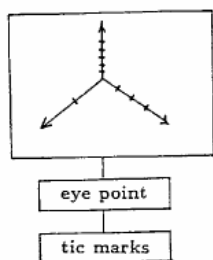


Figure 2.2: Representation of a picture object

The possibility of *activating* objects allows a fast way through command sequences, since as default arguments the possibility of activating objects allows a fast way through command sequences, since as default arguments for object handling always the active object will be assumed. The computation of several smoothing operations of the same (active) workunit does therefore not need the repeated explicit statement of the workunit's name.

Different workunits may be displayed in different picture objects. Figure 2.3 shows a workunit (pointing to the raw data) as a pointcloud together with another workunit showing the smooth regression curve both in one picture object. A density estimate of the marginal density of X is displayed in another picture object (viewport "picture 2") at the upper right corner of the screen.

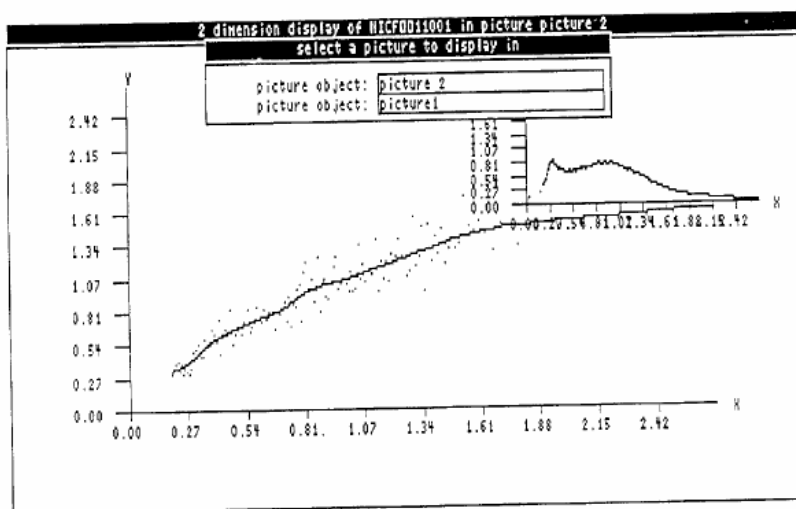


Figure 2.3: Workunits displayed in different picture objects

```

help window level: 4

Trace.hlp)
GENERAL INFORMATION
The ACE algorithm determines the best fitting functions phi[j] in the
following ADDITIVE MODEL

      psi(Y) = sum_{j=1}^p phi[j](X[j]) + error,

Where X[j] denotes the j-th coordinate of the p-dimensional predictor
variable X=(X[1], ..., X[p]).
XploRe expects as input for this manipulation a workunit of the form :

      workunit = (X[1], ..., X[p], Y),

where X and Y denote column vectors. XploRe will create a new workunit
consisting of the fitted functions phi[j], j=1, ..., p and of the fitted
transformation psi.
    
```

Figure 2.4: Example of a help window

Help files can be attached by the system programmer through a stack of "help windows". The designer of the computing environment determines at which analysis stage which "help windows" should appear. The help information is simply obtained by pressing F1. Subsequent pressing of the help key guides through the stack of currently attached help windows. The help windows are in fact internally handled as temporary text objects which are displayed as in Figure 2.4. As more procedures are added to XploRe help-files can be added also. Through a stack mechanism the user can call such help files.

The help windows (and also text objects) can be scrolled backwards and forward by using the PgeDown and PgeUp key. All pulldown menus can be folded and unfolded by successive pressing of the function key F10.

The manipulation of workunits contains currently the following operations:

Regression smoothing

- regressogram
- k-Nearest Neighbour estimation
- super smoothing
- kernel estimation
- weighted averaging using rounded points
- isotonic regression
- running median
- polynomial fitting
- bootstrapping for confidence bounds
- choice of squared error optimal smoothing parameter

Density smoothing

- histogram
- k-Nearest neighbour estimation
- kernel smoothing
- (log)normal fitting
- choice of smoothing parameter

For details on these operations see HÄRDLE (1988).

Additive Models

- Alternating Conditional Expectations (ACE)
BREIMAN and FRIEDMAN (1985)
- Projection Pursuit Regression (PPR)
FRIEDMAN and STUETZLE (1981)
- Recursive Partitioning Regression Trees (RPR)
BREIMAN, FRIEDMAN, OLSHEN, STONE (1984)
- Average Derivative Estimation (ADE)
HÄRDLE and STOKER (1988)

Other manipulations include the possibility to remove missing observations (or ties) or to define new workunits from an existing one according to certain mask attributes.

III. THE INTERACTIVE DISPLAY

Experimental programming techniques rely very much on an interactive display system. Removal, identification and classification of points should be done in an interactive way by just pointing with a cursor to a group of points. This technique is incorporated in XploRe by the *label* and *mask* option of the graphics command menu, see Figure 3.1.

By clicking the "label" field the cursor can be moved to any point on the screen. After pressing ENTER a window pops up that shows the index of the observation (closest in Eukledian distance) together with the coordinate of the workunit. This feature enables the user to see all coordinates of a high dimensional workunit although he might be looking only at one "interesting" point in a two or three dimensional projection. The "mask" field allows the user to interactively define a rectangle of points which he would like to classify into groups 1-9 or invisible. The "unmask" option reverses this action. The *edit* field allows to change the ticmarks and the scaling of the axis and also the display style of the workunit currently shown. The *movoff* is a switch to *movon* which means that all screen information is stored in a movie fashion to disk. By pressing *movie* the saved screens will be shown, this feature allows

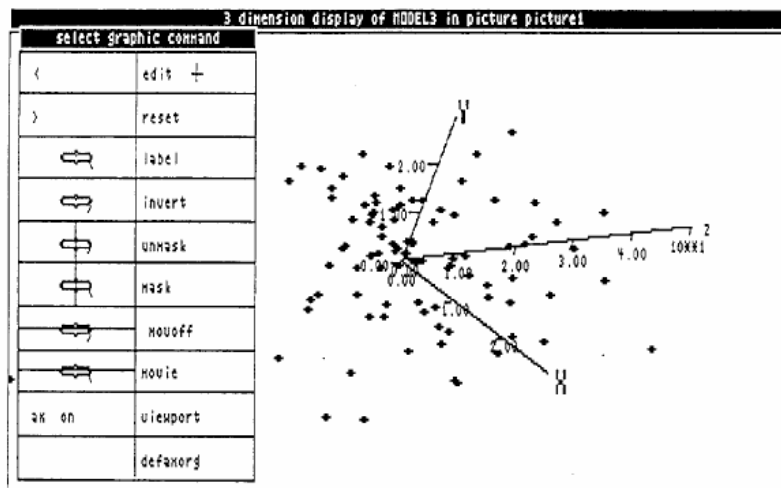


Figure 3.1: Demonstration of label and mask option

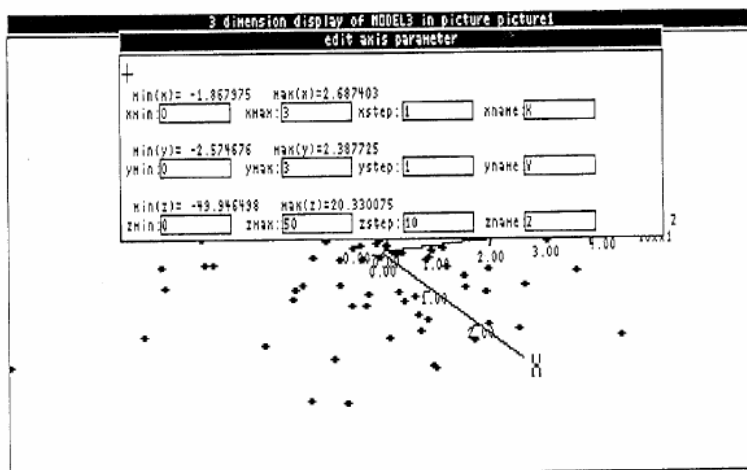


Figure 3.2: Edit command possibilities

tracking of past actions as well as dynamic 3D views of rotating point clouds.

The *viewport* option allows the user to map certain sub-rectangles of the screen to the whole screen. By zooming into a point cloud one may get better understanding of local structures. The *defaxorg* field is for interactive definition of the axis origin. Clicking *ax on* switches to *ax off* which has the effect to display the data without the axis. The six fields above refer to rotations clock- and counterclockwise around each of the three axis in 3D space. The two fields in the upper left corner define the distance of the eyepoint relative to the pointcloud. Clicking successively ">" gives the impression to come closer to the data, whereas "<" makes the distance bigger.

The *edit* field is for locally changing the display style and for inheriting the current picture object ticmarks and axis labelling. Figure 3.2 shows the screen just after clicking "edit" in the situation of Figure 3.1.

The sensitive fields, shown by rectangles, show the current tics. By overwriting in these fields one changes the layout of the axis. The *reset* option gives the standard axis in the cube $[0, \max(x,y,z)]^3$.

IV. INSTALLING NEW PROCEDURES

As an example of how to install own routines I describe how the *running median* primitive was implemented into XploRe. I assume that there is already a procedure *runmed* (y, n, k, s) with input array y , length n , smoothing parameter k and output array s (containing the running median sequence). The user chooses the running median manipulation basically by some mouseclicks and the manipulation refers then to the active workunit object. This workunit has to be sorted by the first column (interpreted as the predictor variable x), then the response variable y has to be stripped off to determine the running median smooth s . It is convenient to build a vector object for this output array s and to create a workunit containing links to the predictor variable x . Inside XploRe these operations would read as follows:

```

procedure dorunmed (wu);
var
  x, y, s: workarray;
  n, k: integer;
  xvec, yvec, svec, newwuobj: objectid;
begin
  quicksort(wu);
  getvector(wu, xvec, x, n, 1);
  getvector(wu, yvec, y, n, 2);
  getparameter(k);
  runmed(y, n, k, s);
  createobj(svec, s, n);
  incvector(svec, s, n);
  createobj(newwu, wuobjpartyp);
  inclink(newwu, xvec, 1);
  inclink(newwu, svec, 2);
end.
    
```

The *getvector* procedure extracts from workunit wu the x and y array. The *createobj* procedure creates an object of the specified type (vectorpartyp, wuobjpartyp). The *incvector* (*inclink*) procedure includes an array (a link) into vector objects (Workunit objects).

V. HIGHER DIMENSIONAL SMOOTHING TECHNIQUES

Nonparametric regression models with more than one predictor variable are handled in XploRe by means of fitting *additive* models. Currently the following models can be fit for a d -dimensional predictor variable (X_1, \dots, X_d)

$$\Psi(Y) = \sum_{j=1}^d \Phi(X_j) + \text{error}$$

and

$$Y = g\left(\sum_{j=1}^d \alpha_j X_j\right) + \text{error}.$$

XploRe uses the ACE-algorithm to find the nonparametric transformations Ψ and $(\Phi_j)_{j=1}^d$, see BREIMAN and FRIED-

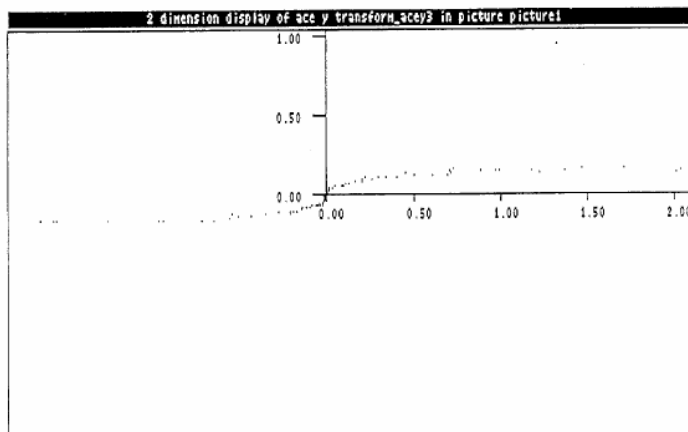


Figure 5.1: Application of the ACE algorithm

MAN (1985). The model exhibiting the "additivity inside", and a nonparametric univariate function g is handled either by Projection Pursuit Regression (PPR), see FRIEDMAN and STUETZLE (1982), or by Average Derivative Estimation (ADE), see HÄRDLE and STOKER (1988). A discrete approximation of the regression curve can be computed using recursive partitioning regression trees (RPR), see BREIMAN et al. (1984). Figure 5.1 shows the transformation $\Psi(y)$ versus y after application of the ACE-algorithm.

The simulated model for this example was

$$Y = (X_1 + X_2)^3 + \text{error}.$$

Clearly the Ψ -transformation recovered the cubic root structure of the data set (as displayed in Figure 3.1). After

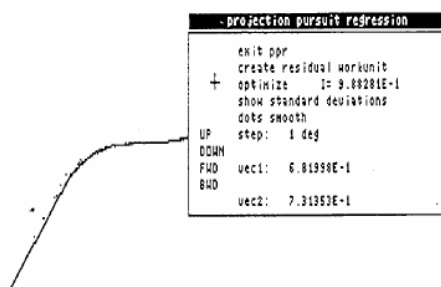


Figure 5.2: Application of the PPR-technique

optimization over projections we find essentially the same structure by the PPR-technique, see Figure 5.2.

A typical output of the RPR-tree algorithm is shown in Figure 5.3. It gives a good graphical expression of the splits (occurring always parallel to some coordinate axis). In a protocol shows XploRe the corresponding mean and the reduction in sample variance.

VI. AVAILABILITY

The program XploRe is available from the author. It fits on a 1.2 MB disk and runs under MS-DOS with almost all video systems (Hercules, CGA, EGA, Olivetti, etc.). The technical report by AERTS and HOLTBERG (1987) describing the systems programmer level of XploRe can be obtained by the author, too.

Acknowledgement

The financial support of the Deutsche Forschungsgemeinschaft and the Koizumi Foundation is gratefully acknowledged. The presentation of the paper improved substantially through discussion with A. Hörmann and R. Shibata.

References

- AERTS, M. and HOLTBERG, A. (1987): Getting Started with XploRe - A Computing Environment for Exploratory Regression and Density Estimation Methods. Technical Report No. A-126, University of Bonn
- BECKER, R.A. and CHAMBERS, J.M. (1984): An Interactive Environment for Data Analysis. Belmont: Wadsworth Press
- BREIMAN, L. and FRIEDMAN, J.H. (1985): Estimating Optimal Transformations for multiple Regression and Correlation (with Discussion). JASA 80, 580-619
- BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R. and STONE, C.J. (1984): Classification and regression trees. Belmont: Wadsworth Press
- FRIEDMAN, J. and STUETZLE, W. (1981): Projection pursuit regression. JASA 76, 817-823

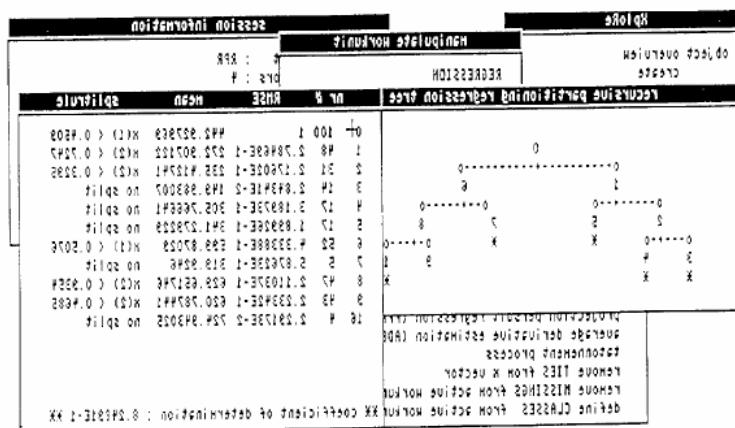


Figure 5.3: Output of the RPP-tree algorithm

HÄRDLE, W. (1988): Applied Nonparametric Regression Book (to appear)

HÄRDLE, W. and STOKER, T. (1988): Investigating multiple regression by the method of averaged derivatives. JASA (to appear)

MCDONALD, J. and PEDERSON, J. (1986): Computing environments for data analysis: part 3: programming environments. Laboratory for Computational Statistics. Stanford University, Technical Report 24

OLDFORD, R.W. and PETERS, S.C. (1985): DINDE: Towards more statistically sophisticated software. Massachusetts Institute of Technology, Technical Report Tr-55

SHIBATA, R. (1981): An optimal selection of regression variables. Biometrika 68, 45-54

SILVERMAN, B.W. (1985): Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). Journal of the Royal Statistical Society (B) 47, 1-45

Resistant Smoothing Using the Fast Fourier Transform

By W. Härdle†

Institut für Wirtschaftstheorie II, West Germany

[Received March 1985. Final revision July 1986]

Keywords: Kernel regression estimation; resistant smoothing; Fast Fourier Transform;**Language**

Fortran 66

Description and Purpose

Suppose $\{(X_j, Y_j)\}_{j=1}^n$ are two-dimensional data points and it is desired to compute the regression curve $r(x) = E(Y|X = x)$ of Y on X . The following curve estimator with kernel function K and bandwidth h

$$r_n^*(x) = \frac{d_n(x)}{f_n(x)} = \frac{n^{-1}h^{-1} \sum_{j=1}^n K(h^{-1}(x - X_j))Y_j}{n^{-1}h^{-1} \sum_{j=1}^n K(h^{-1}(x - X_j))} \quad (1)$$

has been introduced by Nadaraya (1964) and Watson (1964). Brillinger (1977) pointed out the non-resistance to outliers and proposed an M -type smoother. Resistant regression estimates are desirable in data analysis as was pointed out by Velleman and Hoaglin (1981). An application of a resistant smoother to a chemical problem is described in Bussian and Härdle (1984).

In this paper we present an algorithm for the one-step M -type smoother

$$r_n(x) = r_n^*(x) + \frac{n^{-1}h^{-1} \sum_{j=1}^n K(h^{-1}(x - X_j))\psi(\text{res}_j)}{n^{-1}h^{-1} \sum_{j=1}^n K(h^{-1}(x - X_j))\psi'(\text{res}_j)} \quad (2)$$

where $\text{res}_j = Y_j - r_n^*(X_j)$ and $\psi(u) = \max\{-c, \min\{u, c\}\}$, $c > 0$ is Huber's (1981) well known psi function. The boundedness of ψ makes r_n resistant against outliers since large residuals are downweighted. Theoretical aspects of resistant nonparametric regression estimators have been considered by Utreras (1981), Cox (1983) and Härdle (1984) among others. The choice of the bandwidth h is important for the kernel method but in this algorithm the choice of h is left to the user. Bandwidth selection procedures for r_n^* are investigated in Härdle and Marron (1985), Hall (1984) for instance. These procedures require an enormous amount of computation, since the regression estimate has to be calculated for a wide range of bandwidths h . It is therefore desirable to have an efficient numerical method for the computation of formula (2). This is achieved here by the use of the fast Fourier transform. The choice of the cutoff parameter c ,

† Address for correspondence: Institut für Wirtschaftstheorie II, Universität Bonn, Adenauerallee 24-26, D-5300 Bonn, West Germany.

that regulates the amount of downweighting of the residuals, is not part of this algorithm, and has to be supplied by the user. To our knowledge, techniques for adapting the cutoff parameter c are not known.

Numerical Method

The use of the fast Fourier transform in the setting of kernel estimators was suggested by Silverman (1982). Denote the Fourier transformation by F and apply it to d_n , to get

$$F(d_n)(w) = F(K)(hw)D_n(w) \quad (3)$$

where

$$D_n(w) = n^{-1} \sum_{j=1}^n \exp(iwX_j) Y_j.$$

For computational efficiency the Gaussian kernel is implemented here, so

$$F(K)(w) = \exp(-\frac{1}{2}w^2).$$

We use the Fortran subroutine *FASTF* by Monro (1975), Algorithm AS83, for the efficient evaluation of the complex discrete Fourier transform. In a first step $D_n(w)$ is calculated after discretization of X , then formula (3) is computed and d_n is found by inverse transformation. The density estimate f_n is calculated in a similar way. After the first call the transform $D_n(w)$ is retained, so that subsequent calls with different bandwidths can use D_n . Next, the transformed residuals $\psi(\text{res}_j)$, $\psi'(\text{res}_j)$ are found and the nonlinear one-step correction of $r_n(x)$ is computed in the same way as the numerator of r_n^* .

Underflow is avoided by setting $F(d_n)$ equal to zero if the argument is larger than a constant *BIG*. The calculations should be done on an interval that is somewhat larger than the range of X in order to avoid "boundary effects" of the *FFT*. The subroutine *RESSMO* can be called in three modes (*NEWCAL* = 0, 1, 2). In the first mode the transform $D_n(w)$ and the estimated curve $r_n(x)$ are computed, in the second mode the estimate is found with the use of the retained $D_n(w)$, in the third mode it is assumed that the same h as in the previous call is taken but only the cutoff parameter c is changed, so only the one-step correction is to be calculated in this mode.

Structure

*SUBROUTINE RESSMO (X,Y,NIN,CUTOFF,XLO,XHI,BANDW,SMOOTH,NOUT,
NEWCAL,IWK,W,WA,WB,WC,WK,IFault)*

Formal parameters

<i>X</i>	Real array (<i>NIN</i>)	input: contains the x-data
<i>Y</i>	Real array (<i>NIN</i>)	input: contains the y-data
<i>NIN</i>	Integer	input: the sample size
<i>CUTOFF</i>	Real	input: cutoff point of Huber's psi-function, must be greater zero, otherwise <i>IFault</i> = 1
<i>XLO</i>	Real	input: left boundary point of the interval on the estimate is calculated
<i>XHI</i>	Real	input: right boundary point of the interval on which the estimate is calculated
<i>BANDW</i>	Real	input: the bandwidth
<i>SMOOTH</i>	Real array (<i>NOUT</i>)	output: the values of the regression estimate. <i>SMOOTH(I)</i> is an estimate at $XLO + (I-.5)(XHI - XLO)$ <i>NOUT</i>

<i>NOUT</i>	Integer	input: number of points at which the estimate is calculated. <i>NOUT</i> must be less or equal to 1024 (in this version) otherwise <i>IFAULT</i> = 2
<i>NEWCAL</i>	Integer	input: mode parameter 0: compute <i>SMOOTH</i> from <i>X</i> and <i>Y</i> 1: use previously computed Fourier transforms of the <i>x</i> -Data and D_n stored in $WK(3, \cdot) - WK(6, \cdot)$ 2: use previously computed r_n^* stored in $WK(1, \cdot)$
<i>IWK</i>	Integer array (<i>NIN</i>)	input: work area
<i>W</i>	Real array (<i>NIN</i>)	input: work area
<i>WA, WB, WC</i>	Real array (<i>NOUT</i>)	input: work area
<i>WK</i>	Real field (6, <i>NOUT</i>)	input: if <i>NEWCAL</i> = 2, the values of r_n^* at the same grid points as <i>SMOOTH</i>
	$WK(1, \cdot)$	output: the values of r_n^*
	$WK(2, \cdot)$	input: work area
	$WK(3, \cdot)$,	input: if <i>NEWCAL</i> \neq 0, the Fourier transform D_n (real and imaginary part) as previously output.
	$WK(4, \cdot)$	output: the Fourier transform D_n .
	$WK(5, \cdot)$,	input: if <i>NEWCAL</i> \neq 0, the Fourier transform of the <i>x</i> -Data as previously output
	$WK(6, \cdot)$,	output: the Fourier transform of the <i>x</i> -Data
<i>IFAULT</i>	Integer	output: performance indicator 1: <i>CUTOFF</i> is less or equal zero 2: <i>NOUT</i> is not a power of two or greater than 1024 3: <i>XHI-XLO</i> is less or equal zero 4: <i>BANDW</i> is less or equal zero 5: <i>NEWCAL</i> is not 0, 1 or 2 <0: indicates how many times $ res_j \geq c$

Auxiliary routines

Subroutine *RESSMO* calls a subroutine *LINSMO* four times that calculates d_n, f_n as well as the numerator and the denominator of the one-step correction term. Subroutine *FASTF* which performs the forward and reverse discrete Fast Fourier transform is required. This can be the routine of Monro (1975) or equivalent routines.

Restrictions and Remarks

The number *NOUT* must be chosen to be equal to 2^k with k an integer, $3 \leq k \leq 20$ if the routine *FASTF* of Monro (1975) is used. The present version is limited to *NOUT* equal to 2^k , $3 \leq k \leq 10$, since there is not much gain (see **Accuracy**) in using very large values of *NOUT* for moderate sample size. The interval at which the estimate is calculated must not be altered between successive calls with *NEWCAL* > 0. Setting *CUTOFF* = ∞ does not give r_n^* since in this algorithm for structural reasons the residuals res_j are centered at $r_n^*(X_j)$ and not at $r_n^*(x)$. The Nadaraya-Watson estimate r_n^* is provided by the array $WK(1, \cdot)$ anyway. A double precision version can be obtained by using double precision in the declaration statements and by replacing the functions *FLOAT*, *SQRT*, etc. by their double precision counterparts.

Time

After the first call to *RESSMO*, subsequent calls with $NEWCAL > 0$ are performed much faster since $D_n(w)$ need not be computed again. In the even simpler case where with $NEWCAL = 2$ only the one-step correction term is changed by a different *CUTOFF* parameter, r_n^* is left unchanged. If $NEWCAL$ equals zero *FASTF* is called eight times: four times to transform d_n, f_n and the correction term; another four times to calculate the inverse transformation. If $NEWCAL$ equals one *FASTF* is called six times, since $D_n(w)$ and the transformation of the X -data is retained in $WK(3, \cdot) - WK(6, \cdot)$. If $NEWCAL$ equals two, *FASTF* is called four times since only the correction term has to be calculated. Table 1 gives timings for the different calling modes ($NEWCAL = 0, 1, 2$) of *RESSMO*. For comparison, the time consumption using formula (2) directly is also presented there. The computations were done on an IBM 3081D in the university computing centre of Bonn. The timings refer to the following data. NIN data points were generated, the X -data uniformly distributed over $(0, 3)$, the Y -data with density

$$(9/10)\phi(y - r(x)) + (1/90)\phi((y - r(x))/9),$$

$$r(x) = \sin(\pi x),$$

ϕ denoting the standard normal density.

The *CUTOFF* was set to 1.5, $XLO = -1.0$, $XHI = 4.0$ and $BANDW = 0.2$.

Accuracy

The accuracy and smoothness of r_n^* are dependent on the bandwidth h . In Table 2 two measures of accuracy for r_n^* and r_n are shown as functions of h . The averaged square error (*ASE*) and the maximum absolute deviation (*MAD*) were computed at grid points in the

TABLE 1
Timings in econds for calls to *RESSMO* with $NEWCAL = 0, 1, 2$ and for calculations by direct application of the formula (2)

<i>NOUT</i>	<i>NIN</i>	$NEWCAL = 0$	$NEWCAL = 1$	$NEWCAL = 2$	direct
64	100	0.0101	0.0055	0.0042	0.1417
64	200	0.0110	0.0056	0.0044	0.2757
128	100	0.0199	0.0103	0.0084	0.2775
128	200	0.0205	0.0104	0.0089	0.5532
512	100	0.0835	0.0403	0.0375	1.0950
512	200	0.0835	0.0405	0.0382	2.1783

TABLE 2
Maximal absolute deviation (*MAD*) and averaged square error (*ASE*) of the resistant smoother (*RESSMO*) and the Nadaraya-Watson estimator (*NAD WAT*) ($NIN = 200$, $NOUT = 512$, $CUTOFF = 1.5$, $XLO = -1$, $XHI = 4$)

h	<i>MAD</i>		<i>ASE</i>	
	<i>RESSMO</i>	<i>NAD WAT</i>	<i>RESSMO</i>	<i>NAD WAT</i>
0.15	0.88	1.46	0.135	0.394
0.2	0.772	1.13	0.1	0.292
0.25	0.604	1.15	0.087	0.244
0.3	0.62	1.28	0.089	0.219
0.35	0.769	1.34	0.109	0.208
0.4	0.95	1.34	0.131	0.206
0.45	1.07	1.31	0.155	0.212

TABLE 3
 Discretization error of RESSMO
 (NIN = 200, CUTOFF = 1.5, XLO = -1,
 XHI = 4)

h	NOUT = 512	NOUT = 256
.15	1.9×10^{-2}	9.5×10^{-2}
.2	$.87 \times 10^{-2}$	2.9×10^{-2}
.25	$.51 \times 10^{-2}$	1.8×10^{-2}
.3	$.38 \times 10^{-2}$	1.7×10^{-2}
.35	$.35 \times 10^{-2}$	1.4×10^{-2}
.4	$.34 \times 10^{-2}$	1.1×10^{-2}
.45	$.34 \times 10^{-2}$	$.95 \times 10^{-2}$

interval (0, 3.) Table 2 shows the advantage of the one-step smoother r_n over the Nadayara-Watson estimator r_n^* , both the ASE and the MAD are considerably smaller for r_n than for r_n^* . In Table 3 the discretization error of r_n for the same data set that was used above is shown. The errors given are the maximum over the interval (0, 3) of the difference between the estimate and the exact values obtained by direct evaluation of formula (2).

References

- Brillinger, D. R. (1977) In Discussion of "Consistent Nonparametric Regression" by C. J. Stone. *Ann. Statist.*, 5, 622-623.
- Bussian, B. and Härdle, W. (1984) Robust smoothing applied to white noise and single outlier contaminated Raman spectra. *Appl. Spectroscopy*, 38, 309-313.
- Cox, D. D. (1983) Asymptotics for M -type smoothing splines. *Ann. Statist.*, 11, 530-551.
- Härdle, W. (1984) Robust regression function estimation. *J. Mult. Anal.*, 14, 169-180.
- Härdle, W. and Marron, S. (1985) Optimal bandwidth selection in nonparametric kernel regression. *Ann. Statist.*, 13, 1465-1481.
- Hall, P. (1984) Asymptotic properties of integrated square error and cross-validation for kernel estimation of a regression function. *Z. Wahrscheinlichkeitstheorie*, 67, 175-196.
- Huber, P. J. (1981) *Robust Statistics*. New York: Wiley.
- Monro, D. M. (1975) Algorithm AS83. Complex discrete fast Fourier transform. *Appl. Statist.*, 24, 268-272.
- Nadaraya, E. A. (1964) On estimating regression. *Theor. Prob. Appl.*, 9, 141-142.
- Silverman, B. W. (1982) Kernel density estimation using the fast Fourier transform. *Appl. Statist.*, 31, 93-97.
- Utreras, F. (1981) On computing robust splines and applications. *SIAM J. Sci. Stat. Comp.*, 2, 153-163.
- Velleman, P. F. and Hoaglin, D. C. (1981). *Applications, Basics and Computing of Exploratory Data Analysis*. San Francisco: Wadsworth.
- Watson, G. S. (1964) Smooth regression analysis. *Sankhya*, A, 26, 359-372.

```

SUBROUTINE RESSMO(X, Y, NIN, CUTOFF, XLO, XHI, BANDW, SMOOTH,
* NOUT, NEWCAL, INK, W, WA, WB, WC, WK, IFAULT)
C
C   ALGORITHM AS222 APPL. STATIST. (1987) VOL. 36, NO. 1
C
C   INTEGER INK(NIN)
C   REAL X(NIN), Y(NIN), SMOOTH(NOUT), W(NIN), WA(NOUT), WB(NOUT),
* WC(NOUT), WK(6, NOUT)
C
C   DATA ZERO, ONE, TWO, PI, BIG /0.0, 1.0, 2.0, 3.1415, 30.0/
C
C   RANGE = XHI - XLO
C
C   CHECK CUTOFF PARAMETER
C
C   IF (CUTOFF .GT. ZERO) GOTO 1
C   IFAULT = 1
C   RETURN

```

```

C      CHECK IF NOUT IS POWER OF TWO
C
1  NPOW = 8
   M = 2 ** 4
   DO 2 K = 3, 11
   IF (NPOW .EQ. NOUT) GOTO 3
   M = M * 2
2  NPOW = NPOW + NPOW
   IFAULT = 2
   RETURN

C
C      CHECK RANGE
C
3  IF (RANGE .GT. ZERO) GOTO 31
   IFAULT = 3
   RETURN

C
C      CHECK BANDWIDTH
C
31 IF (BANDW .GT. ZERO) GOTO 32
   IFAULT = 4
   RETURN

C
C      CHECK NEWCAL
C
32 IF (NEWCAL .EQ. 0 .OR. NEWCAL .EQ. 1 .OR. NEWCAL .EQ. 2) GOTO 4
   IFAULT = 5
   RETURN

C
4  XSTEP = RANGE / FLOAT(NOUT)
   A = FLOAT(NOUT) / (RANGE * FLOAT(NIN))
   N2 = NOUT / 2
   B = TWO * (PI * BANDW / RANGE) ** 2
   IF (NEWCAL .EQ. 2) GOTO 8

C
   JHI = SQRT(BIG / B)
   JMAX = MINO(N2 - 1, JHI)
   XLO1 = XLO - XSTEP
   IF (NEWCAL .NE. 0) GOTO 51
   DO 5 I = 1, NIN
   IWK(I) = (X(I) - XLO1) / XSTEP
5  W(I) = ONE
51 CONTINUE

C
C      REFRESH FORMER RESULTS
C
   IF (NEWCAL .NE. 1) GOTO 53
   DO 52 J = 1, NOUT
   WA(J) = WK(3, J)
52 WB(J) = WK(4, J)
53 CONTINUE

C
C      FIND NUMERATOR OF NADARAYA-WATSON ESTIMATE
C
   CALL LINSMO(Y, IWK, NIN, SMOOTH, NOUT, NEWCAL, WA, WB, WC, N2,
* JHI, JMAX, A, B, M)

C
C      TRANSFER RESULTS TO WORKAREA
C
   DO 54 J = 1, NOUT
   WK(3, J) = WA(J)
   WK(4, J) = WB(J)
54 WK(1, J) = SMOOTH(J)

C
C      REFRESH FORMER RESULTS
C
   IF (NEWCAL .NE. 1) GOTO 56
   DO 55 J = 1, NOUT
   WA(J) = WK(5, J)
55 WB(J) = WK(6, J)
56 CONTINUE

```

```

C      FIND DENSITY ESTIMATE OF MARGINAL DISTRIBUTION OF X
C
C      CALL LINSMO(W, IWK, NIN, SMOOTH, NOUT, NEWCAL, WA, WB, WC, N2,
*      JHI, JMAX, A, B, M)
C
C      COMPUTE NADARAYA-WATSON ESTIMATE
C
C      DO 7 J = 1, NOUT
C      TEMP = ZERO
C      IF (SMOOTH(J) .GT. ZERO) TEMP = WK(1, J) / SMOOTH(J)
7 WK(1, J) = TEMP
C
C      COMPUTE HUBER'S PSI FROM RESIDUALS
C
C      8 NC = 0
C      DO 11 I = 1, NIN
C      RES = Y(I) - WK(1, IWK(I))
C      IF (RES .GT. (-CUTOFF)) GOTO 9
C      NC = NC + 1
C      W(I) = -CUTOFF
C      GOTO 11
C      9 IF (RES .LT. CUTOFF) GOTO 10
C      NC = NC + 1
C      W(I) = CUTOFF
C      GOTO 11
C      10 W(I) = RES
C      11 CONTINUE
C
C      COMPUTE NONLINEAR CORRECTION
C
C      IFAULT = -NC
C      ICAL = 0
C
C      CALL LINSMO(W, IWK, NIN, SMOOTH, NOUT, ICAL, WA, WB, WC, N2, JHI,
*      JMAX, A, B, M)
C
C      STORE RESULTS
C
C      DO 111 J = 1, NOUT
C      111 WK(2, J) = SMOOTH(J)
C
C      DERIVATIVE OF HUBER'S PSI FROM RESIDUALS
C
C      DO 112 I = 1, NIN
C      TEMP = ZERO
C      IF (W(I) .LT. CUTOFF .AND. W(I) .GT. -CUTOFF) TEMP = ONE
C      112 W(I) = TEMP
C
C      COMPUTE DENOMINATOR OF NONLINEAR CORRECTION
C
C      CALL LINSMO(W, IWK, NIN, SMOOTH, NOUT, ICAL, WA, WB, WC, N2, JHI,
*      JMAX, A, B, M)
C
C      COMPUTE THE FULL ESTIMATOR
C
C      DO 13 J = 1, NOUT
C      TEMP = WK(1, J)
C      IF (SMOOTH(J) .GT. ZERO) TEMP = TEMP + WK(2, J) / SMOOTH(J)
C      13 SMOOTH(J) = TEMP
C
C      RETURN
C      END
C
C      SUBROUTINE LINSMO(Y, IWK, NIN, SMOOTH, NOUT, NEWCAL, WA, WB, WC,
*      N2, JHI, JMAX, A, B, M)
C
C      ALGORITHM AS222 APPL. STATIST. (1987) VOL. 36, NO. 1
C
C      INTEGER IWK(NIN), ITYPE
C      REAL Y(NIN), SMOOTH(NOUT), WA(NOUT), WB(NOUT), WC(NOUT)
C
C      DATA ZERO /0.0/

```

```

      IF (NEWCAL .NE. 0) GOTO 30
C
C      TRANSFORM
C
      DO 10 J = 1, NOUT
      WA(J) = ZERO
10  WB(J) = ZERO
C
      DO 20 I = 1, NIN
      JI = IWK(I)
      IF (JI .LT. 1 .OR. JI .GT. NOUT) GOTO 20
      WA(JI) = WA(JI) + A * Y(I)
C
20  CONTINUE
      M1 = M / 2
      ITYPE = 1
C
      CALL FASTF(WA, WB, M1, ITYPE)
C
C      FILTER FOURIER TRANSFORM
C
30  SMOOTH(1) = WA(1)
      WC(1) = WB(1)
      DO 40 J = 1, JMAX
      C = EXP(-B * FLOAT(J * J))
C
      J1 = J + 1
      J2 = NOUT - J + 1
      SMOOTH(J1) = C * WA(J1)
      WC(J1) = C * WB(J1)
      SMOOTH(J2) = C * WA(J2)
40  WC(J2) = C * WB(J2)
C
C      UNDERFLOW CORRECTION
C
      IF (JHI + 1 - N2) 60, 70, 50
C
50  SMOOTH(N2 + 1) = EXP(-B * FLOAT(N2 * N2)) * WA(N2 + 1)
      WC(N2 + 1) = EXP(-B * FLOAT(N2 * N2)) * WB(N2 + 1)
      GOTO 80
C
60  JHI2 = JHI + 2
      DO 61 J1 = JHI2, N2
      J2 = NOUT - J1 + 2
      SMOOTH(J1) = ZERO
      WC(J1) = ZERO
      SMOOTH(J2) = ZERO
61  WC(J2) = ZERO
C
70  SMOOTH(N2 + 1) = ZERO
      WC(N2 + 1) = ZERO
C
80  CONTINUE
      ITYPE = -1
C
      CALL FASTF(SMOOTH, WC, M1, ITYPE)
C
      RETURN
      END

```

in: Computing Science & Statistics
 Interface '88
 ed. Wefman, Gantz, Härdle

Interactive Smoothing Techniques
 Wolfgang Härdle, Universität Bonn

Abstract

For effective implementation of smoothing techniques a *conditio sine qua non* is an interactive computing environment. We describe some of the logical structures that we find convenient for interactive smoothing. These structures are implemented in XploRe - a computing environment for parameter free regression and density smoothing in high and low dimensions.

0. The Smoothing Analysis Cycle

Smoothing means parameterfree estimation of regression and density curves. If $X \in \mathbb{R}^d, Y \in \mathbb{R}$ denote a pair of random variables, it is the task of regression smoothing to estimate the mean function $m(\cdot) = E(Y|X = \cdot)$ from an independent sample $\{(X_i, Y_i)\}_{i=1}^n$. Density smoothing consists of finding good approximations to the density function $f(\cdot)$ of X from an i.i.d. sample $\{X_i\}_{i=1}^n$. If no parametric restrictions are imposed on these curves the smoothing technique is nonparametric or parameterfree and is typically based on "pooling neighboring information", see Stone (1977).

There exists a wide variety of methods for parameterfree estimation, see e.g. Silverman (1986). These methods have more or less the same asymptotic sharpness but behave quite differently for finite sample size. This is a situation where the computer can be a very good assistant: smoothing means function estimation and therefore different results can only be studied in the form of comparing graphs or tables of values. Another scenario in this setting is to form residuals and to examine them in an iterative way for non-fitted or overfitted structure, see e.g. the backfitting procedure of Hastie and Tibshirani (1987). Here again the computer is a great assistant in trying several alternatives.

Smoothing in dimensions of X bigger than two creates difficulties on the computational and on the statistical side. First of all one cannot study the full fit function without additional "artificial dimensions". Scott (1986) proposes to use time as this dimension and presents changing density contours for dimension $d = 4$. Secondly, in data sets with moderate sample size there is not enough data to perform the "local data pooling" in an effective way. (Theoretically speaking, this means that the rate of convergence of nonparametric smoo-

thers is extremely slow for large dimensions d , see Ibragimov and Hasminski (1982) and Stone (1982).) *Additive models* reduce this dimensionality problem but require quite a bit of machine power e.g. the Projection Pursuit Regression (PPR) algorithm by Friedman and Stuetzle (1981). Interactive control of such an additive model comes into consideration, where one would like to see slightly different projections and corresponding alternative smooth fits in a small neighborhood of some currently favored fit.

Even if a single smoothing method is preferred the choice of smoothing parameter is rather delicate. A wide variety of algorithms yield (asymptotically) "optimal curves" but these can be quite different for finite sample size, see Marron (1986).

Summarizing the above situations we can state that the applied scientist will experiment with different smooth fits and try several alternatives in an iterative way. The typical scenario might be described as follows. The scientist starts with some initial *smooth* curve and then *examines* the graph and perhaps residuals. In a further step he *evaluates* this information perhaps using prior information on forms or structure of the current curve, then he may want to *compare* this current curve with an alternative. This iteration procedure can be called a smoothing analysis cycle as depicted in Figure 0.1.

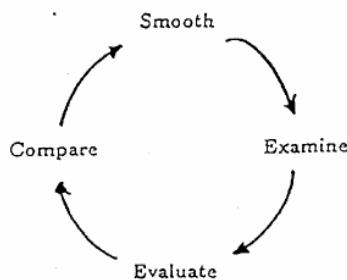


Figure 0.1. The smoothing analysis cycle

This cycle might be performed several times in an improvisational way before one or several satisfactory results are obtained (McDonald and Pederson, 1986). It is obvious that one needs a highly interactive computing environment to go effectively around this cycle.

1. XploRe - an interactive smoothing environment

The computing environment necessary to perform such experimental smoothing falls into three layers (Chambers, 1986):

- α) the individual computer;
- β) the operating system;
- γ) the special logical structures for smoothing.

All three parts interact with each other. Since hardware α) and the system software β) that goes along with it has become affordable even for small institutions the discussion of what to choose for optimization of α) and β) does not seem too relevant to us. In fact we will present the system XploRe as it was developed on a "relatively simple" machine, an IBM AT. The data and program structures γ) for data smoothing and handling seem to be more important to achieve a high degree of interactivity. They should fulfill the following basic requirements.

- (1.1) The interactive system should allow convenient comparison of different fits, preferably in a graphical way.
- (1.2) Certain viewpoints or snapshots (from different "angles") of the data and its smooth should be recordable.
- (1.3) Results, summary statistics or verbalized impressions should be storable on the spot and visible at convenience.
- (1.4) Intermediate stages of a smoothing analysis should be deletable or evocable. Input/Output to or via other layers of the computing environment must be possible.
- (1.5) A dump and a reloading of the current stage of analysis should be possible.

In order to fulfill the above requirements we defined in XploRe the following basic objects:

vector,
workunit,
picture,
text.

Vectors are the simplest objects, they contain an alpha numeric data array of variable length. Workunits are collection of pointers to vectors and may include display and mask attributes. Picture objects are viewports, defining the location and tic marks of the axes in 2D or 3D views. Text objects are sequences of text lines with variable length.

In order to fulfill (1.4) and (1.5) we defined the following basic operations on these objects. Objects can be

created/deleted;
activated/deactivated;
read/written;
manipulated;
displayed.

The concept of the workunit object meets requirements (1.1). In its simplest form a workunit object can be thought of as a data matrix, but the actual realization as a record of pointers to existing vector objects makes it storage space economic. The additional feature of this object to include mask and display information makes exploratory techniques like brushing (Becker and Cleveland, 1986) easy to program. The display information as part of a workunit object makes it convenient to distinguish different functions: Whenever the workunit object is displayed (in a picture object) the corresponding display style information (part of this workunit) is used. This makes it easy to remember different curves. The mask part of this data object can be inherited to children objects (e.g. smooths) of a workunit and makes thus tracing of interesting points through several steps of an analysis possible, see Oldford and Peters (1986) for more information on this inheritance principle and this object oriented approach. A graphical description of workunits is depicted below in Figure 1.1.

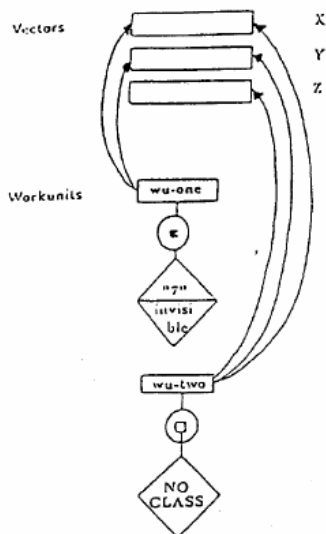


Figure 1.1. Two workunits with mask and display information

Figure 1.1 shows the situation where one wants to analyse a three dimensional data set consisting of vectors X, Y, Z . Workunit *wu-one* consists of the vectors X, Y , another *wu-two* points to all three vectors. When displaying *wu-one* one could have detected some interesting points, which one interactively has marked with the mask "7". Other observations might have been given the mask "invisible". Earlier one might have decided to see the remaining points as stars "*" (except those that have mask "7"). *Wu-two* is shown with square "□" and needles "|" pointing into the (X, Z) plane with no additional mask options.

Picture objects are designed to meet requirement (1.2) and certain information about the location of the 2D or 3D viewpart on the screen, the scaling of all the axes and the location of the axes on the physical screen. This object type is resident until its parts are changed. If one displays a workunit object and has found a reasonable scaling, this current picture object is evokable at later stages. A picture object can be graphically represented as in Figure 1.2.

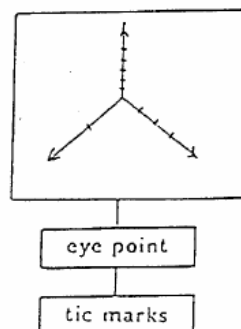


Figure 1.2. A picture object

Different workunits may be displayed in different picture objects. Figure 1.3 below shows a workunit (pointing to the raw data) as a pointcloud together with another workunit showing the smooth regression curve both in one picture object. A density estimate of the marginal density of X is displayed in another picture object (viewport "picture 2") at the upper right corner of the screen.

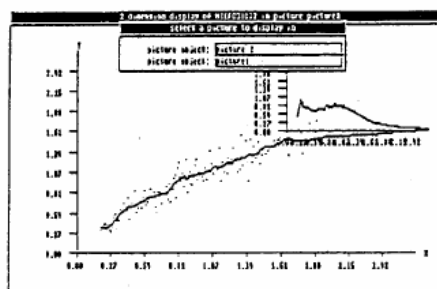


Figure 1.3. Two different picture objects

Text objects are defined according to (1.3). They contain ASCII text lines of variable column length. If such an object is displayed scrolling forward and backward in the actual text are possible. If a text object contains columns of data vectors (as ASCII information) it can be converted into a workunit object (with standard display and mask part) and vice versa.

2. Smoothing Techniques

The basic operations on the four objects have been defined above. All these operations are more or less self-explaining so that we concentrate in this section on the manipulation of workunit and picture objects. The different smoothing techniques entered via this manipulation of an active workunit are described below. The following lists are by no means exhaustive. XploRe (1987) is an open system, more soft work can be included, see section 3.

2.1 Regression Smoothing

- Regressogram (Tukey, 1961).
- k -nearest neighbor estimation (Mack, 1981).
- Supersmoothing (Friedman, 1984).
- Kernel smoothing (Nadaraya, 1964; Watson, 1964).
- WARPing (Härdle and Scott, 1988).
- Isotonic Regression (Barlow et al., 1972).
- Running Median (Tukey, 1977).
- Polynomial Regression (Shibata, 1981).
- Cross-validation (Clark, 1980).

2.2 Density Smoothing

- Histogram.
- k -nearest neighbor estimation (Cover and Hart, 1967).
- Kernel smoothing (Rosenblatt, 1956).
- (Log)Normal fitting.
- L_2 and Kullback Leibler crossvalidation (Marron, 1987).

2.3 Additive Model

- Alternating Conditional Expectations (ACE) (Breiman and Friedman, 1985).
- Projection Pursuit Regression (PPR) (Friedman and Stuetzle, 1981).
- Recursive Partitioning Regression Trees (RPR) (Breiman, Friedman, Olshen and Stone, 1984).
- Average Derivative Estimation (ADE) (Härdle and Stoker, 1988).

2.4 The interactive display

The interactive display features of XploRe allow manipulation of both workunit and picture objects. Removal, identification and classification of points is performed by pointing with a cursor to a group of points. This technique is incorporated in XploRe by the *label* and *mask* option of the graphics command menu, see

Figure 2.1. The mask information will be inherited by the currently displayed workunit object. By clicking the "label" field the cursor can be moved to any point on the screen. After pressing ENTER a window pops up that shows the index of the observation (closest in Eukclidean distance) together with the coordinate of the workunit. This feature enables the user to see all coordinates of a high dimensional workunit although he might be looking only at one "interesting" point in a two or three dimensional projection. The "mask" field allows the user to interactively define a rectangle of points which he would like to classify into groups 1-9 or invisible. The "un-mask" option reverses this action. the *edit* field allows to change the ticmarks and the scaling of the axis and also the display style of the workunit currently shown. The *movoff* is a switch to *movon* which means that all screen information is stored in a movie fashion to disk. By pressing *movie* the saved screens will be shown, this feature allows tracking of past actions.

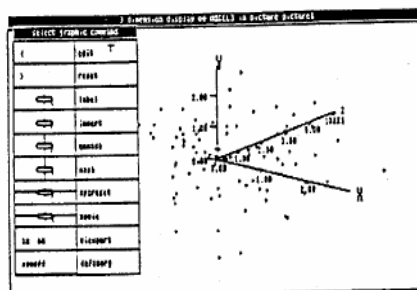


Figure 2.1. The interactive display

The *viewport* option allows the user to map certain sub-rectangles of the screen to the whole screen. The *defazory* field is for interactive definition of the axis origin. Clicking *az on* switches to *az off* which has the effect to display the data without the axis. The six fields above the axis control refer to rotations clock- and counter-clockwise around each of the three axis in 3D space. The two fields in the upper left corner define the distance of the eyepoint relative to the pointcloud. Clicking successively ">" gives the impression to come closer to the data, whereas "<" makes the distance bigger. The 3D graphics have been programmed according to Newman and Sproull (1981).

The *edit* field is for locally changing the display style and for inheriting the current picture object ticmarks and axis labelling. Figure 2.2 shows the screen just after clicking "edit" in the situation of Figure 2.1.

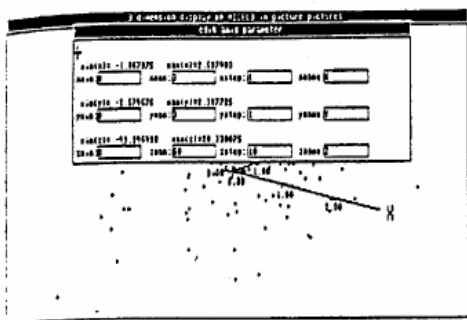


Figure 2.2. Editing the picture object.

The sensitive fields, shown by rectangles, show the current tics. By overwriting in these fields one changes the layout of the axis. The *reset* option gives a standard view in the cube $[0, \max(x, y, z)]^3$.

2.5 Help information

Help files can be attached by the system programmer through a stack of "help windows". The designer of the computing environment determines at which analysis stage which "help windows" should appear. The help information is obtained by pressing F1. Subsequent pressing of the help key guides through the stack of currently attached help windows. The help windows are in fact internally handled as temporary text objects which are displayed as in Figure 2.3.

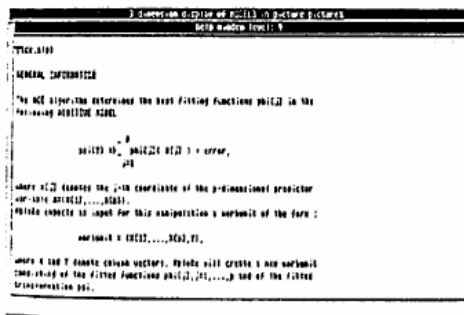


Figure 2.3. A help window

The help windows (and also text objects) can be scrolled backwards and forward by using the PgeDown and PgeUp key. All pulldown menus can be folded and unfolded by successive pressing of the F10 key.

3. Installing own procedures

The system XploRe can be enhanced by installing user written procedures. As an example of how to install own routines we describe how the *running median* primitive was implemented into XploRe. Assume that there is already a procedure *runmed* (y, n, k, s) with input array y , length n , smoothing parameter k and output array s (containing the running median sequence). An optimal algorithm has been given by Härdle, Reinholz and Steiger (1988). The user chooses this manipulation by mouseclicks and by definition the manipulation refers to the active workunit object. This workunit will then be temporarily sorted by the first column (interpreted as the predictor variable x), then the response variable y has to be stripped off to determine the running median smooth s . It is convenient to build a vector object for this output array s and to create a workunit containing links (pointers) to the vector object containing the predictor variable x . In XploRe (respectively TURBO PASCAL) these operations would read as follows.

```

procedure dorunmed (wu);
var
  x,y,s: workarray;
  n,k: integer;
  xvec, yvec, svec, newwuobj: objectid;
begin
  quicksort(wu);
  getvector(wu, xvec, x, n, 1);
  getvector(wu, yvec, y, n, 2);
  getparameter(k); { reads the window size k
                    from the keyboard }
  runmed(y, n, k, s);
  createobj(svec, vectorpartyp, "smooth");
  updatevector (svec, s, n);
  createobj(newwu, wupartyp, "runmed");
  inclink(newwu, xvec);
  inclink(newwu, svec);
end;
    
```

The *getvector* procedure extracts from workunit wu the x and y array. The *createobj* procedure creates an object of the specified type (vectorpartyp, wupartyp). The *updatevector* (*inclink*) procedure includes an array (a link) into vector objects (workunit objects).

Acknowledgement

I would like to thank Wolfgang Rossner who helped in the programming and design of XploRe. The system improved a lot through discussions with David Scott, Anders Holtsberg and Mark Aerts.

References

- Barlow, R.E.; Bartholomew, D.J.; Bremner, J.M. and Brunk, H.D. (1972) *Statistical Inference under Order Restrictions*. Wiley, London.
- Becker, R.A. and Cleveland, W.S. (1986) *Brushing a Scatterplot Matrix: High-Interaction Graphical Methods for Analyzing Multidimensional Data*. Manuscript.
- Breiman, L.; Friedman, J.; Olshen, R. and Stone, C.J. (1984) *Classification and regression trees*. Wadsworth, Belmont.
- Breiman, L. and Friedman, J. (1985) *Estimating Optimal Transformations for Multiple Regression and Correlation*. *J.Amer.Statist. Assoc.*, 80, 580-619.
- Chambers, J.M. (1986) *Computing Environments for Quantitative Applications*. AT & T Bell Labs Stat. Research Reports No. 17.
- Clark, R. M. (1980) *Calibration, Cross-validation and Carbon-14*. II. *J.R.Statist. Soc. A* 143, 177-194.
- Cover, T.M. and Hart, P.E. (1967) *Nearest Neighbor Pattern Classification*. *IEEE Trans. Inf. Theory*, 13, 21-27.
- Friedman, J. and Stuetzle, W. (1981) *Projection Pursuit Regression*. *J.Amer.Statist. Assoc.*, 76, 817-823.
- Friedman, J. and Tibshirani, R. (1984) *The Monotone Smoothing of Scatterplots*. *Technometrics*, 26, 243-250.
- Härdle, W., Reinholz, A. and Steiger, W. (1988) *Optimal Median Smoothing*. Manuscript.
- Härdle, W. and Stoker, T. (1988) *Investigating smooth multiple regression by the method of average derivatives*. *J.Amer.Stat.Assoc.*, submitted.
- Härdle, W. (1988) *Applied Nonparametric Regression*. Book to appear.
- Härdle, W. and Scott, D.W. (1988) *Weighted Averaging using Rounded Points*. Manuscript.
- Hastie, T. and Tibshirani, R. (1987) *Generalized Additive Models: Some Applications*. *J.Amer.Stat.Assoc.*, 82, 371-386.
- Ibragimov, I.A. and Hasminski, R.Z. (1982) *Bounds for the Risk of Nonparametric Regression Estimates*. *Theor. Prob.Appl.*, 27, 84-99.
- Mack, Y.P. (1981) *Local Properties of k-NN Regression Estimates*. *Siam J. Alg. Disc. Meth.*, 2, 311-323.
- Marron, J.S. (1986) *Will the Art of Smoothing ever become a Science?*. in: *Function estimates*, (Marron, ed.) *AMS Contemporary Mathematics* 59.
- Marron, J.S. (1987) *A Comparison of Cross-validation Techniques in Density Estimation*. *Ann.Statist.*, 15, 152-162.
- McDonald, J. and Pederson, J. (1986) *Computing Environments for Data Analysis: Part 3: Programming Environments*. *Laboratory for Computational Statistics, Stanford Technical Report*, 24.
- Newman, W.M. and Sproull, R.F. (1981) *Principles of Interactive Computer Graphics*. Mc Graw-Hill.
- Oldford, R.W. and Peters, S.C. (1985) *DINDE: Towards more Statistically Sophisticated Software*. MIT, Technical Report Tr-55.
- Rosenblatt, M. (1956) *Remarks on some non-parametric estimates of a density function*. *Ann. Math. Statist.* 27, 642-669.
- Scott, D.W. (1986) *Data Analysis in Three and Four Dimensions with Nonparametric Density Estimation*. in "Statistical Image Processing and Graphics" ed. E. Wegman, D. Priest, Marcel Dekker.
- Shibata, R. (1981) *An Optimal Selection of Regression Variables*. *Biometrika* 68, 45-54.
- Silvermann, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.
- Stone, C.J. (1977) *Consistent Non-parametric Regression (with Discussion)*. *Ann. Statist.* 5, 595-645.
- Stone, C.J. (1982) *Optimal Global Rates of Convergence for Nonparametric Regression*. *Ann. Statist.*, 10, 1040-1053.
- Tukey, J.W. (1961) *Curves as Parameters and Touch Estimation*. *Proc 4th Berkeley Symposium*, 681-694.
- Tukey, J.W. (1977) *Exploratory Data Analysis*. Addison, Reading, Massachusetts.
- XploRe (1987) *XploRe - a computing environment for exploratory Regression and density smoothing*. *Wirtschaftstheorie II, Universität D-5300 Bonn*.

Efficient Nonparametric Smoothing in High Dimensions Using Interactive Graphical Techniques

W. Härdle¹, Bonn

Abstract

Smoothing techniques are used to reduce the variability of point clouds. There is great interest not only among applied statisticians but also among applied workers in biostatistics, economics and engineering to model the data in a nonparametric fashion. The benefits of this more flexible modeling come at the cost of greater computation, especially in high dimensions. In this paper several possibilities of smoothing in high dimensions are described using additive models. The algorithms for solving the nonparametric smoothing problems are based on WARPing, i.e. Weighted Averaging using Rounded Points. Interactive graphical techniques are a *conditio sine qua non* for tuning and checking the structure of lower dimensional projections of the data and of smooths produced by the algorithms. Applications of the WARPing technique to a side impact study are shown by smoothing in Projection-Pursuit-type models using Average Derivative Estimation.

¹ This research was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 303.

1. Interactive Smoothing

Smoothing means parameter free re-expression of data points in a form that is easier to understand than the raw point cloud itself. In a *regression smoothing* problem a $(d + 1)$ -dimensional point cloud is observed consisting of response variables $\{Y_i\}_{i=1}^n \in \mathbb{R}$ at predictor variables $\{X_i\}_{i=1}^n \in \mathbb{R}^d$. One is then interested in smoothing the data in order to recover the mean function $m(x) : \mathbb{R}^d \rightarrow \mathbb{R}$. In a *density smoothing* problem a d -dimensional point cloud is observed and data smoothing yields an estimate of the density of the joint density $f(x)$.

There exists a wide variety of smoothing methods, see e.g. Silverman (1986), Härdle (1988b). Correctly compared and tuned these different methods have very similar asymptotic properties, but may exhibit quite different small sample behavior, see Müller (1987), Silverman (1984) and Härdle, Hall and Marron (1988). This is a perfect context in which the interactive graphics tools of an intelligent computing environment are very helpful. From a mathematical viewpoint smoothing involves function and functional estimation and therefore the various results can only be studied in the form of comparing graphs or tables of values. Another scenario where the power of an interactive computing environment is useful is the analysis of residuals : Several algorithms for finding *additive models* are based on iterative schemes based on "intermediate residuals". The *back-fitting algorithm* for Generalized Additive Models described in Hastie and Tibshirani (1986) uses the residuals iteratively.

Smoothing in dimensions bigger than 4 creates problems on the computational and the statistical side. Since parameter free smoothing techniques are based on the idea of *local data pooling* one is usually faced with the problem of sparseness of point clouds. There is a nice illustrating example in the introduction of Friedman and Stuetzle (1981). The sparseness of observations in higher dimensions is sometimes called *curse of dimensionality*, see Huber (1985). A consequence of this curse of dimensionality is the lower rate of convergence of nonparametric smoothers, see Ibragimov and Khas'minskii (1981) and Stone (1982). On the presentation side the problem is that one cannot see the full fit for dimensions greater than three. In four dimensions one can introduce an artificial time dimension (Scott, 1986) but for higher dimensions one has to rely on studying several interesting projections. A complimentary approach is to

perform a Principal Components Analysis (PCA) for dimensionality reduction, see Caussinus(1986). *Additive models* reduce the dimensionality problem on the statistical side by imposing more structure on the functions to be estimated. Despite this the computational burden is still present if not increased since several iterative steps may have to be computed. The necessity for interactive control of smoothing algorithms becomes evident if one summarizes the typical operations. These operations might be performed several times in some sort of "smoothing analysis cycle" until one or several satisfactory results are obtained, see McDonald and Pederson (1986), Härdle (1988a).

SCATTERPLOT SMOOTHING. An elementary building block for higher dimensional model fitting is an efficient scatterplot smoother. Breiman and Friedman (1985) in their ACE-algorithm use a symmetrized k -nearest neighbor algorithm to iteratively compute the "transformations" for Alternating Conditional Expectations.

OPTIMIZATION. Often a measure of "interestingness" is computed and optimized by choice of a tuning parameter. Jones and Sibson (1986) study several indices of "interestingness" for projection pursuit and state that it may be sometimes necessary to run a preliminary PCA in order to reduce computational efforts, see also Caussinus (1987). In the projection pursuit regression setting the optimization is performed over linear combinations of the predictor variables together with one dimensional scatterplot smoothers.

ITERATION. The process of extracting features from residuals might involve changing the "interestingness" functional or a transformation in order to find good approximations to additive models. The CART algorithm for instance iterates recursively to find a Classification And/or Regression Tree of a prescribed complexity, see Breiman, Friedman, Olshen and Stone (1984).

CALIBRATION. A desirable operation in density or regression smoothing is to calibrate the smoothing parameter or to construct confidence bands. Cross validation and related methods (Scott and Terrell, 1987, Härdle, Hall and Marron, 1988) have been used to find good smoothing parameters. Confidence bands have been constructed using the bootstrap in order to calibrate the variability bands, see Härdle and Bowman (1988).

The above elementary operations involve computations that are typically linear in the squared sample size if one uses straightforward implementations.

Improved computation may be obtained by either reducing the number of arithmetic calculations or using techniques that eliminate one or more of the above four steps. In this paper it is demonstrated how WARPing achieves this goal of improved computational efficiency. Furthermore some requirements of a computing environment for performing efficient smoothing in high dimensions are given. It is seen that *XploRe* - a computing environment for *eXploratory Regression and density smoothing* - meets some of these requirements even on a "small scale" machine like an IBM AT.

2. Additive Models and WARPing

Additive models have been studied in order to reduce the dimensionality problem. It is probably easiest to explain the development of this model class starting from binary regression, i.e. the Y -observations are in $\{0,1\}$. A quite common approach to model the conditional probability $m(x) = P(Y = 1|X = x)$ is to assume that there is a known *link function* $G : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$(2.1) \quad m(x) = G(x^T \beta)$$

for some parameter vector β . A well known example is the logistic regression model where G is the distribution function of the logistic distribution. This model is additive but may be unsatisfactory in several ways. First it is rather narrow in the sense that the link function is assumed to be known. Secondly one might argue that it is too restrictive to assume that the regression function depends on a one dimensional link function of certain projections of the predictor variables. The first criticism can be overcome by so called projection pursuit models which have the form

$$(2.2) \quad m(x) = \sum_{l=1}^D g_l(x^T \beta_l),$$

where $g_l : \mathbb{R} \rightarrow \mathbb{R}$ are nonparametric functions operating on projections $x^T \beta_l$. An algorithm to fit such models is described in Friedman and Stuetzle (1981). The second criticism can be overcome by leaving the link function in a suitable

known form but extend the linear projection $x^T \beta_l$. This leads to Generalized Additive Models

$$(2.3) \quad m(x) = G\left(\sum_{l=1}^d g_l(x_l)\right),$$

where as above the $g_l : \mathbb{R} \rightarrow \mathbb{R}$ are nonparametric functions but now operating on the l -th coordinate $x_l = (0, \dots, 1, \dots, 0)^T x$ of x . Hastie and Tibshirani (1987) discuss this model and show how the functions g_l can be estimated by the *local scoring* algorithm. The ACE-model by Breiman and Friedman (1985) is yet another extension where also the function G is nonparametrically estimated.

The ACE-algorithm attempts to find nonparametric functions $G^*, g_l^*, l = 1, \dots, d$ such that

$$(2.4) \quad e^2(G, g_1, \dots, g_d) = \frac{E\{(G(Y) - \sum_{j=1}^d g_j(X_j))^2\}}{EG^2(Y)}$$

is minimized. The ACE-algorithm is highly iterative and makes extensive use of an elementary scatterplot smoother. The computational burden is even bigger if a "supersmoother" - a symmetrized k -nearest neighbor smoother with locally optimized bandwidth choice - is implemented, see Friedman (1986). Similar computational costs occur for the other additive models. In order to make additive model fits accessible for the average user one should start optimizing the scatterplot smoothing routines. The WARPing approach is a promising technique to achieve computational efficiency.

The WARPing technique is based on *Weighted Averaging using Rounded Points*. It is easiest to demonstrate this approach in the univariate density smoothing context. It is well known that the histogram may show quite different shapes if the origin of the histogram defining classes is changed, see Scott (1986). A natural way of eliminating this effect is to construct an average of histograms over a collection of origin choices. In particular, if there are q origins $\{x_{0,k} = kh/q, k = 0, \dots, m-1\}$ for a histogram $HG_h(x, x_{0,k})$ with bin width h , then the WARPed histogram is simply

$$(2.5) \quad \hat{f}_{h,q} = q^{-1} \sum_{k=0}^{q-1} HG_h(x, x_{0,k}).$$

It is straightforward to see by partial summation that this density estimate is simply a weighted average of points aggregated into smaller bins of width $\delta = h/q$. If the parameter q tends to infinity the density estimate (2.5) approaches the kernel estimate with triangular kernel. This motivates a generalization of this technique to other weighting schemes. More specifically one constructs a weighting sequence $w_q(k)$ that sum to one and estimates the density by

$$(WARP) \quad \hat{f}(x) = q^{-1} \sum_{k=1-q}^{q-1} RP_{i(x)+k},$$

where $i(x)$ is the bin in which x falls and RP_i is the frequency of *Rounded Points* in the i -th bin. This technique is also applicable to scatterplot smoothing and is described in detail in Härdle and Scott (1988).

Consider the ACE algorithm again. In this context the WARPed kernel smoother has to perform $2\delta^{-1}q + n$ operations for each of the elementary scatterplot smoothing fits. The number n comes from discretizing the data into rounded points. The effective cost on the rounded points is equal to $2\delta^{-1}q$. Depending on δ and q it can be made highly efficient compared to ordinary kernel smoothing or symmetrized k -nearest neighbor smoothing. The latter will take $n \log(n) + 2n$ operations, the $n \log(n)$ cost coming from sorting the one dimensional X variables for recursive updating of local linear fits, see Friedman and Stuetzle (1982). Also the Projection Pursuit Regression (PPR) algorithm requires an efficient scatterplot smoother as a basic tool. The PPR algorithm searches first for the best pair (β, g) such that with β suitably normalized the residual sum of squares $\sum_{i=1}^n (Y_i - g(X_i^T \beta))^2$ is minimized. Again this task is done iteratively. In a further step residuals are fitted and the same procedure is applied to the set of estimated residuals. Note that if β is standardized such that $E_X(dg/d(x^T \beta)) = 1$ then the terms in the projection pursuit model (2.2) can be estimated by the average derivate

$$(2.6) \quad \delta = E_X(m'(X)),$$

where m' denotes the vector of partial derivatives. *Average Derivative Estimation (ADE)* is a technique for estimating δ . An estimator for δ is given by

$$(2.7) \quad \hat{\delta} = n^{-1} \sum_{i=1}^n Y_i (-\hat{f}'_h(X_i) / \hat{f}_h(X_i)).$$

Again the density estimates in this formula can be estimated using WARPing. The estimator $\hat{\delta}$ is then used to construct a two dimensional scatterplot of $\{(\hat{\delta}^T X_i, Y_i)\}$. A smoother applied to this scatterplot yields an estimate of the nonparametric function g in (2.2). This technique also applies to partial linear models, see Spiegelman (1976), Rice (1986), Heckman (1986). Theoretical properties of the ADE technique are given in Härdle and Stoker (1988).

3. The Desirable Computing Environment

The computing environment necessary to perform smoothing techniques falls into three layers (Chambers, 1986) : The individual computer, the operating system and the special logical structures designed for effective smoothing. It is clear that all three parts interact with each other and that the degree of interaction varies from machine to machine. There is no useful purpose here to discuss the interaction between hardware and operating system here since sophisticated machines become affordable nowadays even on the desktop level. However, the minimal requirement on the hardware side for good performance of smoothing techniques seems to be a Personal Computer with color display and at least 200*640 pixels. Most important is the user interface which should allow various graphical controls desirably mouse oriented. The logical structures of such an intelligent computing environment should meet the following basic requirements in order to allow a high degree of interactivensness with the data or intermediate stages of an analysis.

- (3.1) The interactive system must enable fast and informative comparison of different fits, preferably in a graphical way (colors, linestyle,...).
- (3.2) Since a data set might show an interesting structure only from a certain angle, certain viewpoints or snapshots of the data and its current fits should be recordable.
- (3.3) Intermediate stages of the analysis should be storable and evocable in a further analysis step.
- (3.4) The system should be open to new programs and macros. The macro language or the new programs should have well defined system interfaces.

Oldford and Peters (1986) propose the object oriented approach in order to meet such requirements. In the object oriented approach data is not only seen as a data matrix but as a more complex object that not only carries the pure numbers but also display information, linestyle, masks and other data analytic features. The system S by Becker and Chambers (1984) is a good example for an open analysis system. Also in XploRe (1987) the object oriented approach is taken. Moreover the *inheritance principle* is employed in order to facilitate graphical interpretation of the results. This principle allows the user of the system to inherit certain attributes from existing objects (data) to offspring objects (data).

Suppose for instance one has decided to mask part of a certain data set with a marker "1". (This can be done interactively through mouse control defining a rectangle.) This mask information can be inherited to any smoother applied to that data. The smooth fit as an offspring object can thus show the marker "1" at the same data points where it was in the raw data. This technique makes it simple to trace influences of points or to understand certain local structures of a nonparametric fit.

In XploRe there exist four different object types: vector-, workunit-, text- and picture-objects. Workunits are collections of vector objects and represent point clouds. Workunits are established by pointers to existing vector objects and include various data analytic attributes, for details see Härdle (1987). Text objects are containing ASCII information and have been included for formatted input/output and for display of intermediate numerical results and summary statistics. Picture objects have been defined to allow handling of different projections or aspects of data in different viewports or sections of the screen. Figure 1 shows two different picture objects (viewports). Three workunit objects are shown in two picture objects. The big plot (picture object *picture 1*) shows two vector objects of a ten dimensional workunit as dots. The smooth curve is another workunit created by a k -nearest neighbor. The second picture object in the right upper corner shows a workunit object containing the marginal density of the X variable of first workunit.

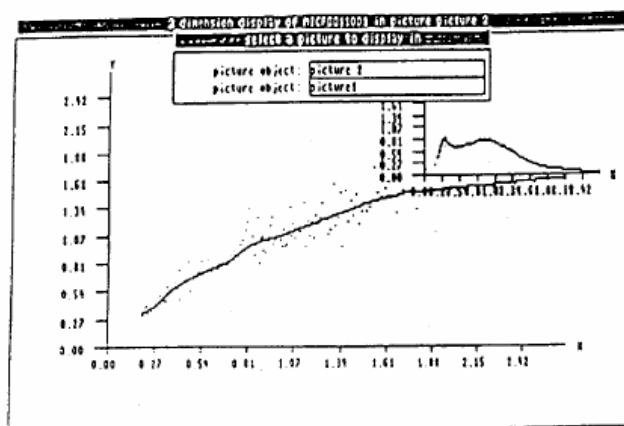


Figure 1. Three workunit objects in two picture objects.

The smoothing parameters for these plots have been selected by cross-validation.

4. An Application to Side Impact Data

The above described techniques have been applied to some side impact data. The object of the study was to compare the behavior of dummies and Post-Mortem-Human-Subjects (PMHS), see Kallieris, Mattern and Härdle (1986). The WARP technique was used to compute a three dimensional density estimate of three biomechanical variables. The object was to see the distribution of these variables for dummies and for PMHS. There were 31 dummy variables and 58 PMHS variables. Figure 2 shows the three dimensional density of these variables for dummies with a contour at level 5 percent of the modal level.

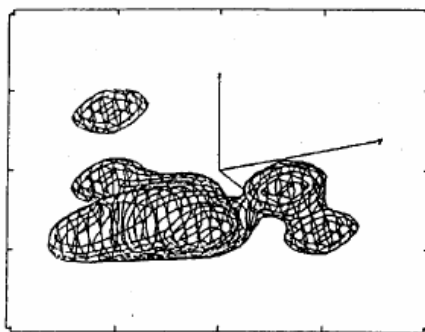


Figure 2. WARPed density smoother for dummies

Figure 3 shows the density of the same variables for PMHS at the same contour level.

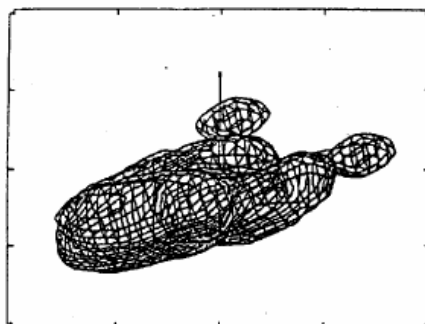


Figure 3. WARPed density smoother for PMHS

Interestingly there are two clusters at the "drumsticks" of the "frozen duck" shape in Figure 3. Both density contours are plotted on the same scale and show a very different shape due to the fact that the frozen duck has higher z -values and lower x -values than the "flying duck" (Figure 2).

The average derivative technique has been applied to this data set as well. Let Y be the random variable with $Y = 1$ indicating fatal injury and $Y = 0$ indicating survival. Figure 4 shows the density contours for the variables ($AGE, VEL, T12RM$) for the two groups with $Y = 0$ or 1 respectively. The contour in the foreground corresponds to $Y = 1$.

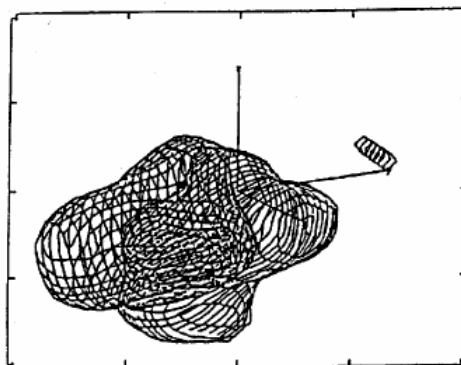


Figure 4. WARP density contours for PMHS

Setting up the above PPR model with unknown link function g we have applied the ADE technique to find projections $\hat{\delta}$. Figure 5 displays the projected predictor variables $\hat{\delta}^T X$ versus the response Y . The solid line indicates the WARP kernel smoother with biweight kernel.

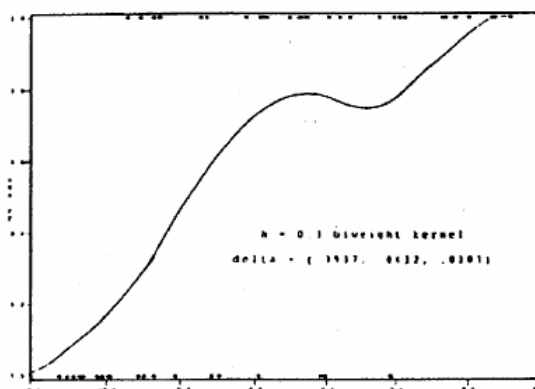


Figure 5. WARPed ADE technique and WARPed kernel smoother.

The group of outlying observations could be identified as observations with low AGE variable.

References

- Becker, R.A. and Chambers, J.M. (1984). *S: An Interactive Environment for Data Analysis and Graphics*. Wadsworth, Belmont.

- Breiman, L. ; Friedman, J. ; Olshen, R. and Stone, C.J. (1984). Classification and Regression Trees. *Wadsworth, Belmont*.
- Breiman, L. and Friedman, J. (1985). Estimating Optimal Transformations for Multiple Regression and Correlation (with discussion). *J.Amer.Stat.Assoc.*, *80*, 580-619.
- Caussinus, H. (1986). Models and Uses of Principal Component Analysis. *Proc. Multidimensional Data Analysis, Cambridge, DSWO Press*.
- Caussinus, H. (1987). Discussion of "What is Projection Pursuit?" by Jones, M.C. and Sibson, R. *J.Royal Statist. Soc.(A)*, *150*, 26.
- Chambers, J.M. (1986). Computing Environments for Quantitative Applications. *ATT Bell Labs Stat. Research Reports 17*.
- Friedman, J. and Stuetzle, W. (1981). Projection Pursuit Regression. *J.Amer.Stat.Assoc.*, *76*, 817-823.
- Friedman, J. and Stuetzle, W. (1982). Smoothing of Scatterplots. *Tech. Report Orion 3. Dept. Statistics, Stanford University*.
- Härdle, W. (1988a). Interactive Smoothing Techniques. in: *Proceedings of the Interface Conference, 1988, Reston Virginia, E. Wegman, Ed*.
- Härdle, W. (1988b). Applied Nonparametric Regression. *Cambridge University Press, to appear*.
- Härdle, W. and Bowman A. (1988). Bootstrapping in Nonparametric Regression : Local Adaptive Smoothing and Confidence Bands. *J. Amer. Statist. Assoc.*, *83*, 102-110.
- 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*, 86-101.
- Härdle, W. and Scott, D.W. (1988). Smoothing in Low and High Dimensions by Weighted Averaging using Rounded Points. *Statistical Science, submitted*.
- Härdle, W. and Stoker, T. (1988). Investigating Smooth Multiple Regression by the Method of Average Derivatives. *J. Amer. Statist. Assoc.*, *submitted*.
- Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical*

- Heckman, N. (1986). Spline Smoothing in a Partly Linear Model. *J. Royal Statist. Soc. (B)*, 48, 244-248.
- Huber, P.J. (1985). Projection Pursuit (with discussion). *Ann. Statist.*, 13, 435-375.
- Ibragimov, I.A. and Khasm'inskii, R.Z. (1981). Asymptotic Quality Boundaries of Regression Estimation in L_p . *Zap. Nauch. Sem. Lomi.*, 97, 88-101.
- Jones, M.C. and Sibson, R. (1987). What is Projection Pursuit? (with discussion). *J. Royal Statist. Soc. (A)*, 150, 1-39.
- Kallieris, D.; Mattern, R. and Härdle, W. (1986). Belastbarkeitsgrenze und Verletzungsmechanik des angegurteten PKW-Insassen beim Seitenaufprall. Phase II: Ansätze zur Verletzungsprädiktion. *FAT Schriftenreihe 60, Forschungsvereinigung Automobiltechnik e. V. (FAT)*.
- McDonald, J. and Pederson, J. (1986). Computing Environments for Data Analysis, Part 3: Programming Environments. *Laboratory for Computational Statistics, Stanford Technical Report 24*.
- Müller, H.G. (1987). Weighted Local Regression and Kernel Methods for Nonparametric Curve Fitting. *J. Amer. Statist. Assoc.*, 82, 231-238.
- Oldford, R.W. and Peters, S.C. (1985). DINDE : Towards more Statistically Sophisticated Software. *MIT, Technical Report 55*.
- Rice, J.A. (1986). Convergence Rates for Partially Splined Models. *Statistics and Probability Letters*, 4, 203-208.
- Scott, D.W. (1986). Data Analysis in 3 and 4 Dimensions with Nonparametric Density Estimation. in: *Statistical Image Processing*, E. Wegman and D. DePriest, eds., Marcel Dekker, New York, 291-305.
- Scott, D.W. and Terrell, G.R. (1987). Biased and Unbiased Crossvalidation in Density Estimation. *J. Amer. Statist. Assoc.*, 82, 1131-1146.
- Silverman, B.W. (1984). Spline Smoothing: The equivalent Variable Kernel Method. *Ann. Statist.*, 12, 898-916.
- Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. *Chapman and Hall, London*.

Stone, C.J. (1982). Optimal Global Rates of Convergence for Nonparametric Regression. *Ann.Statist.*, 10, 1040-1053.

XploRe (1987). XploRe - a computing environment for eXploratory Regression and density smoothing. *Wirtschaftstheorie II, Universität D-5300 Bonn.*

How Far Are Automatically Chosen Regression Smoothing Parameters From Their Optimum?

WOLFGANG HÄRDLE, PETER HALL, and J. S. MARRON*

We address the problem of smoothing parameter selection for nonparametric curve estimators in the specific context of kernel regression estimation. Call the "optimal bandwidth" the minimizer of the average squared error. We consider several automatically selected bandwidths that approximate the optimum. How far are the automatically selected bandwidths from the optimum? The answer is studied theoretically and through simulations. The theoretical results include a central limit theorem that quantifies the convergence rate and gives the differences asymptotic distribution. The convergence rate turns out to be excruciatingly slow. This is not too disappointing, because this rate is of the same order as the convergence rate of the difference between the minimizers of the average squared error and the mean average squared error. In some simulations by John Rice, the selectors considered here performed quite differently from each other. We anticipated that these differences would be reflected in different asymptotic distributions for the various selectors. It is surprising that all of the selectors have the same limiting normal distribution. To provide insight into the gap between our theoretical results and these simulations, we did a further Monte Carlo study. Our simulations support the theoretical results, and suggest that the differences observed by Rice seemed to be principally due to the choice of a very small error standard deviation and the choice of error criterion. In the example considered here, the asymptotic normality result describes the empirical distribution of the automatically chosen bandwidths quite well, even for small samples.

KEY WORDS: Bandwidth selection; Curve estimation; Kernel regression.

1. INTRODUCTION

Regression smoothing is a method for recovering the mean function from noisy data Y_1, \dots, Y_n of the form

$$Y_i = m(x_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where the ϵ_i are independent, identically distributed, mean-zero observation errors. There are several methods for estimating the regression function, m , which are closely related to moving averages; that is, to estimate $m(x)$, average the Y_i that have x_i close to x . The width of the neighborhood over which averaging is performed, often called the bandwidth or smoothing parameter, controls the smoothness of the resulting estimate. This is illustrated by Figure 1, which shows a solid curve $m(x)$, with 75 Y_i 's coming from adding simulated noise (the same in all three cases). The dashed lines are three different weighted moving averages, in order of increasing size of bandwidth. More details concerning Figure 1 may be found in Section 4.

It is apparent from Figure 1 that choice of bandwidth is very important to this type of estimation. Note that in Figure 1a the estimate has features of the Y_i 's that would be quite different for another realization of the Y_i 's. This is caused by undersmoothing or taking the window width too small. On the other hand, Figure 1c is clearly over-smoothed, with part of the peak averaged away. In this article, we consider several automated (i.e., data driven)

smoothing parameter (bandwidth) selectors and study the amount of noise inherent to them.

Proposed methods for choosing the bandwidth (window size) are based on estimates of the prediction error. For instance, the cross-validation technique provides estimates of the prediction error based on so-called "leave one out" estimators of the regression function (see Clark 1975; Härdle and Marron 1985a). Several other selectors are based on adjustments of the residual sum of squares, which yield an unbiased estimate of the prediction error (see Craven and Wahba 1979; Härdle and Marron 1985b; Rice 1984).

Section 2 gives the precise formulation of these selectors and contains the theoretical results, which quantify the differences between these and the optimal bandwidth. Remarks pertinent to these theoretical results are in Section 3. Section 4 contains simulation results that give additional insight into the meaning of the theoretical results. Some concluding remarks are in Section 5. The proofs are in the Appendix.

2. BEHAVIOR OF DATA-DRIVEN BANDWIDTHS

To simplify the presentation, assume the design points are equally spaced on the unit interval (i.e., $x_i = i/n, i = 1, \dots, n$) and assume that the ϵ_i have common variance σ^2 . The kernel estimator proposed by Priestley and Chao (1972) is, in this setting,

$$\hat{m}_h(x) = n^{-1}h^{-1} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) Y_i,$$

where h is the bandwidth. The kernel, K , is taken here to be a symmetric, compactly supported probability density with [roughly (see the Appendix for a precise formulation)] a second derivative.

* Wolfgang Härdle is Principal Researcher, Wirtschaftstheoretische Abteilung II, Universität Bonn, D-5300 Bonn 1, West Germany. Peter Hall is Reader in Statistics, Australian National University, Canberra, Australian Capital Territory 2601, Australia. J. S. Marron is Assistant Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27514. The research of Härdle and Hall was supported by U.S. Air Force Office of Scientific Research Grant S-49620-R2-C-0144. Härdle's research was also supported by Deutsche Forschungsgemeinschaft Grants SFB-123 and SFB-303. Marron's research was supported by National Science Foundation Grant DMS-8400602.

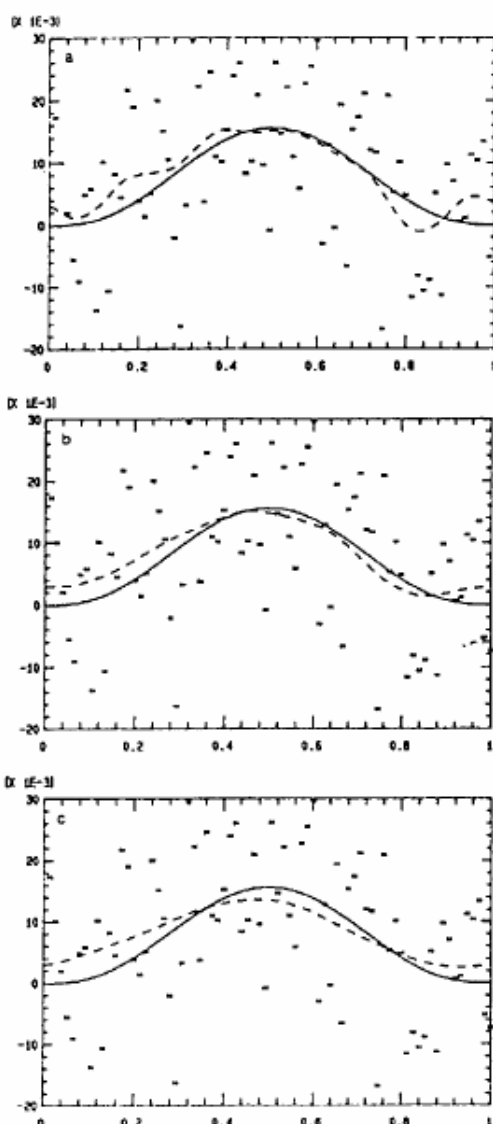


Figure 1. Seventy-Five Simulated Regression Observations From Solid Curve $m(x)$, Plus Dashed Moving-Average Estimates of m With Window Widths: (a) $h = 26$; (b) $h = 39$; (c) $h = 66$.

The optimal bandwidth is taken here to be \hat{h}_0 , the minimizer of the average squared error (ASE)

$$d_A(h) = n^{-1} \sum_{i=1}^n [\hat{m}_h(x_i) - m(x_i)]^2 w(x_i).$$

The weight function w is introduced to allow elimination (or at least significant reduction) of boundary effects (see Gasser and Müller 1979) by taking w to be supported on

a subinterval of the unit interval. If one does not object to the assumption that m is circular (i.e., m and its first two derivatives agree at the endpoints 0 and 1), then w may be taken to be identically 1. Another candidate for the optimal bandwidth is \hat{h}_0 , the minimizer of the mean average squared error (MASE)

$$d_w(h) = E[d_A(h)].$$

We call \hat{h}_0 the optimal bandwidth because it makes \hat{m}_h as close as possible to m for the data set at hand, instead of for the average over all possible data sets. See Scott and Terrell (1987) for a different view of this subject.

How fast may we expect \hat{h}_0 and \hat{h}_0 to tend to 0? If m^* is uniformly continuous, then under the assumption that the moments of the ϵ_i exist, $d_A(h)$ and $d_w(h)$ are both approximately

$$d_w^*(h) = n^{-1} h^{-1} \sigma^2 \int w \int K^2 - h^4 \left(\int u^2 K/2 \right)^2 \int (m^*)^2 w,$$

in the sense that

$$\sup_{h \in H_n} \left(\left| \frac{d_A(h) - d_w^*(h)}{d_w^*(h)} \right| + \left| \frac{d_w(h) - d_w^*(h)}{d_w^*(h)} \right| \right) \rightarrow 0 \quad (2.1)$$

in probability as $n \rightarrow \infty$, where $H_n = [n^{-1+\delta}, n]$, for arbitrarily small $\delta > 0$ (see Marron and Hardle 1986). A consequence of (2.1) is that \hat{h}_0 and \hat{h}_0 are each roughly equal to the unique minimizer of d_w^* , $h_0^* = C_0 n^{-1/5}$, where

$$C_0 = \left(\sigma^2 \int w \right) \left(\int K^2 \right) / \left(\int u^2 K \right)^2 \int (m^*)^2 w \quad (2.2)$$

that is,

$$\hat{h}_0/h_0^*, \hat{h}_0/h_0^* \rightarrow 1 \quad (2.3)$$

in probability. A sketch of the proof of (2.1) and (2.3) is given in the Appendix.

Most bandwidth selectors are based on minimization of some function of h , which is related to the residual sum of squares

$$p(h) = n^{-1} \sum_{i=1}^n \{Y_i - \hat{m}_h(x_i)\}^2 w(x_i).$$

By taking expectations, it can be seen that, as an estimator of the prediction error, $p(h)$ is biased in such a way that its minimizer will not have desirable properties of the type described in (2.3), which is not surprising, since $p(h)$ uses the same set of data to construct an estimate and to assess it. This problem can be handled by multiplying $p(h)$ by a correction factor $\Xi(n^{-1}h^{-1})$, which may be random or nonrandom. Simple examples include (a) generalized cross-validation (Craven and Wahba 1979),

$$\Xi_{CV}(n^{-1}h^{-1}) = (1 - n^{-1}h^{-1}K(0))^{-2};$$

(b) Akaike's information criterion (Akaike 1974),

$$\Xi_{AIC}(n^{-1}h^{-1}) = \exp(2n^{-1}h^{-1}K(0));$$

(c) finite prediction error (Akaike 1970),

$$\Xi_{FPE}(n^{-1}h^{-1}) = (1 + n^{-1}h^{-1}K(0))(1 - n^{-1}h^{-1}K(0));$$

(d) a model selector of Shibata (1981),

$$\Xi_S(n^{-1}h^{-1}) = 1 + 2n^{-1}h^{-1}K(0);$$

and (e) the bandwidth selector T of Rice (1984),

$$\Xi_T(n^{-1}h^{-1}) = (1 - 2n^{-1}h^{-1}K(0))^{-1}.$$

Note that each of these has a Taylor expansion (in the variable $n^{-1}h^{-1}$) of the form

$$\Xi(n^{-1}h^{-1}) = 1 + 2n^{-1}h^{-1}K(0) + O(n^{-2}h^{-2}). \quad (2.4)$$

So it makes sense to define a general bandwidth selector

$$G(h) = \rho(h)\Xi(n^{-1}h^{-1}),$$

where the correction factor $\Xi(n^{-1}h^{-1})$ is of the form (2.4). Other bandwidth selectors are also of essentially this form, but it takes more work to see this. An important example is the cross-validation (CV) function introduced by Clark (1975) (in this setting):

$$CV(h) = n^{-1} \sum_{i=1}^n \{Y_i - \hat{m}_i(x_i)\}^2 w(x_i),$$

where $\hat{m}_i(x_i)$ is a "leave one out" version of \hat{m} ; that is, the observation (x_i, Y_i) is left out in constructing \hat{m} . Priestley and Chao (1972) gave a method of adapting \hat{m} to the fact that the x_i are now not equally spaced. We show in the Appendix that

$$CV(h)/\rho(h) = 1 + 2n^{-1}h^{-1}K(0) + O_p(n^{-2}h^{-2}) \quad (2.5)$$

uniformly over $h \in H_n$. Hence $CV(h)$ can also be thought of as a special case of $G(h)$. This last statement can easily be shown, by essentially the same method, to hold also for bandwidth selectors based on unbiased risk estimation, such as

$$R(h) = n^{-1} \sum_{i=1}^n \{ \{Y_i - \hat{m}_i(x_i)\}^2 + n^{-1}h^{-1}K(0)\{Y_i - Y_{i-1}\}^2 I_{|b|>h} \} w(x_i),$$

(see Rice 1984). The aforementioned list of automatic smoothing parameter selectors is not exhaustive (see, e.g., Li 1985, 1987; Mallows 1973).

In view of the asymptotic equivalence of these bandwidth selectors, one would expect their performances to be about the same, at least for large n . Indeed, it can be shown that the minimizers of all of these (let \hat{h} denote a generic one) are asymptotically optimal (i.e., the ratio of loss to minimum loss tends to 1):

$$d_A(\hat{h})/d_A(\hat{h}_0) \rightarrow 1 \quad (2.6)$$

in probability or (nearly equivalently)

$$\hat{h}/\hat{h}_0 \rightarrow 1 \quad (2.7)$$

in probability (see Härdle and Marron 1985a; Rice 1984). A major objective of this article is to study how fast the convergence in (2.6) and (2.7) occurs, with a view toward trying to distinguish the various bandwidth selectors and, in particular, trying to quantify the differences appearing in Rice's (1984) table 1.

The first part of this is accomplished by the following theorem.

Theorem 1. Under the preceding assumptions (summarized at the beginning of the Appendix),

$$n^{3/10}(\hat{h} - \hat{h}_0) \rightarrow N(0, \sigma_1^2)$$

$$n\{d_A(\hat{h}) - d_A(\hat{h}_0)\} \rightarrow C_1\chi^2 \quad (2.8)$$

in distribution, where σ_1^2 and C_1 (defined in the Appendix) are independent of the particular choice of \hat{h} .

Note that by (2.2), (2.3), and (2.7), all of \hat{h} , \hat{h}_n , h_n , and h_n^* are tending to 0 at the rate $n^{-1/5}$. Hence (2.8) says that the relative difference between \hat{h} and \hat{h}_0 is of the very slow order $n^{-3/10}$. Although this rate seems at first glance to be excruciatingly slow, it should not be too disappointing, because it is of the same order as the difference between \hat{h}_0 and h_0 , as demonstrated by our second theorem.

Theorem 2. Under the preceding assumptions (summarized at the beginning of the Appendix),

$$n^{3/10}(h_0 - \hat{h}_0) \rightarrow N(0, \sigma_2^2)$$

and

$$n\{d_A(h_0) - d_A(\hat{h}_0)\} \rightarrow C_2\chi^2$$

in distribution, where σ_2^2 and C_2 are defined in the Appendix.

3. DISCUSSION AND REMARKS

Remark 3.1. An important consequence of Theorems 1 and 2 is that they imply that the "plug-in" method of choosing h [where one substitutes estimates of the unknown parts of (2.2)], even if one knew exactly the unknowns σ^2 and $\int (m'')^2$, has an algebraic rate of convergence no better than that of the \hat{h} 's given before. Hence the additional noise involved in estimating these unknown parts in practice, especially the second derivative part in the case where m is not very smooth, seems to cast considerable doubt on the applicability of the plug-in estimator. A further advantage of the methods of bandwidth selection proposed in this article is that they automatically adapt to the case $m'' = 0$, whereas plug-in methods either are not defined or come up against an artificial upper bound that makes them well-defined but has no practical relevance.

Remark 3.2. Since the bandwidths \hat{h} converge to the optimum \hat{h}_0 at the same algebraic rate as h_0 , it is natural to compare them by studying the asymptotic variances σ_1^2 and σ_2^2 . By comparing σ_1^2 and σ_2^2 in Lemma 4 (see the Appendix) using the Parseval identity, we see $\sigma_1^2 \leq \sigma_2^2$, so h_0 is closer to \hat{h}_n than \hat{h} is in terms of asymptotic variance. But the limit theorems 1 and 2 can be joined to give a

single limit theorem. Hence

$$\liminf_{n \rightarrow \infty} \Pr[d_A(h_n) > d_A(\hat{h})] > 0;$$

that is, for some data sets, \hat{h} will perform better than h_n . The exact form of the asymptotic covariance of this joint limit theorem is given in the Appendix. It can be shown in many interesting special cases (a sufficient condition is that K be in L^2 and have a nonnegative Fourier transform), including the setting used for the simulation study in Section 4, that this covariance is negative. Observe that this tempers Remark 3.1 by implying that \hat{h} will tend to be on the side of h_n that is away from h_0 . Another consequence of the joint limit theorem is that the bandwidth parts of Theorems 1 and 2 can be added to give Rice's (1984) theorem 2.3.

Remark 3.3. When we did these calculations we were surprised to note that the asymptotic variance σ_1^2 is independent of the particular function $\Xi(n^{-1}h^{-1})$, especially in view of the simulations of Rice (1984). In Section 4 we see that the phenomena observed by Rice are mostly caused by his particular setting, and often these selectors are not so different.

Remark 3.4. A technical advantage of Theorems 1 and 2 over previous results of this type (see Hall and Marron 1987a; Rice 1984) is that the range of bandwidths under consideration has been extended from $[an^{-1/3}, bn^{-1/3}]$ to $[n^{-1+\epsilon}, n]$. This provides more security and theoretical underpinning for consideration of h both large and small. This range is reasonable because $h \approx n^{-1}$ corresponds to no smoothing at all, and $h \approx 1$ corresponds to averaging over the entire sample. See Nolan and Pollard (1987) for techniques to extend this range even further.

Remark 3.5. Several extensions of Theorems 1 and 2 are straightforward. These include the following:

1. The assumption that the errors are identically distributed can be relaxed to the assumption that ϵ_i has variance $V(x_i)$, where the function V is uniformly continuous. The only change in the results is in the constants; for example, in C_0 the expression $\sigma^2 \int w$ is replaced by $\int Vw$. Similar replacements are easily calculated for σ_1^2, σ_2^2 , and the other expressions given in the Appendix.
2. The design points x_i need not be univariate. In the multivariate case where the x_i have dimension p , the exponents of convergence in the first parts of Theorems 1 and 2 change from $3/10$ to $(p + 2)/[2(p + 4)]$; the second parts remain the same except for the values of C_1 and C_2 . The changes in the constants are easily calculated. Since it is unlikely to have an equally spaced design in this case, the estimators discussed in extension 4 (following) would be more appropriate here.
3. The kernel K can be allowed to take on negative values to exploit the well-known higher rates of convergence possible in that case. In particular, if K is of order k , that is,

$$\int K = 1, \quad \int xK = \dots = \int x^{k-1}K = 0, \quad \int x^k K > 0,$$

and if m has a uniformly continuous k th derivative, then the exponents of convergence in the first parts of Theorems 1 and 2 change from $3/10$ to $3/2(2k + 1)$, whereas (again) the second parts are essentially unchanged and (again) the new constants are easily calculated. The rates given here and in the preceding extension 2 are rather paradoxical because they say that when m is easier to estimate, it is harder to select \hat{h} . See Marron (1986) for a discussion of this.

4. The Priestley-Chao \hat{m} may be replaced by several other kernel-type estimators, including those of Nadaraya (1964), Watson (1964), and Gasser and Müller (1979).

Remark 3.6. By estimating the unknown parts in the expressions in the Appendix for σ_1^2 , Theorem 1 can be used to provide approximate confidence intervals for \hat{h}_0 , which can be useful for suggesting a reasonable range of bandwidths to consider for choosing the smoothing parameter by an interactive trial and error approach. Of course the comments in Remark 3.1 serve to put some substantial limitations on this approach.

Remark 3.7. It is conjectured that the slow relative rate of convergence of \hat{h} to \hat{h}_0 that was observed in Section 2 is in fact the best possible in the minimax sense. This was made precise (in the related density estimation setting) by Hall and Marron (1987b). The implication is that although all of the procedures given in this article give a slow rate of convergence, there is no point in searching for a procedure that gives a faster rate [although, of course, improvements in the constant coefficient are certainly possible; see Scott and Terrell (1987) for an interesting possibility in this direction].

4. SIMULATIONS

Following Rice (1984, sec. 4), we generated 100 samples of $n = 75$ pseudorandom normal variables, ϵ_i , with mean 0 and standard deviation $\sigma = .0015$. These were added to the regression curve $m(x) = x^3(1 - x)^3$, which has the nice effect of allowing a circular design (i.e., when estimating near $i = 1$, for $i \leq 0$, let $x_i = x_{-i}$, and similarly at the other end) to eliminate boundary effects. The kernel function was taken to be

$$K(x) = (15/8)(1 - 4x^2)^2 1_{[-.5, .5]}(x).$$

Table 1 contains the results when the selectors introduced in Section 2 were used to find \hat{h} . The entries show the number of times out of 100 that either the ratio of MASE's (the actual d_w as opposed to $d_{\hat{w}}$).

$$d_w(\hat{h})/d_w(h_n), \tag{4.1}$$

or the ratio of ASE's,

$$d_A(\hat{h})/d_A(h_n), \tag{4.2}$$

exceeded the value of the column heading.

The Rice rows are copied from the study of Rice (1984), who only worked with d_w , and are included to provide some assurance against programming errors and to allow some understanding of how things change when one works

Table 1. Number of Exceedances of the Column Headings (by ratios of error criteria) for the Various Bandwidth Selectors: 100 Data Sets of Size 75, With Error Standard Deviation $\sigma = .0015$

		1.05	1.1	1.2	1.4	1.6	1.8	2	4	6	8
T	Rice	28	17	4	1	0	0	0	0	0	0
	MASE	30	13	3	1	0	0	0	0	0	0
	ASE	63	51	29	13	10	7	2	0	0	0
CV	Rice	33	22	7	1	0	0	0	0	0	0
	MASE	38	22	3	2	0	0	0	0	0	0
	ASE	73	53	30	14	10	7	3	0	0	0
R	Rice	36	21	6	3	1	0	0	0	0	0
	MASE	32	19	8	6	4	3	3	0	0	0
	ASE	65	52	32	17	12	10	8	2	0	0
GCV	Rice	33	21	8	4	1	1	0	0	0	0
	MASE	34	17	12	7	5	5	4	0	0	0
	ASE	64	50	35	19	14	11	10	2	1	0
FPE	Rice	45	38	28	25	22	21	21	21	18	18
	MASE	40	24	20	11	8	7	0	0	0	0
	ASE	65	51	38	20	16	12	12	2	0	0
AIC	Rice	46	27	18	16	14	13	11	4	4	4
	MASE	40	24	20	14	11	9	0	0	0	0
	ASE	64	51	38	22	17	12	12	2	1	0
S	Rice	66	57	50	43	42	42	41	41	19	19
	MASE	68	63	56	47	45	43	0	0	0	0
	ASE	75	69	62	56	43	32	25	5	1	0

NOTE: The "Rice" rows are from Rice (1984). The "MASE" and "ASE" rows are for the ratios (4.1) and (4.2), respectively.

with a different data set. The MASE rows show our reproduction of Rice's simulations. Note that they correspond about as one might expect, except for the selectors AIC, FPE, and S. These are much different because these selectors have a trivial minimum at $h = n^{-1}K(0) = .025$, the "no smoothing" point, where $m(x_i) = Y_i$.

The poor performance of S may be somewhat surprising, since S satisfies (2.4) with $O(n^{-1}h^{-2}) = 0$. This and the poor performances of FPE and AIC can be understood by observing that (in the present setting) $p(h)$ has a second-order 0 at the no-smoothing point. Note that GCV counters this by using a correction factor, $\Xi_{GCV}(n^{-1}h^{-1})$, with a second-order pole at this same point, and T achieves a similar, although stronger effect with a pole to the right, at $h = 2n^{-1}K(0) = .05$. On the other hand, FPE has only a single pole, whereas AIC and S have no poles, at the no-smoothing point.

We believe these trivial minima at the no-smoothing point did not cause a complete disaster, for FPE, AIC, and S, in Rice's study because the iterative minimizer that he used would typically go to a local minimum that provided a reasonable bandwidth estimate [except when the curve $G(h)$ had no local minima, which occurred in our study about the same number of times as Rice reported a ratio exceeding 8]. We could not duplicate this because we used a minimizer that evaluated the function on a grid and took the minimizer, and hence always gave the no-smoothing point as the minimum. This choice of minimization algorithm was motivated by concern over local minima. We did indeed discover local minima, up to 5 in the worst case. To make the selectors FPE, AIC, and S work at all well, we minimized them over only the interval $h \in [.1, 1]$, where this interval was chosen by examining the

functions of h and asking, "What range will make them work well on the average?" Of course this cannot be done in practice, but we feel it is instructive to see what happens when we give these selectors their best chance. This local minima problem may also contribute to the difference between Rice's results and ours for the other selectors as well.

Table 1's ASE rows show how the same set of selected bandwidths performed when the d_A ratio was used instead of the d_W ratio. The same rough ordering of selectors is preserved, but the relative differences are much less.

We feel that the main reason the simulation results of Rice (1984) showed such a big difference in the performance of these selectors is that σ was chosen to be only .0015. When this is plugged into asymptotic formulas such as d_W^* , it indicates that this setting requires what perhaps may be thought of as an unnaturally small amount of smoothing; that is, reasonable h 's may tend to be smaller than "usual." A possible interpretation of this is that the setting chosen by Rice may be one where the asymptotics of the theory presented here take a rather long time to "take effect." To see if this was actually the case, we repeated the study with $\sigma = .011$ (chosen because it makes the minimizer of d_W^* roughly $\frac{1}{2}$). The results are in Table 2, the format of which is similar to that of Table 1.

The same general ordering of selectors that Rice observed still holds up here (except that both here and in Table 1, FPE and AIC have traded places, and in Table 2 GCV seems slightly better than R), but we feel our asymptotic result that the performance of these selectors is roughly the same holds up quite well (with the possible exception of S) in the present setting. Hence it seems to us that the dramatic differences in selectors observed by Rice may be expected to disappear in situations not slanted toward undersmoothing.

To investigate this further, we repeated the $\sigma = .011$ simulations with n increased to 500. A representative subset of the selectors consisting of T, GCV, AIC, and S was considered. For each data set in this setting, the selected bandwidths were essentially the same (i.e., within .01 to

Table 2. Number of Exceedances by Ratios of Error Criteria: 100 Data Sets of Size 75, With Error Standard Deviation $\sigma = .011$

		1.05	1.1	1.2	1.4	1.6	1.8	2	4	6	8
T	MASE	49	32	20	7	4	3	3	0	0	0
	ASE	70	59	48	33	27	22	16	5	5	4
CV	MASE	49	35	26	6	5	5	5	2	2	0
	ASE	75	64	53	35	29	24	17	6	6	4
GCV	MASE	48	36	24	13	9	7	7	3	2	0
	ASE	75	65	50	37	30	25	19	8	8	6
R	MASE	50	37	28	15	11	11	10	4	0	0
	ASE	73	63	49	36	32	26	21	12	10	8
FPE	MASE	50	39	31	19	16	13	10	2	0	0
	ASE	76	66	49	46	32	27	22	12	9	9
AIC	MASE	50	41	33	20	17	13	10	4	0	0
	ASE	76	65	50	42	33	28	22	12	10	9
S	MASE	63	57	48	37	34	31	30	20	0	0
	ASE	82	70	60	54	48	43	38	23	14	12

Table 3. Numbers of Exceedances by Ratios of Error Criteria: 100 Data Sets of Size 500, With Error Standard Deviation $\sigma = .011$

		1.05	1.1	1.2	1.4	1.6	1.8	2	4	6	8
GCV	MASE	44	29	18	13	8	5	4	1	0	0
	ASE	74	62	50	39	29	23	17	7	6	3

.02 of each other), and they were ordered as

$$\hat{h}_1 \leq \hat{h}_{AIC} \leq \hat{h}_{GCV} \leq \hat{h}_T.$$

We feel the similarity is because the differences between the selectors only show up in the $O(n^{-2}h^{-2})$ part of (2.4), and this gets small very rapidly with increasing n . Note that the no-smoothing point here, $h = n^{-1}K(0) = .00375$, is well below the left endpoint of the grid we considered in our minimization algorithm (.01, .02, . . . , .60). Table 3 shows the analog of Tables 1 and 2 in this setting. Since the selectors all give nearly the same bandwidth, only the ratios for GCV are shown. These are only slightly better than the GCV part of Table 2 and are roughly comparable with the T part of Table 2. This seems to illustrate the very slow rates of convergence given in Theorems 1 and 2.

Figure 1 gives an indication of what our results mean in terms of the actual curves, for one of the 100 data sets (with $\sigma = .011$ and $n = 75$). Recall that in each panel the solid curve is $m(x)$. The curve of dashes in Figure 1a is $\hat{m}_h(x)$ with $h = .26$, the minimizer of S for that data set. In Figure 1b, the curve of dashes is $\hat{m}_h(x)$ with $h = .39$, the minimizer of ASE. The dashed curve in Figure 1c is $\hat{m}_h(x)$ with $h = .66$, the minimizer of each of the other automatic selectors. This particular data set was chosen because, though it was far from the worst, most of the other data sets gave better performances of the automatic selectors.

To investigate how well our central limit theorems were describing the situation in the finite sample case, Epanechnikov kernel density estimates (see Rosenblatt 1971), with bandwidth chosen by the cross-validation method of Rudemo (1982) and Bowman (1984), were constructed based on the samples from the distributions of (a) $\hat{h}_0 - h_0$, (b) $\hat{h}_T - h_0$, and (c) $\hat{h}_T - h_0$, where \hat{h}_T is the minimizer of Rice's $T(h)$. Figure 2 shows these curves as solid lines, with the dashed lines showing a parametric normal fit (i.e., the normal density with the sample mean and variance) for each of these observation sets.

Observe that the scale in Figure 2a is twice that of Figure 2b. This shows that, in Theorems 1 and 2, σ_2 is nearly half the size of σ_1 . (See the discussion of Table 6 for more on this topic.)

Figures 2a and 2b demonstrate that even for only $n = 75$, the asymptotic normality of Theorems 1 and 2 holds to what we feel is a remarkable degree. Furthermore, they also do a good job of illustrating the departure of the data distributions from normality; in particular, there is a skewness in the direction of a slightly heavy right tail (the height of the peak is also lower than the parametric fit, which is expected from a kernel density estimate).

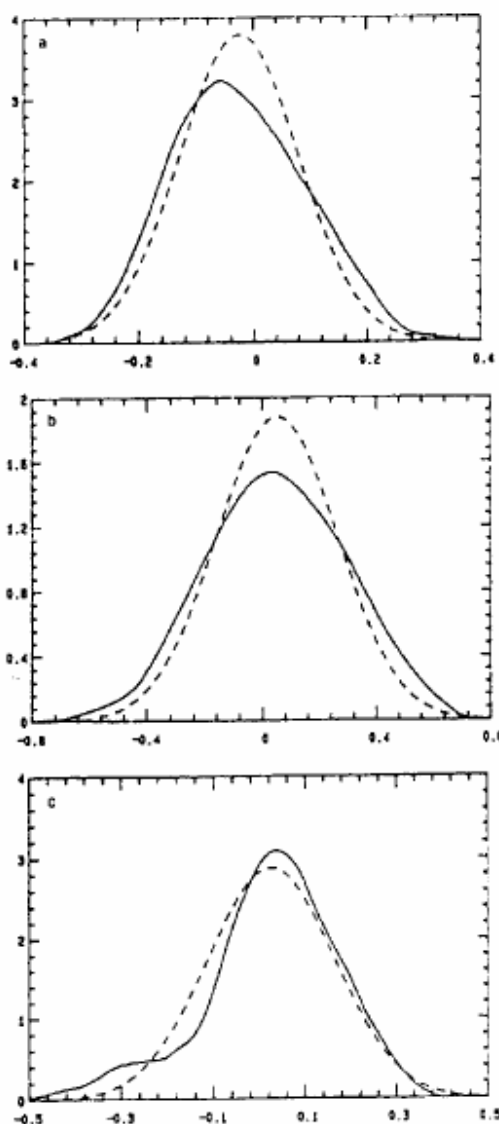


Figure 2. Kernel-Density Estimates (solid curve) and Parametric Normal Fit (dashed curve) for the Data Sets of 100 Simulated Observations From the Distributions of (a) $\hat{h}_0 - h_0$, (b) $\hat{h}_T - h_0$, and (c) $\hat{h}_1 - h_0$.

Figure 2c is remarkable, both for the shape of the left side and because it is actually taller than the mode of the parametric normal fit, indicating substantial leptokurtosis. A possible interpretation of this is that the normal asymptotic distribution, which can be computed for $\hat{h}_T - h_0$, takes a much larger sample size for a realistic description of what is happening. In view of the computations in the Appendix, this could be because the terms that drive the limiting distributions of $\hat{h}_0 - h_0$ and $\hat{h}_T - h_0$ have a simple structure as a sum of uncorrelated martingales, whereas

Table 4. Measures of Departures From Normality for Simulations From Asymptotically Normal Distributions

	Skewness	Kurtosis	D:Normal	Prob. > D
n = 75				
$\hat{h}_A - \hat{h}_0$.29018	-.47242	.08307	.088
$\hat{h}_T - \hat{h}_0$	-.07598	-.45908	.04796	>.15
$\hat{h}_S - \hat{h}_0$	-.80269	.91670	.11471	<.01
n = 500				
$\hat{h}_A - \hat{h}_0$.26145	-.81649	.10580	<.01
$\hat{h}_T - \hat{h}_0$	-.03317	-.87094	.05833	>.15
$\hat{h}_S - \hat{h}_0$	-.85577	.22605	.13198	<.01

their sum, $\hat{h}_T - \hat{h}_0$, has a much more complicated structure.

To allow a more conventional analysis of the departures from normality of these three distributions, Table 4 summarizes the usual statistics.

These were computed by the SAS procedure UNIVARIATE. Observe that the three pictures of Figure 2 show quite clearly the skewness and kurtosis computed for the case $n = 75$. The third column gives the Kolmogorov distance to the best Gaussian fit for each difference. The fourth column contains the observed significance of the Kolmogorov test of the hypothesis that the data are indeed normal. Here again the statement (from looking at the pictures) that the data sets of Figures 2a and 2b are much closer to normally distributed is supported by the computations. Another interesting feature of Table 4 is that the Kolmogorov distance increases when n is increased to 500.

Table 5 adds some insight into how the different selectors compare with each other for the data of Tables 2 and 3. The first two columns contain the sample mean and standard deviation of the bandwidths minimizing the quantity listed at the left. Note that the mean for the automatically selected bandwidths is nearly a decreasing function of the ordering given in Table 2. Also, the selector whose mean matches best with \hat{h}_0 is the rather poorly performing FPE, which is not surprising in view of Rice's

Table 5. Summary Statistics for Automatically Chosen and Optimal Bandwidths From 100 Data Sets

\hat{h}	$\mu_n(\hat{h})$	$\sigma_n(\hat{h})$	$\rho_n(\hat{h}, \hat{h}_0)$	$\rho_n(\hat{h}, \hat{h}_{GCV})$
n = 75				
ASE	.51000	.10507	1.00000	-.46602
T	.56035	.13845	-.50654	.85076
CV	.57297	.15411	-.47494	.87106
GCV	.52929	.16510	-.46602	1.00000
R	.52482	.17852	-.40540	.83565
FPE	.49790	.17846	-.45879	.76829
AIC	.49379	.18169	-.46472	.76597
S	.39435	.21350	-.21965	.52915
n = 500				
ASE	.36010	.07198	1.00000	-.31463
T	.32740	.08558	-.32243	.99869
GCV	.32580	.08864	-.31463	1.00000
AIC	.32200	.08865	-.30113	.97373
S	.31840	.08886	-.29687	.97308

comment that our error criteria d_A and d_W penalize more heavily for h small, or in other words the good performance of the selector T shows up quite well in Table 5 as a bias toward \hat{h} too big. Another interesting feature is that the standard deviation of the selected bandwidths increases as a function of Table 2 performance.

The last two columns of Table 5 show the sample correlation coefficients for the selected bandwidth with \hat{h}_0 and \hat{h}_{GCV} , the minimizer of GCV (chosen because it seemed the most representative), respectively. In interpreting these numbers keep in mind that all minima were computed for the same 100 sets of 75 (500, respectively) observations. The negative correlations of \hat{h}_0 with the \hat{h} cause the ASE values to be much worse than the MASE values in Tables 1, 2, and 3; they also cause the much larger scale on the x axis of Figure 2b in comparison to Figures 2a and 2c. The high correlation between \hat{h}_{GCV} and each \hat{h} is of course expected because of the similar character of these bandwidth selectors. A paradoxical feature is that the correlation of the various \hat{h} 's with \hat{h}_0 tends to be more negative for the bandwidths that perform better.

Table 6 provides another way of checking how well the asymptotic theory corresponds to the simulations. For both sample sizes, the first lines give the observed values (from our 100 data sets) of the given standard deviations and correlation. The second lines give the values of these predicted by the asymptotic theory. These are all quite close when compared to what one might expect from reading the proofs in the Appendix.

Note that $\sigma(\hat{h}_{GCV} - \hat{h}_0)$ and $\sigma(\hat{h}_0 - \hat{h}_0)$ are rescalings of σ_1 and σ_2 , respectively. So Table 6 more precisely quantifies the observation made in the discussion of Figure 2, that σ_1 appeared to be roughly twice the size of σ_2 .

For theoretical results with simulations in settings related to the present, see Scott and Terrell (1987) and Wahba (1985).

5. CONCLUSIONS

We believe the most important lesson to be learned from these results is that even though automatic smoothing methods contain a good deal of useful information, they are subject to quite a bit of noise. Hence it seems reasonable first to choose a bandwidth by a method such as Rice's T and then to look at plots of the estimated regression function for that bandwidth as well as ones on either side (the confidence intervals described in Remark 3.6 could be useful here).

For the problem of which bandwidth selectors to use, we have a slight preference for T , but the statement of

Table 6. Comparison of Empirically Observed Statistics From Simulations With Values Predicted by Asymptotic Theory

	$\sigma(\hat{h}_{GCV} - \hat{h}_0)$	$\sigma(\hat{h}_0 - \hat{h}_0)$	$\rho(\hat{h}_{GCV}, \hat{h}_0)$
Simulations, $n = 75$.1738	.1051	-.4660
Theory, $n = 75$.2057	.1348	-.2729
Simulations, $n = 500$.1289	.0719	-.3146
Theory, $n = 500$.1146	.0764	-.2729

Rice (1984) that "these results are suggestive, but far from conclusive" (p. 1229) seems pertinent here as well. One recommendation that clearly can be made is that for kernel regression estimation, FPE, AIC, and *S* should not be used because of their trivial minimum at the no-smoothing point (note that these were designed for model selection and not kernel regression estimation).

APPENDIX: ASSUMPTIONS AND PROOFS

Summary of Assumptions for Theorems 1 and 2. (a) The errors, ϵ_i , are iid with mean 0 and all other moments finite. (b) The kernel function, K , is a symmetric, compactly supported probability density with a Hölder continuous second derivative. (c) The regression function, m , has a uniformly continuous, integrable second derivative.

Proof of Theorems 1 and 2. The proof of Theorem 2 is based on the expansion

$$0 = d_n^*(\hat{h}_n) = d_n^*(h_n) + D^*(\hat{h}_n) = (h_n - h_0)d_n^*(h^*) + D^*(\hat{h}_n), \tag{A.1}$$

where h^* is between \hat{h}_n and h_n , where $D(h) = d_n(h) - d_n^*(h)$, and where D^* , d_n^* , and d_n denote the derivatives with respect to h of D , d_n^* , and d_n , respectively. The proof of Theorem 1 is based on the expansion

$$G(h) = [d_n(h) + \hat{\sigma}^2 + \delta_1(h)] \times [1 + 2n^{-1}h^{-1}K(0) + O_p(n^{-2}h^{-2})], \tag{A.2}$$

where $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n [m(x_i) - Y_i]^2 w(x_i)$ and $\delta_1(h) = 2n^{-1} \sum_{i=1}^n [m_h(x_i) - m(x_i)][m(x_i) - Y_i]w(x_i)$. Let $\delta_1(h) = \delta_1(h) + 2n^{-1}h^{-1}K(0)\hat{\sigma}^2$.

To analyze the expressions (A.1) and (A.2), we use the following lemmas. Notation used there includes

$$r_n(h) = n^{-1}h^{-1} + h^2, \quad L(u) = -uK'(u),$$

$$K_n(u) = h^{-1}K(u/h), \quad L_n(u) = h^{-1}L(u/h),$$

$$b_n(x) = n^{-1} \sum_{i=1}^n K_n(x - x_i)m(x_i) - m(x),$$

and

$$c_n(x) = n^{-1} \sum_{i=1}^n L_n(x - x_i)m(x_i) - m(x).$$

Lemma 1. For $l = 1, 2, \dots$ there is a constant C_l so that

$$\sup_{h \in H_n} E[r_n(h)^{-1}h^{2l}D^*(h)]^2 \leq C_l \tag{A.3}$$

and

$$\sup_{h \in H_n} E[r_n(h)^{-1}h^{2l}\delta_1(h)]^2 \leq C_l. \tag{A.4}$$

Furthermore, there is an $\eta_l > 0$ and a constant C_l so that

$$E[r_n(h)^{-1}h^{2l}[D^*(\hat{h}) - D^*(h^*)]^2] \leq C_l \left(\frac{h^* - h}{h}\right)^{\eta_l} \tag{A.5}$$

and

$$E[r_n(h)^{-1}h^{2l}[\delta_1(\hat{h}) - \delta_1(h^*)]^2] \leq C_l \left(\frac{h^* - h}{h}\right)^{\eta_l} \tag{A.6}$$

whenever $h, h^* \in H_n$ with $h \leq h^*$ and $|h - h^*|/h \leq 1$.

Lemma 2. For any $\eta_l \in (0, 1/10)$,

$$\sup_{h \in H_n} [r_n(h)^{-1}h^{2l}[D^*(\hat{h}) + \delta_1(\hat{h})]^2] = O_p(n^{-\eta_l}). \tag{A.7}$$

Furthermore, if $h_n \rightarrow h^*$ tends to a constant, then

$$\sup_{h \in H_n} [r_n(h)^{-1}h^{2l}[D^*(\hat{h}) - D^*(h_n)] + |\delta_1(\hat{h}) - \delta_1(h_n)|] = o_p(1). \tag{A.8}$$

Lemma 3. For some $\epsilon > 0$, $|\hat{h}_n - h_0| + |\hat{h} - h_0| = O_p(n^{-\epsilon/2})$.

Lemma 4.

$$n^{\epsilon/2} \begin{bmatrix} D^*(\hat{h}_n) \\ \delta_1(\hat{h}_n) \end{bmatrix} \rightarrow N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_w \\ \sigma_w & \sigma_1^2 \end{bmatrix} \right)$$

in distribution, where (letting $*$ denote convolution)

$$\sigma_1^2 = \frac{8}{C_1^2} \sigma^2 \left[\int w^2 \right] \left[\int (K^*K - K^*L)^2 \right] + 4C_2^2 \sigma^2 \left[\int u^2 K \right]^2 \left[\int (m^*)^2 w^2 \right],$$

$$\sigma_2^2 = \frac{8}{C_1^2} \sigma^2 \left[\int w^2 \right] \left[\int (K - L)^2 \right] + 4C_2^2 \sigma^2 \left[\int u^2 K \right]^2 \left[\int (m^*)^2 w^2 \right],$$

and

$$\sigma_w = \frac{-8}{C_1^2} \sigma^2 \left[\int w^2 \right] \left[\int (K - L)(K^*K - K^*L) \right] - 4C_2^2 \sigma^2 \left[\int u^2 K \right]^2 \left[\int (m^*)^2 w^2 \right].$$

Lemma 5. For any constants a and b ,

$$\sup_{h \in H_n} |D^*(h)| = o_p(n^{-\epsilon/2}).$$

To finish the proof of the first part of Theorem 2, note first that

$$n^{\epsilon/2} d_n^*(h^*) \rightarrow C_1, \tag{A.9}$$

where $C_1 = (2/C_1)\sigma^2[\int K^2][\int w] + 3C_2[\int u^2 K^2][\int (m^*)^2 w]$. It follows from (A.1), (A.8), and Lemma 3 that $D^*(\hat{h}_n) = D^*(h_n) + o_p(n^{-\epsilon/2})$. Hence, by Lemma 4, $n^{\epsilon/2} D^*(\hat{h}_n) \rightarrow N(0, \sigma_1^2)$. Thus, applying Lemma 2 and (A.9) to (A.1) gives

$$0 = (\hat{h}_n - h_0)C_1 n^{\epsilon/2} + D^*(\hat{h}_n) + o_p(n^{-\epsilon/2}), \tag{A.10}$$

from which it follows that $n^{\epsilon/2}(\hat{h}_n - h_0) \rightarrow N(0, \sigma_1^2)$, where $\sigma_1^2 = \sigma_1^2/C_1^2$.

To prove the second part of Theorem 2, note that by Lemma 5,

$$d_n(h_n) - d_n(\hat{h}_n) = 4(h_n - \hat{h}_n)^2 d_n^*(h^*) + o_p(n^{-2}),$$

where h^* is between h_n and \hat{h}_n . Hence $n[d_n(h_n) - d_n(\hat{h}_n)] \rightarrow C_2 \mathcal{Z}$ in distribution, where $C_2 = C_1 \sigma_2^2/2$.

The proof of the first part of Theorem 1 takes slightly more work than the proof of the first part of Theorem 2. For $h \in [an^{-1}, bn^{-1}]$ (where a and b are arbitrary constants), differentiating (A.2) gives

$$0 = G^*(\hat{h}) = [d_n^*(\hat{h}) - \delta_1^*(\hat{h})][1 + O_p(n^{-1})] + [d_n(\hat{h}) + \hat{\sigma}^2 + \delta_1(\hat{h})][-2n^{-1}h^{-2}K(0) + O_p(n^{-1})], \tag{A.11}$$

which may be written as

$$0 = d_n^*(\hat{h}_n) + \delta_1^*(\hat{h}_n) + o_p(n^{-1}). \tag{A.12}$$

Working on (A.12) as in (A.1) and (A.10) gives

$$0 = (\hat{h}_n - h_0)C_1 n^{\epsilon/2} + D^*(\hat{h}_n) + \delta_1^*(\hat{h}_n) + o_p(n^{-\epsilon/2}).$$

which after subtracting (A.10) yields

$$-\delta_2(h_0) = (\hat{h} - \hat{h}_0)C_1n^{-2\alpha} + o_p(n^{-2\alpha}).$$

Hence, by Lemma 4, $n^{1/2}(\hat{h} - \hat{h}_0) \rightarrow N(0, \sigma_1^2)$, where $\sigma_1^2 = \sigma_1^2/C_1^2$.

The proof of the second part of Theorem 1 is so similar to the above that only the result is given: $n[d_2(\hat{h}) - d_2(h_0)] \rightarrow C_2\mathcal{L}$ in distribution, where $C_2 = C_2\sigma_1/2$.

The asymptotic covariance discussed in Remark 3.2 is seen, in the aforementioned way, to be $\sigma_{12} = \sigma_1\omega/C_1$.

Proof of Lemma 1. From here on, for notational simplicity we take $w(x) = 1$. Note that $D_1(h) = -(h/2)D'(h)$ can be expanded into

$$D_1(h) = S_1(h) + S_2(h) + S_3(h), \quad (A.13)$$

where

$$\begin{aligned} S_1 &= S_{11} - S_{12}, & S_2 &= S_{21} + S_{22}, & S_3 &= S_{31} - S_{32}, \\ S_{11}(h) &= 2n^{-1} \sum_{i,j} \left[n^{-1} \sum_{i,j} K_h(x_i - x_j)K_h(x_i - x_j) \right] \epsilon_i \epsilon_j, \\ S_{12}(h) &= n^{-1} \sum_{i,j} \left\{ n^{-1} \sum_{i,j} [K_h(x_i - x_j)L_h(x_i - x_j) \right. \\ &\quad \left. + L_h(x_i - x_j)K_h(x_i - x_j)] \right\} \epsilon_i \epsilon_j, \\ S_{21}(h) &= n^{-1} \sum_{i,j} \left[n^{-1} \sum_{i,j} K_h(x_i - x_j)[2b_h(x_i) - c_h(x_i)] \right] \epsilon_i, \\ S_{22}(h) &= n^{-1} \sum_{i,j} \left[n^{-1} \sum_{i,j} L_h(x_i - x_j)[-b_h(x_i)] \right] \epsilon_i, \\ S_{31}(h) &= n^{-1} \sum_{i,j} \left[n^{-1} \sum_{i,j} K_h(x_i - x_j)^2 \right] (\epsilon_i^2 - \sigma^2), \end{aligned}$$

and

$$S_{32}(h) = n^{-1} \sum_{i,j} \left[n^{-1} \sum_{i,j} K_h(x_i - x_j)L_h(x_i - x_j) \right] (\epsilon_i^2 - \sigma^2).$$

Note that

$$\begin{aligned} -\delta_2(h)/2 &= n^{-1} \sum_{i,j} \epsilon_i \left[n^{-1} \sum_{i,j} K_h(x_i - x_j) \epsilon_j \right. \\ &\quad \left. + b_h(x_i) - n^{-1}h^{-1}K(0)\epsilon_i \right] \\ &= 2n^{-1} \sum_{i,j} \sum_{i,j} K_h(x_i - x_j)\epsilon_i \epsilon_j + n^{-1} \sum_{i,j} b_h(x_i)\epsilon_i. \end{aligned}$$

Now, as before write

$$\delta_2(h) = (h/2)\delta_2'(h) = T_1 + T_2, \quad (A.14)$$

where

$$T_1 = 2n^{-1} \sum_{i,j} \sum_{i,j} [K_h(x_i - x_j) - L_h(x_i - x_j)]\epsilon_i \epsilon_j,$$

and $T_2 = n^{-1} \sum_{i,j} [b_h(x_i) - c_h(x_i)]\epsilon_i$. The proof of (A.3) is very similar in spirit to that of (A.5), but it is easier, so only the proof of (A.5) will be given. First, write

$$S_{11}(h) = n^{-1} \sum_{i,j} \sum_{i,j} A_{ij}(h)\epsilon_i \epsilon_j,$$

where $A_{ij}(h) = n^{-1} \sum_{i,j} K_h(x_i - x_j)K_h(x_i - x_j)$. Note that the sum over l ranges only over at most a multiple of nh indexes, due to the compactness of support of K . Standard arguments

show that

$$|A_{ij}(h) - A_{ij}(h')| \leq Ch^{-1}|(h - h')/h|,$$

where here and following C denotes a generic constant. By theorem 2 of Whittle (1960), for a generic constant C ,

$$\begin{aligned} E[|r_n(h)^{-1}h^{-1/2}[S_{11}(h) - S_{11}(h')]|^2] \\ \leq Cr_n(h)^{-2}h^{-1}n^{-1/2} \left[\sum_{i,j} |A_{ij}(h) - A_{ij}(h')|^2 \right] \\ \leq Cr_n(h)^{-2}h^{-1}n^{-1/2}(n^2h^2|(h - h')/h|^2h^{-2})^l \\ \leq C|(h - h')/h|^{2l}. \end{aligned}$$

By similar bounds on the other terms in the decomposition of $D_1(h)$, it may be seen that

$$E[r_n^{-1}(h)h^{1/2}[D_1(h) - D_1(h')]|^2] \leq C|(h' - h)/h|^{2l},$$

from which (A.5) is an easy consequence. The proofs of (A.4) and (A.6) are the same in spirit and hence are omitted.

Proof of Lemma 2. By Hölder continuity and compactness of the support of K and L , there is a $\rho > 0$ large enough so that

$$\sup_{|h-h'| \leq n^{-\rho}} |D'(h) - D'(h')| = O(n^{-\rho}).$$

Hence it is sufficient to restrict the supremum in the statement of Lemma 2 to a set H'_n , which is a subset of H_n , so that $\#(H'_n) \leq n^{\rho+1}$ and so that for any $h \in H_n$ there is an $h' \in H'_n$ with $|(h - h')/h| \leq n^{-\rho}$. By Bonferroni's inequality, Whittle's inequality, and (A.3),

$$\begin{aligned} \Pr\left\{ \sup_{h \in H_n} |r_n(h)^{-1}h^{1/2}n^{-1/2}D'(h)| > \epsilon \right\} \\ \leq \#(H'_n) \sup_{h \in H'_n} E[|r_n^{-1}(h)h^{1/2}n^{-1/2}D'(h)|^2] \\ \leq Cn^{\rho+1}(n^{-\rho})^{2l} \rightarrow 0 \end{aligned}$$

by taking l sufficiently large, which proves the D part of (A.7). The proofs of the δ_2 part of (A.7) and of (A.8) use the same type of partitioning argument with (A.4), (A.5), and (A.6), respectively.

Proof of Lemma 3. By (2.3) and (A.7)

$$d'_2(h_0) = d'_2(\hat{h}_0) - d'_2(\hat{h}_0) = d'_2(h_0) - d'_2(\hat{h}_0) + o_p(n^{-2\alpha}).$$

But by (A.7)

$$d'_2(h_0) = D'(h_0) + O_p(n^{-2\alpha}).$$

Thus, setting $\epsilon = -\eta_2 + 1/10$,

$$d'_2(\hat{h}_0) - d'_2(h_0) = O_p(n^{-2\alpha}).$$

But $d'_2(h_0) - d'_2(\hat{h}_0) = (h_0 - \hat{h}_0)d''_2(h^*)$; so, by (A.9), $|\hat{h}_0 - h_0| = O_p(n^{-1/2-\alpha})$. By the same method it can be shown that $|\hat{h}_0 - h_0| = O_p(n^{-1/2-\alpha})$.

Proof of Lemma 4. This proof is very close to the proofs of lemmas 3.4 and 3.5 of Hall and Marron (1987a) [and it makes use of a martingale central limit theorem of the type developed by Hall (1984)]. The major difference is that the variances of the terms in the expansion (A.13) satisfy

$$\begin{aligned} n^2h_0 \text{var}(S_1(h_0)) &\rightarrow 2\sigma^4 \left[\int (K^*K - K^*L)^2 \right] \left[\int w^2 \right], \\ nh_0^2 \text{var}(S_2(h_0)) &\rightarrow \sigma^2 \left[\int w^2K \right]^2 \left[\int (m^*n)^2 \right], \end{aligned}$$

and

$$\text{var}(S_3(h_0)) = O_p(n^{-1/2});$$

and for (A.14),

$$n^2 h_n \text{var}(T_1(h_n)) \rightarrow 2\sigma^2 \left[\int (K - L)^2 \right] \left[\int w^2 \right]$$

and

$$n h_n^2 \text{var}(T_2(h_n)) \rightarrow \sigma^2 \left[\int u^2 K \right]^2 \left[\int (m^* w)^2 \right].$$

A little care must be taken with the martingale structure in the case when w is not identically 1, but this case is handled by writing (e.g., in the part involving S_{11})

$$\sum_{i=1}^n \sum_{j=1}^n = \sum_{i=1}^n \sum_{j=i}^n + \sum_{i=1}^n \sum_{j=1}^{i-1}$$

The form of the asymptotic covariance σ_w follows from $n^2 h_n \text{cov}(S_1(h_n), T_1(h_n)) \rightarrow$

$$2\sigma^2 \left[\int (K - L)(K^* K - K^* L) \right] \left[\int w^2 \right]$$

and

$$n h_n^2 \text{cov}(S_2(h_n), T_2(h_n)) \rightarrow \sigma^2 \left[\int u^2 K \right]^2 \left[\int (m^* w)^2 \right].$$

Proof of (2.1) and (2.3). The proof of (2.1) follows by an argument easier than that used in the proof of Lemma 2. A consequence of (2.1) is

$$\left(\left| \frac{d\hat{m}_n(\hat{h}_n) - d\hat{m}_n(h_n^*)}{d\hat{m}_n(h_n^*)} \right| + \left| \frac{d\hat{w}_n(\hat{h}_n) - d\hat{w}_n(h_n^*)}{d\hat{w}_n(h_n^*)} \right| \right) \rightarrow 0,$$

from which (2.3) follows.

Proof of (2.5). Write

$$CV(h) = p(h) + I + II, \tag{A.15}$$

where

$$I = 2n^{-1} \sum_{i=1}^n [Y_i - \hat{m}(x_i)] [\hat{m}(x_i) - \hat{m}_n(x_i)] w(x_i)$$

and

$$II = n^{-1} \sum_{i=1}^n [\hat{m}(x_i) - \hat{m}_n(x_i)]^2 w(x_i).$$

But it is straightforward to verify that

$$p(h) = \sigma^2 \left[\int W \right] + O_p(n^{-1} h^{-1}),$$

$$I = 2n^{-1} h^{-1} K(0) \sigma^2 \left[\int w \right] + O_p(n^{-1} h^{-1}),$$

and

$$II = O_p(n^{-1} h^{-2})$$

uniformly over $h \in H_n$. Hence (A.15) gives (2.5).

[Received December 1985. Revised May 1987.]

REFERENCES

Akaike, H. (1970). "Statistical Predictor Information," *Annals of the Institute of Statistical Mathematics*, 22, 203-217.

— (1974). "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716-723.

Bowman, A. (1984). "An Alternative Method of Cross-Validation for the Smoothing of Density Estimates," *Biometrika*, 71, 353-360.

Clark, R. M. (1975). "A Calibration Curve for Radio Carbon Dates," *Antiquity*, 49, 251-266.

Craven, P., and Wahba, G. (1979). "Smoothing Noisy Data With Spline Functions," *Numerische Mathematik*, 31, 377-403.

Gasser, T., and Müller, H. G. (1979). "Kernel Estimation of Regression Functions," in *Smoothing Techniques in Curve Estimation (Lecture Notes in Mathematics, 757)*, Heidelberg: Springer-Verlag, pp. 23-68.

Hall, P. (1984). "Central Limit Theorem for Integrated Square Error of Multivariate Nonparametric Density Estimators," *Journal of Multivariate Analysis*, 14, 1-16.

Hall, P., and Marron, J. S. (1987a). "Extent to Which Least-Squares Cross-Validation Minimises Integrated Square Error in Nonparametric Density Estimation," *Theory of Probability and Related Fields*, 74, 567-581.

— (1987b). "On the Amount of Noise Inherent in Bandwidth Selection for a Kernel Density Estimator," *The Annals of Statistics*, 15, 163-181.

Hardie, W., and Marron, J. S. (1985a). "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation," *The Annals of Statistics*, 13, 1465-1481.

— (1985b). "Asymptotic Nonequivalence of Some Bandwidth Selectors in Nonparametric Regression," *Biometrika*, 72, 481-484.

Li, K. C. (1985). "From Stein's Unbiased Risk Estimates to the Method of Generalized Cross-Validation," *The Annals of Statistics*, 13, 1352-1377.

— (1987). "Asymptotic Optimality for C_p , C_c , Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958-975.

Mallows, C. L. (1973). "Some Comments on C_p ," *Technometrics*, 15, 661-675.

Marron, J. S. (1986). "Will the Art of Smoothing Ever Become a Science?" in *Function Estimates (Contemporary Mathematics ser., no. 59)*, Providence, RI: American Mathematical Society, pp. 169-178.

Marron, J. S., and Härdle, W. (1986). "Random Approximations to Some Measures of Accuracy in Nonparametric Curve Estimation," *Journal of Multivariate Analysis*, 20, 91-113.

Nadaraya, E. A. (1964). "On Estimating Regression," *Theory of Probability and Its Application*, 9, 141-142.

Nolan, D., and Pollard, D. (1987). "U-Processes: Rates of Convergence," *The Annals of Statistics*, 15, 780-799.

Priestley, M. B., and Chao, M. T. (1972). "Non-parametric Function Fitting," *Journal of the Royal Statistical Society, Ser. B*, 34, 385-392.

Rice, J. (1984). "Bandwidth Choice for Nonparametric Regression," *The Annals of Statistics*, 12, 1215-1230.

Rosenblatt, M. (1971). "Curve Estimates," *Annals of Mathematical Statistics*, 42, 1815-1842.

Rudemo, M. (1982). "Empirical Choice of Histograms and Kernel Density Estimators," *Scandinavian Journal of Statistics*, 9, 65-78.

Scott, D. W., and Terrell, G. R. (1987). "Biased and Unbiased Cross-Validation in Density Estimation," *Journal of the American Statistical Association*, 82, 1131-1146.

Shibata, R. (1981). "An Optimal Selection of Regression Variables," *Biometrika*, 68, 45-54.

Wahba, G. (1985). "A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem," *The Annals of Statistics*, 13, 1378-1402.

Watson, G. S. (1964). "Smooth Regression Analysis," *Sankhyā, Ser. A*, 26, 359-372.

Whittle, P. (1960). "Bounds for the Moments of Linear and Quadratic Forms in Independent Variables," *Theory of Probability and Its Applications*, 5, 302-305.

DAVID W. SCOTT*

1. INTRODUCTION

The authors have provided a deeper understanding of the theory and practice of data-based methods in nonparametric regression. The clear and forthright presentation is strongly influenced by Rice (1984). Hall and Marron (1987a,b) compiled parallel results for nonparametric density estimators. George Terrell and I have chosen to work primarily on cross-validation in the context of density estimation (Scott and Terrell 1987), which is in some ways harder and in other ways easier than regression. My comment focuses on two questions that apply to both regression and density estimation. The first question is: What are the relative merits of choosing h_0 or \hat{h}_0 as the "optimal bandwidth"? The second question, which refers to the authors' Remark 3.1, is: What role, if any, should "plug-in" methods for choosing h have? These issues were touched on for density estimation in Scott and Terrell (1987). We have also compared cross-validation algorithms for very large samples and presented one such example. We suggest that having two rather different cross-validation algorithms can be very useful in practice.

Designing experiments for nonparametric regression is harder than for density estimation. Only the shape of the density curve f must be specified in the latter situation, whereas regression requires at least three choices: shape of the regression curve m , noise level σ^2 , and the form of the noise distribution. How interactions among these choices affects regression estimates is not completely understood. On the other hand, regression is much easier than density estimation because actual residuals are available with the data.

2. CHOICE OF OPTIMAL BANDWIDTH DEFINITION

Although the error criterion introduced in regression is average squared error (ASE), most theoretical work has examined the mean average squared error (MASE). In Section 2 the authors argue that h_0 (minimizing ASE) should be taken as the definition of optimal bandwidth rather than \hat{h}_0 (minimizing MASE). This appears noncontroversial, but some discussion is appropriate. With some beautiful theorems the authors characterize the asymptotic joint distribution of some data-based choices, \hat{h} , of the smoothing parameter and the optimal ASE choice h_0 .

Many density estimates have been examined using both h_0 and \hat{h}_0 [optimal mean integrated squared error (MISE) and integrated squared error (ISE) smoothing param-

eters], and I presume that experience roughly translates to the regression setting, given the similar asymptotic error expansions. Without question a (regression) estimate using \hat{h}_0 is "better" than an estimate constructed with h_0 (assuming for the moment that both are observable). How much better is \hat{h}_0 , and how is the improvement achieved? In our density-estimation simulations, we found the improvement was only occasionally large and was small relative to the range of observed integrated squared errors. The curious feature was the manner by which the estimate was "improved" as h varied between h_0 and \hat{h}_0 . For samples where the sample standard deviation $\hat{\sigma}$ was less than σ , \hat{h}_0 tended to be larger than h_0 , and vice versa (the correlation between \hat{h}_0 and $\hat{\sigma}$ was approximately $-.65$ for simulations with Gaussian data). Thus samples with too small a variance had improved squared error after "flattening," whereas samples with too large a variance were improved by "roughening." In the regression context, with data measured on an equally spaced grid, we believe a similar phenomenon may be observed based on the level. For example, if the residuals between the data and the curve $m(t)$ [not $\hat{m}(t)$] are positively correlated with the curve $m(t)$ [i.e., the data lie above $m(t)$ near peaks and below near troughs], then increasing the smoothing parameter beyond h_0 will reduce the error. For negatively correlated residuals [data below $m(t)$ at peaks and above at troughs], roughening the curve reduces error.

In practice it is difficult for a cross-validation value \hat{h} to mimic the behavior of h_0 . To do so would require, in part, guessing whether $\hat{\sigma} > \sigma$ in the density case and the signs of residuals in the regression case. We might expect \hat{h} and h_0 to be negatively correlated (with h_0 in between), which is indeed the case in both regression and density estimation (see Table 6 and Remark 3.2). On the other hand, the correlation is not very large and usually h_0 and \hat{h}_0 are not too far apart.

To summarize, we do not prefer MASE to ASE philosophically, but we believe it will not be possible in practice to find cross-validation algorithms that actually estimate \hat{h}_0 rather than h_0 . It might be desirable and feasible to design an algorithm so that \hat{h} tends toward h_0 .

3. THE ROLE FOR "PLUG-IN" ESTIMATES

For a nonnegative symmetric kernel density estimator

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right),$$

it is well known that the MISE is approximately

$$\text{MISE}(h) = R(K)/nh + \frac{1}{4}h^4\mu_2^2 R(f''), \quad (1)$$

* David W. Scott is Professor, Department of Statistics, Rice University, Houston, TX 77251. The research for this article was performed while the author was on sabbatical at Stanford University's Department of Statistics. It was supported by Office of Naval Research Contract N00014-85-K-0100 and Army Research Office Contract DAAG-29-85-K-0212. The author thanks Luc Devroye and Brad Efron for helpful discussions.

where $\mu_2 = \int x^2 K(x) dx$ and $R(\phi) = \int \phi(x)^2 dx$ [see $d_{CV}^2(h)$ before eq. (2.1)]. A "plug-in" (or biased cross-validation) procedure attempts to estimate the unknown $R(f^*)$ in Equation (1). Using $R(\hat{f}^*)$, which varies with h , is not satisfactory, since \hat{f}^* is inconsistent. On the other hand, it may be shown (Scott and Terrell 1987) that

$$E[R(\hat{f}^*)] = R(f^*) + R(K^*)/nh^5 + O(h^7).$$

Since $h^* = O(n^{-1/5})$, $R(\hat{f}^*)$ asymptotically has a fixed bias. Substituting $R(\hat{f}^*) - R(K^*)/nh^5$ for $R(f^*)$ in (1) gives a consistent but slightly biased (but asymptotically unbiased) estimate of the MISE as a function of h , which can be numerically minimized. I have proven that asymptotically the distance between this "biased" cross-validation choice for $h(\hat{h}_{BCV})$ and h_0 is much less than the distance between \hat{h} and h_0 [where \hat{h} or \hat{h}_{UCV} is the corresponding "unbiased" or least-squares cross-validation choice from Rudemo (1982) and Bowman (1984)]. Since \hat{h} and \hat{h}_{UCV} are negatively correlated, it is not unusual for \hat{h}_{BCV} to be closer to h_0 than \hat{h}_{UCV} is. I present an example using these two criteria in Section 4.

The authors express doubt that a plug-in estimate would work well in regression, since there are two unknowns, σ^2 and $R(m^*)$. Assume the authors' simplest case (at the beginning of their sec. 2), where $m(t)$ is periodic on the unit interval and the weight function $w(x) = 1$. Then it may be shown that

$$E[R(\hat{m}^*)] = R(m^*) + R(K^*)\sigma^2/nh^3 + o(1),$$

so a plug-in estimate becomes

$$M\hat{A}SE(h) = \frac{R(K) - \mu_2^2 R(K^*)/4}{nh} \sigma^2 + \frac{1}{4} h^4 \mu_2^2 R(\hat{m}^*),$$

which may be computed if σ^2 is known. Otherwise, $\hat{\sigma}^2(h)$ may be estimated by computing $\hat{m}(h)$ and adjusting the residuals in the usual manner (Belsley, Kuh, and Welsch 1980).

The plug-in regression procedure is more complicated here than in the density-estimation case. Whether it pro-

vides as good an estimate of h_0 as in the density-estimation case is unknown. Plug-in procedures do not have the desirable automatic adaptivity enjoyed by the unbiased cross-validation procedures.

4. AN EXAMPLE

Experimentation with large data sets indicates the value of having two distinct cross-validation procedures. As an example, consider the steel-surface data of Bowyer (1980) analyzed by Silverman (1986). These data were provided by Bernard Silverman. The 15,000 points are given as 500 bin counts over the interval (0, 50), so the bin width $\delta = .1$. The density estimator used is the averaged shifted histogram (see Scott 1985) with the triweight kernel

$$K(t) = (35/32)(1 - t^2)^3 I_{[-1,1]}(t).$$

The unbiased and biased cross-validation functions are plotted in my Figure 1 for $h = m\delta$ ($1 \leq m \leq 35$) and are minimized for $h = 1.3$ and $h = 1.2$, respectively. Notice that the level of the curve in Figure 1a is shifted by a fixed amount; otherwise, the two graphs have comparable scales. The data-based density estimate is shown in Figure 2b. Observe that the unbiased cross-validation function is relatively indifferent among h 's in the range [.3, 2.0], a fairly rare event (see Figs. 2a and 2c). Perhaps unexpectedly, a clear preference is not always made with such large data sets, but I believe that many interesting and practical questions remain about applying cross-validation procedures. S functions to perform these calculations are available from me (scotttdw@rice.edu).

5. CONCLUDING REMARKS

Given the relatively slow convergence of cross-validation methods, it is very helpful to compare several different cross-validation algorithms. When these agree we are satisfied. When they do not, we investigate further. The methods preferred by the authors are all highly correlated. Development of a complementary biased cross-validation procedure in regression may be as useful as in density

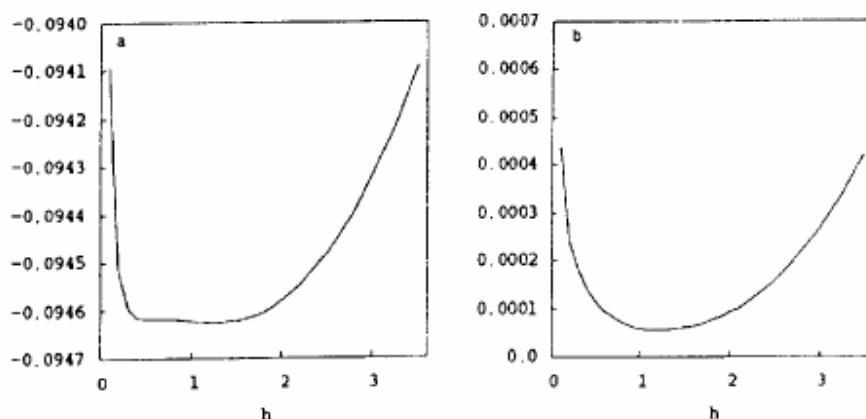


Figure 1. Cross-Validation Functions: (a) Unbiased. (b) Biased.

Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands

WOLFGANG HÄRDLE and ADRIAN W. BOWMAN*

The operation of the bootstrap in the context of nonparametric regression is considered. Bootstrap samples are taken from estimated residuals to study the distribution of a suitably recentered kernel estimator. The application of this principle to the problem of local adaptive choice of bandwidth and to the construction of confidence bands is investigated and compared with a direct method based on asymptotic means and variances. The technique of the bootstrap is to replace any occurrence of the unknown distribution in the definition of the statistical function of interest by the empirical distribution function of the observed errors. In a regression context these errors are not directly observed, although their role can be played by the residuals from the fitted model. In this article the fitted model is a kernel nonparametric regression estimator. Since nonparametric smoothing is involved, an additional difficulty is created by the bias incurred in smoothing. This bias, however, can be estimated in a consistent fashion. These considerations suggest the way in which the distribution of the nonparametric estimate about the true curve at some point of interest may be approximated by suitable recentering of the nonparametric estimates based on bootstrap samples. The bootstrap samples are constructed by adding to the observed estimate errors, which are randomly chosen without replacement from the collection of recentered and bias-corrected residuals from the original data. A theorem is proved to establish that the bootstrap distribution approximates the distribution of interest in terms of the Mallows metric. Two applications are considered. The first uses bootstrap sampling to approximate the mean squared error of the nonparametric estimate at some point of interest. This can then be minimized over the smoothing parameter to adapt the degree of smoothing applied at any point to the local behavior of the underlying curve. The second application uses the percentiles of the approximate distribution to construct confidence intervals for the curve at specific design points. In both of these cases the performance of the bootstrap is compared with a simple "plug-in" method based on direct estimation of the terms in an asymptotic expansion. The performances of the two methods are in general very similar. The bootstrap, however, has the slight advantage of not being as sensitive as the direct method to second derivatives near 0 in the local adaptive smoothing problem. In addition, in the construction of confidence intervals the bootstrap is able to reflect features such as skewness but falls slightly short of target confidence intervals as a result of inaccuracies in centering when the second derivative of the curve is high.

KEY WORDS: Regression smoothing; Resampling techniques.

1. INTRODUCTION

The bootstrap is a resampling technique whose aim is to gain information on the distribution of an estimator. In nonparametric regression there are several ways in which such information could be of considerable assistance. One application could be in choosing the parameter that controls the degree of smoothing that is applied to the data. Another area of interest is the construction of confidence intervals for the curve. Discussion of the first problem is usually directed toward methods that asymptotically minimize a global criterion such as mean integrated squared error. When estimating the regression function at a particular point, however, it would be helpful to tailor our choice of smoothing parameter to the features exhibited near that point. For example, near a peak a relatively small value of smoothing parameter is appropriate, whereas on an approximately linear section a larger value should be used.

The construction of confidence intervals extends the use of nonparametric smoothing beyond its role as a point estimator, often constructed with the sole purpose of giving visual information on the shape of the underlying regression curve. It would be very helpful to obtain, through confidence intervals, an impression of the variability of

the estimator, providing a useful scale against which unusual features of the estimated curve may be assessed.

This article investigates the use of the bootstrap in providing approximations to a suitably centered distribution of kernel estimators of nonparametric regression curves. From the bootstrap distribution an estimate of local mean squared error is available, enabling a good choice of a local smoothing parameter to be made. Confidence bands for the true curve can also be derived from the bootstrap distribution. Both of these problems could be tackled in a more direct way by estimating the asymptotic means and variances of the estimators. Such an approach is simpler and can be very effective. One advantage of the bootstrap is that it does reflect the presence of nonstandard features such as skewness, although in the simulations of Section 3 the bootstrap proved to be slightly less effective than the direct method in attaining the target coverage probability of confidence bands.

The structure of the data is assumed to be of the following form.

Condition 1. $Y_i = m(x_i) + \epsilon_i$ ($i = 1, \dots, n$), where $E(\epsilon_i) = 0$ and the design points x_i are equally spaced. For simplicity we assume that $x_i = (i - 1)/n$, where n is the total number of observations. Extensions to other patterns of design points are possible. m is a twice continuously differentiable function, and the errors ϵ_i are independent, with distribution F and constant variance σ^2 .

* Wolfgang Härdle is Principal Researcher, Department of Economics, University of Bonn, D-5300 Bonn 1, West Germany. Adrian W. Bowman is Lecturer, Statistics Department, University Gardens, The University, Glasgow G12 8QW, Scotland. The authors are indebted to Steve Portnoy, Dennis Cox, and Mike Titterton for stimulating discussions and to an associate editor and referee for helpful suggestions. This research was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 123 and Sonderforschungsbereich 303.

We adopt the estimator of m originally proposed by Priestley and Chao (1972), namely

$$\hat{m}(x) = \hat{m}(x; h) = n^{-1}h^{-1} \sum_{i=1}^n K((x - x_i)/h)y_i,$$

and make the following assumptions on K .

Condition 2. The kernel function K is a symmetric probability density with bounded support that is Lipschitz continuous and has been parameterized so that $\int u^2 K(u) du = 1$.

Under Conditions 1 and 2, we have for any x in a subinterval $[\eta, 1 - \eta]$ ($\eta > 0$),

$$E_r \hat{m}(x) = m(x) + \frac{h^2}{2} m''(x) + o(h^2)$$

$$\text{var}_r(\hat{X}) = n^{-1}h^{-1} \sigma^2 \int K^2(u) du + o(n^{-1}h^{-1}) \quad (1)$$

as $n \rightarrow \infty, h \downarrow 0$.

These asymptotic expressions indicate that an appropriate choice of the smoothing parameter h for estimation of $m(x)$ should be influenced by the local curvature of m , as expressed in the second derivative $m''(x)$. When $|m''(x)|$ is large, small values of h are required to keep the bias low, whereas when $|m''(x)|$ is small, large values of h are appropriate to deflate the variance. Local adaptive smoothing aims to balance these effects in a way that is appropriate for each particular location.

Section 2 discusses the general application of the bootstrap in the context of nonparametric regression. It is shown that the bootstrap works when an appropriate correction term is introduced. Section 3 discusses local adaptive smoothing, and Section 4 deals with confidence bands; Sections 3 and 4 give numerical examples and describe a small simulation study. Some brief discussion is given in Section 5.

2. THE BOOTSTRAP IN NONPARAMETRIC REGRESSION

The technique of the bootstrap is to replace any occurrence of the unknown distribution F in the definition of the statistical function of interest by the empirical distribution function F_n of $\{e_i\}$. Since we cannot observe F_n , we need an initial estimate \hat{m} of the regression function from which to estimate residuals $\hat{e}_i = Y_i - \hat{m}(x_i)$. Special attention, however, must be paid to observations near the boundary of the interval $[0, 1]$. Since \hat{m} has a slower rate of convergence near the boundary (Gasser and Müller 1979), it is advisable to use residuals only from an interior subinterval $[\eta, 1 - \eta]$ ($0 < \eta < \frac{1}{2}$), which contains the point x . The residuals need not necessarily have mean 0, so, to let the resampled residuals reflect the behavior of the true observation errors, they should first be recentered as

$$\tilde{e}_i = \hat{e}_i - \frac{1}{[(1 - 2\eta)n]} \sum_i \hat{e}_i,$$

where, to exclude boundary effects, $\eta n + 1 \leq i \leq (1 - \eta)n - 1$.

Bootstrap residuals e_i^* are then created by sampling with replacement from $\{\tilde{e}_i\}$, giving bootstrap observations $y_i^* = \hat{m}(x_i) + e_i^*$. A bootstrap estimator m^* of m is then obtained by smoothing $\{Y_i^*\}$ rather than $\{Y_i\}$.

We define the bootstrap principle to hold if the distributions of $m^*(x)$ and $\hat{m}(x)$, when suitably normalized, become close as the sample size n increases. Specifically, we shall examine convergence of these distributions in the Mallows metric, following Bickel and Freedman (1981).

Since the variance of $\hat{m}(x)$ converges to 0 at the rate $n^{-1}h^{-1}$, as shown previously, we consider $\sqrt{nh}(\hat{m}(x) - m(x))$. It is important, however, to note that $\hat{m}(x)$ is a biased estimator of $m(x)$ and that if h is chosen to balance this bias against the standard deviation of \hat{m} then the variance and squared bias will have the same speed of convergence to 0. It is necessary, therefore, to ensure that this behavior is mirrored in the distribution of the bootstrap estimator $m^*(x)$.

The following approximate decomposition into a variance and bias part is helpful in understanding bootstrapping in this context:

$$\hat{m}(x) - m(x) \approx n^{-1}h^{-1} \sum_{i=1}^n K((x - x_i)/h)e_i + (h^2/2)m''(x). \quad (2)$$

In the bootstrap, any occurrence of e_i is replaced by e_i^* and we have

$$m^*(x; h, g) = n^{-1}h^{-1} \sum_{i=1}^n K((x - x_i)/h)(\hat{m}(x_i; g) + e_i^*),$$

where the pilot bandwidth g is used to produce residuals $\hat{e}_i = Y_i - \hat{m}(x_i; g)$. Note that in the definition of m^* there are two levels of smoothing involved. It is clearly helpful to have a good initial estimate $\hat{m}(x_i; g)$, giving reasonable residuals. Cross-validatory choice of g is a strong candidate, since Rice (1984) and Härdle and Marron (1985) showed that this produces estimators that asymptotically minimize the mean integrated squared error. In this article cross-validation will be used to choose the pilot bandwidth g .

The distribution of the bootstrap estimator is centered around its expectations (under the bootstrap distribution). This expectation is

$$n^{-2} \sum_{i=1}^n \sum_{j=1}^n h^{-1}K\left[\frac{x - x_i}{h}\right] g^{-1}K\left[\frac{x_i - x_j}{g}\right] Y_j.$$

Using Conditions 1 and 2, it can be shown that the convolution term is approximated by the integral

$$K_1(v; h, g) = \int h^{-1}K(u/h)g^{-1}K\left[\frac{u - v}{g}\right] du.$$

Therefore, center m^* around

$$\hat{m}_e(x; h, g) = n^{-1} \sum_i K_1((x - x_i); h, g)Y_i,$$

for reasons of computational efficiency. The kernel K_1 corresponds to the density of the sum of the two inde-

pendent random variables hZ_1, gZ_2 , where Z_1 and Z_2 have density K . K_1 can, therefore, be computed analytically.

The bias component in (2) may be estimated by employing a consistent estimator of $\hat{m}^*(x)$. For example, a consistent kernel estimator is

$$\hat{m}^*(x) = n^{-1}l^{-3} \sum_{i=1}^n K_{(2)} \left[\frac{x - x_i}{l} \right] Y_i,$$

where $K_{(2)}$ satisfies

$$\int K_{(2)}(u) du = \int u K_{(2)}(u) du = 0, \\ \int u^2 K_{(2)}(u) du = 2$$

and $l \rightarrow 0$ and $nl^5 \rightarrow \infty$. To study the distribution of $(nh)^{-1/2}(\hat{m}(x) - m(x))$, we will, therefore, use the following bootstrap approximation:

$$\sqrt{nh} \left(m^*(x; h, g) - \hat{m}_c(x; h, g) + \frac{h^2}{2} \hat{m}^*(x) \right).$$

We make the following assumption on the smoothing parameters.

Condition 3. $\{h\}$ and $\{g\}$ are sequences of smoothing parameters that tend to 0 at the rate $n^{-1/5}$.

This is the rate entailed by choosing h to balance integrated squared bias against variance [see (1)] and was shown by Stone (1980) to be the optimal rate under our conditions on the regression function m .

At first sight the two levels of smoothing have some obvious similarities to twicing. Stuetzle and Mittal (1979), however, derived some asymptotic theory for twiced kernel estimators and showed that twicing is equivalent to using $2\hat{m} - \hat{m}_c$ as an estimator for m . Twicing is, therefore, different from bootstrapping.

The Mallows metric $d_2(F, G)$ between the distributions F and G is defined to be the infimum of $E\{(X - Y)^2\}^{1/2}$ over pairs of random variables X and Y having marginal distributions F and G , respectively. We shall adopt the convention that where random variables appear in the arguments of d_2 they represent the corresponding distributions.

Theorem 1. Under Conditions 1-3, the bootstrap principle holds in the following form:

$$d_2 \left(\sqrt{nh} \{ \hat{m}(x; h) - m(x) \}, \sqrt{nh} \left\{ m^*(x; h, g) - \hat{m}_c(x; h, g) + \frac{h^2}{2} \hat{m}^*(x) \right\} \right) \rightarrow 0,$$

as $n \rightarrow \infty$. Proof of this theorem is given in the Appendix.

This theorem shows that the bootstrap principle holds when resampling is carried out from the residuals $y_i - \hat{m}(x_i; g)$. Since an estimate of bias is already employed in the recentring of the distribution, we may also bias correct the residuals, so that resampling takes place from $(y_i - \hat{m}(x_i; g) + \frac{1}{2}g^2\hat{m}^*(x_i))$. It is easy to see that the theory

of Theorem 1 follows through without difficulty in this case; the bias component of the Mallows metric is the only part that is affected. This gives the following corollary.

Corollary 1. Theorem 1 holds when resampling is carried out from bias-corrected residuals.

The advantage of this is that the bootstrap distributions reflects the true error distribution more faithfully. Section 4 gives an example where the bootstrap is able to respond to skewness in the error distribution. Such a feature is less easily identified when the residuals are not corrected for bias. For the remainder of the article, bias correction of residuals will be assumed.

The mean squared error $MSE(x; h) = E_r\{\hat{m}(x, h) - m(x)\}^2$ may be estimated from the bootstrap method by

$$M\hat{S}E(x; h) = \int \left(m^*(x; h, g) - m_c(x; h, g) + \frac{h^2}{2} \hat{m}^*(x; g) \right)^2 d\hat{F}_n,$$

where \hat{F}_n denotes the empirical distribution function of $\{\hat{e}_i\}$. Denote by \hat{h} the bandwidth that minimizes $M\hat{S}E(x; h)$ over a range of smoothing parameters $H_n \subset (an^{-1/5}, bn^{-1/5})$ ($0 < a < b$), with cardinality $\#H_n$. The following theorem can be proved by using methods similar to those of Rice (1984) and Härdle and Marron (1985).

Theorem 2. If the conditions of Theorem 1 are in force and if, for some $D > 0$, $\#H_n n^{-1/5} \leq D$, then \hat{h} is asymptotically optimal in the sense that

$$\frac{MSE(x; \hat{h})}{\inf_{h \in H_n} MSE(x; h)} \rightarrow 1,$$

as $n \rightarrow \infty$.

3. LOCAL ADAPTIVE SMOOTHING

The bootstrap principle allows the estimation of mean squared error at specific estimation points x . To adapt the smoothing to local features this estimated mean squared error can be minimized over a range of smoothing parameters. A more direct estimator of mean squared error is obtained by plugging in estimates for the unknown quantities in expressions (1), namely

$$\frac{h^4}{4} \{\hat{m}^*(x)\}^2 + n^{-1}h^{-1} \int K^2(u) du \hat{\sigma}^2,$$

which in turn provides what we will term a "plug-in" estimate of the optimal local smoothing parameter. An estimate of σ^2 is provided by

$$\hat{\sigma}^2 = \{(1 - 2\eta)n\}^{-1} \sum_i \left\{ Y_i - \hat{m}(x_i) + \frac{h^2}{2} \hat{m}^*(x_i) \right\}^2.$$

This shares with the bootstrap the need for an estimate of bias. The bootstrap attempts to remove some of the dependence on such asymptotic formulas by simulating from the data to provide an estimate of the variance part of the mean squared error. Alternative estimates $\hat{\sigma}^2$ were dis-

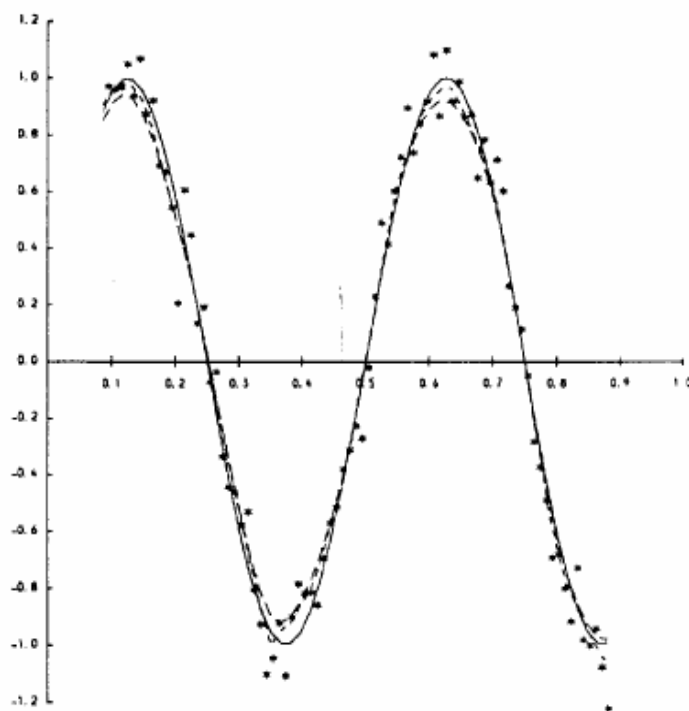


Figure 1. Data Simulated From the Curve $m(x) = \sin(4\pi x)$. With $N(0, (.1)^2)$ Error Distribution. The curve is shown by a solid line, global smoothing is shown by a dashed line, and local adaptive smoothing is shown by a fine dashed line.

cussed by Rice (1984) and Gasser, Sroka, and Jennen-Steinmetz (1986).

Figure 1 displays some data simulated by adding normally distributed errors, with standard deviation .1, to the curve $m(x) = \sin(4\pi x)$ evaluated at $x = (i - \frac{1}{2})/100$ ($i = 1, \dots, 100$). To avoid problems with edge effects, the curves and data have been plotted only over an interior region of $(0, 1)$. Cross-validation was used to select a good global smoothing parameter ($g = .03$; sum of squares based on an interior region to avoid edge effects). The resulting estimate of the regression function shows the problems caused by bias at the peaks and troughs, where $|m''(x)|$ is high.

Estimation of derivatives and appropriate smoothing parameters was discussed by Gasser and Müller (1984), who showed that a larger smoothing parameter will be required to obtain a good estimator of $m''(x)$. The asymptotic formulas given by these authors suggest that, for sample size 100, a simple but reasonable smoothing parameter for estimation of $m''(x)$ is obtained by approximately doubling the cross-validated one, g . To use a level of smoothing that deviates greatly from this rule of thumb would require the assumption that $m''(x)$ is extremely smooth or extremely rough, since higher-order derivatives enter the asymptotic formulas with only a very small power. Here we will use $2g$. [Notice that estimation of $m''(x)$ requires

the smoothing parameter to converge to 0 at a slower rate than in estimation of $m(x)$. The proposal to use $2g$ is tailored to sample size 100 and, in general, an approximate formula such as $1.5g n^{1/10}$ might be used.]

Figure 2 plots the local smoothing parameters obtained by minimizing the bootstrap estimate of mean squared error over a grid of smoothing parameters near g . For comparison, the asymptotically optimal local smoothing parameters are also plotted, and it can be seen that an appropriate pattern of local smoothing has been achieved. Again, to avoid edge effects, residuals near 0 and near 1 were not included in the bootstrap sampling and the curve was evaluated over the corresponding interior region. Comparison with the "plug-in" local smoothing parameters also reveals very little difference. Since the bootstrap is estimating only the variance of $\hat{m}(x)$, its performance is not markedly superior to the direct method. The estimate of the regression curve produced by the local parameters is also displayed in Figure 1, where it can be seen that this estimate is considerably nearer the true curve at most of the peaks and troughs. Since the two estimates based on local smoothing are virtually indistinguishable, only the bootstrap one has been plotted. A normal kernel with truncated support was used in this example because it has the helpful property that the convolution kernel $K_1(u; h, g)$ is well approximated by a normal density, with

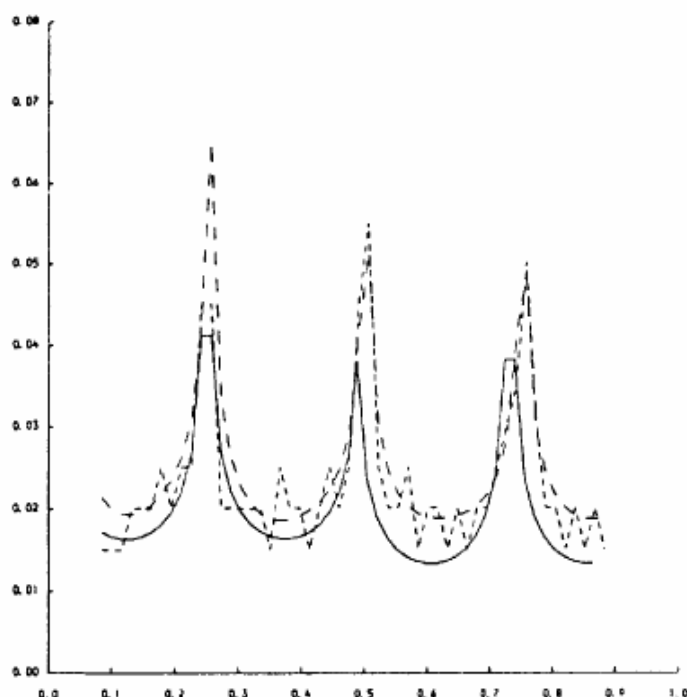


Figure 2. Local Smoothing Parameters for the Simulated Data of Figure 1. Asymptotically optimal is shown by a solid line, direct estimation is shown by a dashed line, and bootstrap is shown by a fine dashed line.

variance $h^2 + g^2$. In addition, the kernel $K_{(2)}$ used in the estimation of m'' can be taken to be the second derivative of the normal kernel.

To quantify the comparisons, 10 simulations were carried out and the squared errors of each estimated curve were averaged over the simulations and over the design points. The results for global smoothing, local smoothing by bootstrapping, and local smoothing by direct estimation were .000997, .000582, and .000581, respectively. This confirms the improved performance of local adaptive smoothing over global smoothing, and the similar results of the bootstrap and direct methods.

The asymptotically optimal local smoothing parameter contains the factor $m''(x)^{-2}$ and so takes very large values when m'' is close to 0. The curve on Figure 2 is truncated because the grid over which the parameters have been calculated does not contain the points .25, .5, or .75, where m'' is exactly 0. One disadvantage of the "plug-in" method of local smoothing, compared with the bootstrap, is that the factor $m''(x)^{-2}$ is present, causing oversensitivity at locations where m'' is near 0, as Figure 2 shows.

4. CONFIDENCE BANDS

In addition to estimation of mean squared error, the bootstrap principle allows the construction of pointwise confidence bands for the true regression curve, since Theo-

rem 1 shows that the bootstrap distribution approximates the distribution of $(nh)^{1/2}(m\hat{h} - m)$. A confidence interval for the curve at a specific point x may be obtained by bootstrap sampling and calculation of the empirical quantiles. This contrasts with a direct approach based on asymptotic normality with estimated bias and variance.

In this section, global smoothing based on the cross-validated bandwidth is used, both for the original data and for the bootstrap samples. The use of a global bandwidth allows results to be pooled across different points on the curve. The potential advantages of local adaptive smoothing in the context of confidence intervals are not clear, since a simple bias correction can be added to the globally smoothed curve, as suggested by the material of Section 2.

Figures 3 and 4 display the bias-corrected estimate with nominal 95% pointwise confidence intervals at 32 estimation points for the example described in Section 3. For clarity, the figures are drawn on recentered scales by subtracting the true regression curve. Figure 3 shows that when the error distribution is normal there is very little difference between the intervals produced by the bootstrap and the direct method, as would be expected from the discussion of mean squared error estimation in Section 2. The empirical confidence levels (coverage relative frequencies, averaged over design points) of the two methods

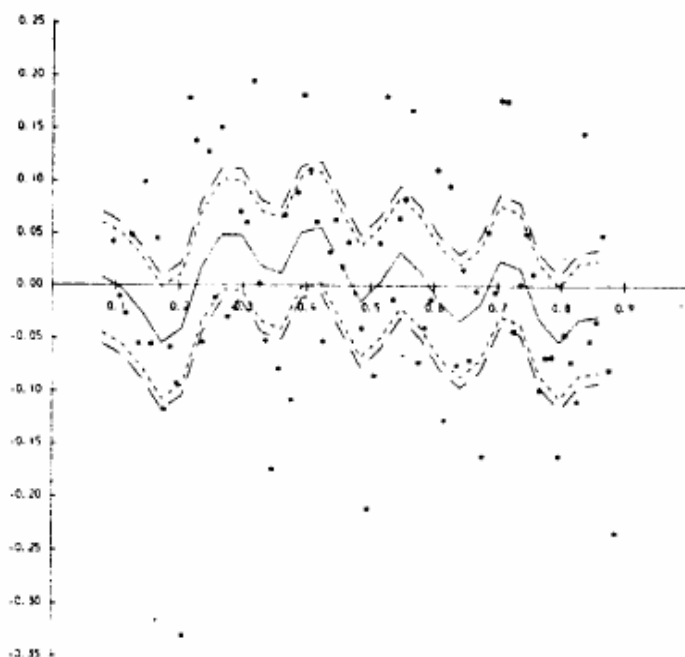


Figure 3. Pointwise Confidence Bounds, Recentered by Subtraction of the True Curve $m(x) = \sin(4xx)$, With Normal Errors, Scaled in Each Case to Have Mean 0 and Standard Deviation .1. Point estimate is shown by a full line, directly estimated bands are shown by a dashed line, and bootstrap estimation is shown by a fine dashed line.

are 87% for the bootstrap and 92% for the direct method. This is based on 100 simulations of original data followed in each case by 100 bootstrap simulations.

Since the direct method is based only on an estimate of error variance, we may expect the bootstrap to perform better in the presence of skewness. This is investigated in Figure 4, where the error distribution in the simulations is exponential, shifted to have mean 0 and scaled to have standard deviation .1. Here the bootstrap reflects the variation of the estimate about its mean more satisfactorily than the direct method and the asymmetry is clearly apparent. (Notice that positive skewness of the bootstrap distribution about its mean leads to negative skewness of the confidence interval.) The empirical confidence levels, however, remain at 86% for the bootstrap and 92% for the direct method.

The slightly low empirical confidence levels are due to the imperfect estimation of bias, as can be seen by the fact that the target confidence levels are very nearly achieved when the true second derivative is employed in the bias-correction term. The differences between the bootstrap and the direct methods can be explained by the fact that the bootstrap correctly reflects the variation of the estimate about its mean, and so the confidence intervals have a smaller width than the direct method. In both cases, however, the intervals are slightly incorrectly centered because of the bias estimation. In the direct method this also

leads to a slightly inflated estimate of variance, which counteracts the incorrect centering and achieves a confidence level close to the target one.

A potential advantage of the bootstrap is that it can be applied to the construction of uniform confidence bands. Bootstrap sampling can be used to approximate the distribution of $\sup | \hat{m}(x) - m(x) |$, which in general is not amenable to theoretical treatment without the further assumption of normality, as in Knafik, Sacks, and Ylvisaker (1985) and Hall and Titterington (1986). This could be done in practice by examining the estimated regression curve over a very fine grid. With the methods described previously, however, the achieved confidence levels are unacceptably low as a result of the slight incorrect centering already discussed. The simultaneous bands are particularly sensitive to this effect because it needs only one point of the true curve to lie outside the confidence bands for coverage to fail. Practical use of this approach for simultaneous bands awaits a more satisfactory estimate of second derivatives.

5. DISCUSSION

The theory of Section 2 shows that the bootstrap is successful, in an asymptotic sense, in estimating features of the distribution of a nonparametric regression estimator. The numerical results of Sections 3 and 4 show that bootstrapping does not always perform better than a simple

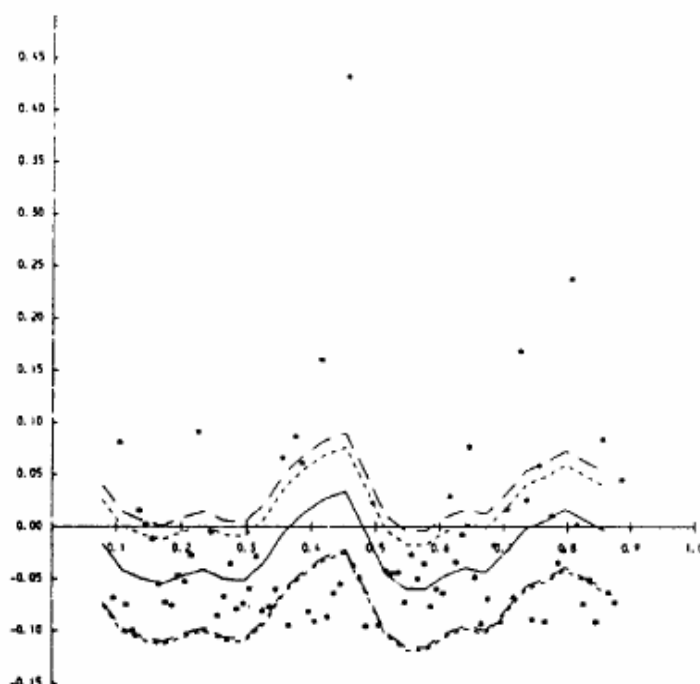


Figure 4. Pointwise Confidence Bounds, Recentered by Subtraction of the True Curve $m(x) = \sin(4\pi x)$. With Exponential Errors, Scaled in Each Case to Have Mean 0 and Standard Deviation .1. Point estimate is shown by a full line, directly estimated bands are shown by a dashed line, and bootstrap estimation is shown by a fine dashed line.

plug-in estimator. The root problem is estimation of bias and, in particular, estimation of the second derivative of the regression curve. The example used in the article is a particularly exacting one since the second derivative is large in several places along the curve. The methods described in the article will work more successfully for smoother curves.

In summary, the performances of the bootstrap and direct methods in the two problems discussed in the article may be compared by first observing that the direct method is simpler but based on asymptotic normality, whereas the bootstrap requires more computational effort in an attempt to reflect nonnormal features of the underlying distribution. In the local adaptive smoothing problem the bootstrap has a slight advantage of not being as sensitive to second derivatives near 0. In the construction of confidence intervals the bootstrap is able to reflect features such as skewness but falls slightly short of target confidence levels as a result of inaccuracies in centering when the second derivative of the curve is high.

APPENDIX: PROOF OF THEOREM 1

An application of Bickel and Freedman's (1981) lemma 8.8, using the modulus norm on the real line, allows the Mallows metric to be split up into a variance part

$$V_n = d_2(\sqrt{nh}\{\hat{m}(x; h) - E_r\hat{m}(x; h)\}, \sqrt{nh}\{m^*(x; h, g) - E^*m^*(x; h, g)\}).$$

where E^* denotes expectation with respect to the bootstrap sampling, and a squared bias part

$$nh\{b_n(x) - b_n^*(x)\}^2,$$

where

$$b_n(x) = E_r\hat{m}(x; h) - m(x)$$

and

$$b_n^*(x) = E^*m^*(x; h, g) - \hat{m}_n(x; h, g) + \frac{h^2}{2} \hat{m}''(x).$$

The variance part may be handled by fairly straightforward application of some of the results of Bickel and Freedman. Their lemma 8.9 shows that

$$\sup_{x \in \mathcal{H}_n} d_2(\sqrt{nh}\{\hat{m}(x; h) - E_r\hat{m}(x; h)\}, \sqrt{nh}\{m^*(x; h, g) - E^*m^*(x; h, g)\})$$

is bounded above by

$$\sup_{x \in \mathcal{H}_n} nh \sum_{y \in \mathcal{H}_n} \alpha_n(x; h)^2 \cdot d_2(y - m(x), y^* - \hat{m}(x; g))^2. \quad (A.1)$$

where $\alpha_n(x; h)$ denotes $n^{-1/2}h^{-1}K((x - x_i)/h)$.

Let F denote the distribution function of the errors ϵ_i , let F_n denote the empirical distribution function (edf) of $\{\epsilon_i\}$ where i is such that $x_i \approx x$ or $x_i \approx 1 - x$. Let \tilde{F}_n denote the edf of the uncentered residuals $\{\tilde{\epsilon}_i\}$, and let \hat{F}_n denote the edf of the centered residuals $\{\hat{\epsilon}_i\}$. The bound (A.1) may be denoted by

$$\sup_{x \in \mathcal{H}_n} nh \sum_{y \in \mathcal{H}_n} \alpha_n(x; h)^2 \cdot d(F, \hat{F}_n)^2.$$

if we adopt the convention on d_1 made previously. It suffices to show that $d_1(F_n, F_n)$ converges in probability to 0. Since d_1 is a metric, we have that

$$d_1(F_n, \hat{F}_n)^2 \leq 2(d_1(F_n, F_n))^2 + d_1(F_n, \hat{F}_n)^2.$$

$d_1(F_n, F_n) \rightarrow 0$ by lemma 8.4 of Bickel and Freedman.

The general result for random variables U and V ,

$$d_1(U, V)^2 = d_1(U, V - E(V))^2 + E^2(U - V) - E^2(U),$$

can be proved by a slight amendment of the proof of Bickel and Freedman's lemma 8.8. An application of this to $d_1(F_n, \hat{F}_n)^2$, with $U = F_n$ and $V = \hat{F}_n$, yields

$$d_1(F_n, \hat{F}_n)^2 = d_1(F_n, \hat{F}_n)^2 - \left\{ ((1 - 2\eta)n)^{-1} \sum_i (\epsilon_i - \xi_i) \right\}^2 + \left\{ ((1 - 2\eta)n)^{-1} \sum_i \epsilon_i \right\}^2$$

and hence

$$E_1 d_1(F_n, \hat{F}_n)^2 \leq E_1 d_1(F_n, \hat{F}_n)^2 + \frac{\sigma^2}{(n - 2\eta n)}.$$

By definition of the Mallows metric $d_1(F_n, \hat{F}_n)^2$, we may consider the joint distribution of $\{\epsilon_i\}$ and $\{\xi_i\}$, which puts probability $(n - 2\eta n)^{-1}$ at each $\{\epsilon_i, \xi_i\}$ to establish that

$$E_1 d_1(F_n, \hat{F}_n)^2 \leq E_1 \left\{ (n - 2\eta n)^{-1} \sum_i (\epsilon_i - \xi_i)^2 \right\} = (n - 2\eta n)^{-1} \sum_i \text{MSE}(x_i),$$

where $\text{MSE}(x_i)$ denotes the mean squared error of $\hat{m}(x_i; g)$.

The convergence result is now established by combining these inequalities and observing that $\sum_i \alpha_i(x; h)^2$ converges uniformly over $h \in H_n$ to 0 with speed $n^{-1}h^{-1}$ (Priestley and Chao 1972).

To deal with the bias part, denote by $\alpha_i(x; h)$ the weight $n^{-1}h^{-1}K((x - x_i)/h)$ and by $\beta_i(x; h, g)$ the weight $n^{-1}K((x - x_i; h, g))$. Then

$$\hat{m}(x; h) = \sum_i \alpha_i(x; h) Y_i$$

and

$$b_n(x) = \sum_i \alpha_i(x; h) m(x_i) - m(x).$$

The bootstrap bias is

$$\begin{aligned} b_n^*(x) &= E^* \left\{ \sum_i \alpha_i(x; h) Y_i^* \right\} - \hat{m}(x; h, g) + \frac{h^2}{2} \hat{m}''(x) \\ &= \sum_i \alpha_i(x; h) \hat{m}(x_i; g) - \hat{m}(x; h, g) + \frac{h^2}{2} \hat{m}''(x) \\ &= \sum_i \alpha_i(x; h) \sum_j \alpha_j(x_i; g) Y_j - \sum_j \beta_j(x; h, g) Y_j \\ &\quad + \frac{h^2}{2} \hat{m}''(x). \end{aligned}$$

By writing $Y_j = m(x_j) + \epsilon_j$ and combining the bias components, we have

$$\begin{aligned} b_n(x) - b_n^*(x) &= \sum_i \alpha_i(x; h) m(x_i) - m(x) \\ &\quad - \sum_i \alpha_i(x; h) \sum_j \alpha_j(x_i; g) m(x_j) \\ &\quad + \sum_j \beta_j(x; h, g) m(x_j) - \frac{h^2}{2} \hat{m}''(x) \\ &\quad - \sum_i \alpha_i(x; h) \sum_j \alpha_j(x_i; g) \epsilon_j + \sum_j \beta_j(x; h, g) \epsilon_j. \end{aligned}$$

Consider first the ϵ_j terms. These may be gathered together as

$$T_{\epsilon} = - \sum_i \left\{ \sum_j \alpha_i(x; h) \alpha_j(x_i; g) - \beta_i(x; h, g) \right\} \epsilon_i,$$

which has mean 0 and variance

$$n^{-1} \sum_i \left\{ n^{-1} \sum_j h^{-1} K((x - x_i)/h) g^{-1} K((x_i - x_i)/g) - K_i((x - x_i); h, g) \right\}^2 \sigma^2. \quad (A.2)$$

Since $K_i(x - x_i; h, g) = \int h^{-1} K((x - y)/h) g^{-1} K((x_i - y)/g) dy$, we may use the mean-value theorem and the Lipschitz continuity of the kernel to show that an upper bound for (A.2) is provided by $C_1 n^{-1} (n^{-1} g^{-1} (h^{-1} + g^{-1}))^2$ for some constant C_1 . This shows that (A.2) is $o(n^{-1})$ under Condition 3.

Consider now the other terms of $b_n(x) - b_n^*(x)$, which may be grouped together as

$$\begin{aligned} T_2 &= \sum_i \alpha_i(x; g) m(x_i) - m(x) - \sum_i \alpha_i(x; h) \sum_j \alpha_j(x_i; g) m(x_j) \\ &\quad + \sum_i \alpha_i(x; h) m(x_i) + \sum_j \beta_j(x; h, g) m(x_j) \\ &\quad - \sum_i \alpha_i(x; g) m(x_i) - \frac{h^2}{2} \hat{m}''(x). \quad (A.3) \end{aligned}$$

The first two terms of (A.3) may be written as $(g^2/2) m''(x) + o(g^2)$, and the second two terms may be written as $-\sum_i \alpha_i(x; h) (g^2/2) m''(x_i) + o(g^2)$, where $o(g^2)$ is uniform in i since m'' is uniformly continuous. Boundary problems cannot occur since x was assumed to be between η and $1 - \eta$ and the kernel K has compact support. So, for n large enough, no design point with $x_i \geq 1 - \eta/2$ or $x_i \leq \eta/2$ enters the approximations of the terms in formula (A.3).

The first four terms of (A.3) can be replaced by the following integral:

$$-g^2/2 \int h^{-1} K((x - y)/h) \{m''(y) - m''(x)\} dy + o(g^2).$$

Splitting this integral and using continuity of m'' as in Parzen (1962), it can be shown that this term is $o(g^2)$.

Since the variance associated with the density $K(\cdot; h, g)$ is $(h^2 + g^2)$, the remaining terms of (A.3) may be written as

$$\begin{aligned} \frac{(h^2 + g^2)}{2} \cdot m''(x) - \frac{g^2}{2} m''(x) - \frac{h^2}{2} \hat{m}''(x) + o_p(h^2 + g^2) \\ = \frac{h^2}{2} \{m''(x) - \hat{m}''(x)\} + o_p(h^2 + g^2). \end{aligned}$$

Since $\hat{m}''(x)$ is a consistent estimator of $m''(x)$, these terms are $o_p(h^2 + g^2)$.

Collecting together all of the terms of the squared bias part of the Mallows metric, we now have that this converges in probability to 0. This general result allows us to use the bootstrap to investigate the distribution of any quantities of interest.

[Received June 1985. Revised April 1987.]

REFERENCES

Bickel, P. J., and Freedman, D. A. (1981). "Some Asymptotic Theory for the Bootstrap." *The Annals of Statistics*, 9, 1196-1217.
 Gasser, T., and Müller, H.-G. (1979). "Kernel Estimation of Regression Functions." in *Smoothing Techniques for Curve Estimation*, eds T. Gasser and M. Rosenblatt, Springer Lecture Notes, 757, 23-68.
 ——— (1984). "Estimating Regression Functions and Their Derivatives by the Kernel Method." *Scandinavian Journal of Statistics*, 11, 171-185.

- Gasser, T., Sroka, L., and Jennen-Steinmetz, C. (1986), "Residual Variance and Residual Pattern in Nonlinear Regression," *Biometrika*, 73, 625-634.
- Hall, P., and Titterton, D. M. (1986), "On Confidence Bands in Nonparametric Density Estimation and Regression," unpublished manuscript.
- Härdle, W., and Marron, J. S. (1985), "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation," *The Annals of Statistics*, 13, 1465-1481.
- Knäfl, G., Sacks, J., and Ylvisaker, D. (1985), "Confidence Bands for Regression Functions," *Journal of the American Statistical Association*, 80, 683-691.
- Parzen, E. (1962), "On Estimation of a Probability Density Function and Mode," *Annals of Mathematical Statistics*, 33, 1065-1076.
- Priestley, M. B., and Chao, M. J. (1972), "Non-parametric Function Fitting," *Journal of the Royal Statistical Society, Ser. B*, 34, 385-392.
- Rice, J. (1984), "Bandwidth Choice for Nonparametric Regression," *The Annals of Statistics*, 12, 1215-1230.
- Stone, C. J. (1980), "Optimal Rates of Convergence for Nonparametric Estimators," *The Annals of Statistics*, 8, 1348-1360.
- Stuetzle, W., and Mittal, Y. (1979), "Some Comments on the Asymptotic Behavior of Robust Smoothers," in *Smoothing Techniques for Curve Estimation*, eds. T. Gasser and M. Rosenblatt, Springer Lecture Notes, 757, 191-195.

STRONG UNIFORM CONSISTENCY RATES FOR ESTIMATORS OF CONDITIONAL FUNCTIONALS¹

BY W. HÄRDLE, P. JANSSEN AND R. SERFLING

*Universität Bonn, Limburgs Universitair Centrum and
The Johns Hopkins University*

Strong uniform consistency rates are established for kernel type estimators of functionals of the conditional distribution function, under general conditions. The present treatment unifies a number of specific problems previously studied separately in the literature. Some of these applications we treat in detail, including regression curve estimation, density estimation, estimation of conditional df's, L -smoothing and M -smoothing. Various previous results in the literature are extended and/or sharpened.

1. Introduction, basic formulation and applications. Let (X, Y) be a bivariate random vector with joint df $F(x, y)$, joint density $f(x, y)$, conditional df $F(y|x)$ for Y given X , conditional density $f(y|x)$ for Y given X and marginal density $f_0(x)$ for X , x and $y \in \mathbb{R}$. Let $\{\beta_t, t \in I\}$ be a family of real-valued measurable functions on \mathbb{R} for which it is desired to estimate

$$(1.1) \quad r_t(x) = E\{\beta_t(Y)|X = x\} = \int \beta_t(y) dF(y|x),$$

with a good almost sure (a.s.) convergence rate holding uniformly for $t \in I$ and $x \in J$, where I is a possibly infinite, or possibly degenerate, interval in \mathbb{R} and J is a possibly infinite interval in \mathbb{R} . In general, we may think of this type of problem as one of nonparametric estimation of linear functionals of the conditional df $F(y|x)$. As will be seen from the examples, such a problem may arise in nonparametric regression and related contexts, either as a given target problem or as a technical problem to which a given target problem becomes reduced.

Expressing $r_t(x)$ in the form $r_t(x) = d_t(x)/f_0(x)$, with

$$(1.2) \quad d_t(x) = \int \beta_t(y) f(x, y) dy,$$

we shall consider estimators of the form

$$(1.3a) \quad r_{tn}(x) = d_{tn}(x)/f_n(x),$$

with

$$(1.3b) \quad f_n(x) = (nh_n)^{-1} \sum_{i=1}^n K_{0n} \left(\frac{x - X_i}{h_n} \right)$$

Received July 1986; revised January 1988.

¹Research supported by the U.S. Department of Navy under Office of Naval Research Contract N00014-79-C-0801 and by NATO under Research Grant 0034/87. Reproduction in whole or in part is permitted for any purpose of the United States Government. Research of the third author also supported under a U.S. Senior Scientist Award by the Alexander von Humboldt-Stiftung.

AMS 1980 subject classifications. Primary 62G05; secondary 60F15.

Key words and phrases. Strong uniform consistency rates, nonparametric kernel estimators, density estimation, L -smoother, M -smoother, regression function estimation.

and

$$(1.3c) \quad d_{tn}(x) = (nh_n)^{-1} \sum_{i=1}^n \beta_t(Y_i) K_n \left(\frac{x - X_i}{h_n} \right),$$

where $(X_1, Y_1), \dots, (X_n, Y_n)$ are independent observations on F , $\{K_{0n}\}$ and $\{K_n\}$ are sequences of *kernel* functions $K: \mathbb{R} \rightarrow \mathbb{R}$ and $\{h_n\}$ is a sequence of positive constants (*bandwidths*) tending to 0 as $n \rightarrow \infty$. We recognize $f_n(\cdot)$ to be in the form of the familiar Rosenblatt-Parzen type of density estimator for $f_0(\cdot)$, except that we consider a sequence $\{K_{0n}\}$ instead of a fixed kernel K_0 .

The kernels under consideration may be smooth or discrete, although we shall give some emphasis to the discrete case. Since smooth kernels become discretized in computations with data, this case has considerable relevance to estimators actually computed in practice. We consider sequences instead of fixed kernels K_0 and K in order to include the case that K_{0n} and K_n are step-function kernels providing increasingly close approximation to given smooth kernels. The sequences $\{K_{0n}\}$ and $\{K_n\}$ may be selected to coincide, but this is not necessary, and we avoid such an assumption in order to provide greater flexibility in applications.

Under suitable restrictions, we shall establish the uniform a.s. rate

$$(1.4) \quad \sup_{t \in I} \sup_{x \in J} |r_{tn}(x) - r_t(x)| = O \left(\max \left\{ \left(\frac{\log n}{nh_n} \right)^{1/2}, h_n^\alpha \right\} \right) \quad \text{a.s., } n \rightarrow \infty,$$

where α is the order of uniform local Lipschitz (uLL) conditions imposed on $f_0(\cdot)$ and $\{d_t(\cdot), t \in I\}$. [Under stronger smoothness conditions, the component h_n^α in (1.4) can be improved.] By allowing a *family* of $\beta(\cdot)$ functions instead of a single one, we obtain a very useful type of extension of previous results in the literature, and our results also yield certain improvements in previously considered special cases.

Section 2 provides general theory, in which the most fundamental results are Theorems 2.1 and 2.2. These yield, in particular, a new result on density estimation (Corollary 2.1), our main result on (1.4) (Theorem 2.3) and a corollary giving conditions under which (1.4) provides the rate $O((n/\log n)^{-\alpha/(2\alpha+1)})$. Optimality of this rate, in the case of nonparametric regression function estimation, is shown in Stone (1982).

The crucial role of (1.4) in establishing uniform strong consistency rates for a variety of estimators involving conditional functionals may be seen from the following examples, which will be treated technically in Section 3 by systematically applying the theory of Section 2.

EXAMPLE 1. Nonparametric regression function estimation. This corresponds to (1.1) with the single $\beta(\cdot)$ function $\beta(y) = y$, in which case $r(x) = E(Y|X = x)$, the classical regression function, and $r_n(\cdot)$ represents the classical Nadaraya-Watson estimator [Nadaraya (1964) and Watson (1964)]. For general background, see Collomb (1981) and Mack and Silverman (1982), with whose results we make comparison in Section 3.

EXAMPLE 2. Nonparametric scale curve estimation. A nonparametric approach to the problem of heteroscedasticity in linear models involves estimation of the conditional variances

$$v(x) = E(Y^2|X = x) - [E(Y|X = x)]^2$$

[see Carroll (1982)]. Here the first component is given by (1.1) with the single function $\beta(y) = y^2$, and the second component is handled by Example 1. (Higher-order conditional moments may be treated similarly.)

EXAMPLE 3. The conditional df. The conditional df itself, i.e., the function $F(t|x)$, $t \in \mathbb{R}$, is given by (1.1) with $\beta_t(y) = I(y \leq t)$, $y \in \mathbb{R}$, $t \in I = \mathbb{R}$. The corresponding estimator $F_n(t|x)$ given by (1.3a) has been treated by Collomb (1980). He proved consistency results, without rates, which are uniform in x and pointwise in t . A Glivenko–Cantelli type theorem for $F_n(t|x)$, uniform in t and pointwise in x , is given in Stute (1986). Besides the intrinsic interest of the additional information provided by (1.4) in this case, such a result also plays a fundamental role in obtaining uniform strong consistency rates in other problems, as in Example 5.

EXAMPLE 4. The marginal density f_0 . With the single trivial function $\beta(y) \equiv 1$, $d_n(\cdot)$ given by (1.3c) becomes a density estimator for $f_0(\cdot)$. A key theoretical tool (Theorem 2.2) in Section 2 concerns the behavior of $\sup_t \sup_x |d_{tn}(x) - d_t(x)|$ and, for this choice of $\{\beta_t, t \in I\}$, yields new results on density estimation [see Corollary 2.1 and Remark 2.3(i)].

EXAMPLE 5. L -smoothing. Denote by $F^{-1}(v|x) = \inf\{y: F(y|x) \geq v\}$, $0 < v < 1$, the conditional quantile function associated with $F(\cdot|x)$ and consider estimation of a conditional L -functional

$$l(x) = \int_0^1 J(v) F^{-1}(v|x) dv.$$

For $J(v) \equiv 1$, $l(x)$ reduces to the regression function $r(x)$ considered in Example 1. The same occurs in the case $J(v) = I\{p \leq v \leq 1 - p\}/(1 - 2p)$, where $0 < p < 1/2$, with $f(y|x)$ symmetric about $r(x)$. Letting $F_n(t|x)$ denote the estimator of $F(t|x)$ considered in Example 3, we consider for $l(x)$ the estimator $l_n(x)$ produced by substituting $F_n^{-1}(v|x)$ for $F^{-1}(v|x)$. In our treatment in Section 3, we obtain uniform strong consistency rates for trimmed L -smoothers by reduction of the problem to an application of results obtained for Example 3.

EXAMPLE 6. M -smoothing. For any given real function $\psi(\cdot)$, a corresponding M -functional $T_\psi(\cdot)$ may be defined on df's G by letting $T_\psi(G)$ denote the solution t_0 of the equation

$$\int \psi(y - t_0) dG(y) = 0.$$

[The case $\psi(x) = x$ yields $T_\psi(G) = \int y dG(y)$, the mean functional.] In the case

that $G(\cdot)$ is symmetric about θ , any antisymmetric ψ yields $T_\psi(G) = \theta$. Thus, for a class of such $\psi(\cdot)$, a class of competing estimators of θ is given by $T_\psi(\hat{G})$, with \hat{G} estimating G .

Adapting this to regression curve estimation, we let $r(x)$ be as in Example 1 and assume that, for each $x \in J$, the conditional density $f(y|x)$ is symmetric about $r(x)$. Then, for antisymmetric ψ , $r(x)$ is the solution of the preceding equation with $G(\cdot)$ replaced by $F(\cdot|x)$ and an estimator $r_{\psi n}(x)$ is given by solving this equation with $G(\cdot)$ replaced by $F_n(\cdot|x)$ defined as in Example 3. For suitable choice of $\psi(\cdot)$, the function $r_{\psi n}(x)$ for estimation of $r(x)$, $x \in J$, is more resistant to the presence of outliers than is the estimator $r_n(x)$ treated previously. We call $r_{\psi n}(x)$, $x \in J$, the M -smoother corresponding to ψ . Pointwise consistency of M -smoothers has been treated by Stone (1977), Tsybakov (1983) and Härdle (1984). Uniform *weak* consistency rates have been established by Härdle and Luckhaus (1984), by reduction, with $\beta_t(y) = \psi(y - t)$, $y \in \mathbb{R}$, to the analogous problem for the estimators $r_{tn}(x)$ of $r_t(x)$, for t in a small neighborhood of $r(x)$. Following this approach, we establish a.s. uniform consistency rates for M -smoothers in Section 3.

Our method in Section 2 will be to handle $r_{tn} - r_t$ via the decomposition

$$(1.5) \quad r_{tn} - r_t = R_{tn} + S_{tn},$$

with $R_{tn} = (d_{tn} - d_t)/f_n$ and $S_{tn} = d_t(f_0 - f_n)/(f_0 f_n)$. As noted in Example 4, results for $f_n - f_0$ may be obtained by specialization of results for $d_{tn} - d_t$. Thus our treatment of $r_{tn} - r_t$ via (1.5) will flow from study of $d_{tn} - d_t$. For this we shall provide key foundational results in Theorems 2.1 and 2.2, from which our target results, Corollary 2.1, Theorem 2.3 and Corollary 2.2 will be derived. We shall deal with the stochastic component $d_{tn} - Ed_{tn}$ of $d_{tn} - d_t$ by analyzing, in effect, the modulus of continuity of a certain randomly weighted empirical df. The related bias component $Ed_{tn} - d_t$ will be handled by imposing mild local Lipschitz conditions. Without such conditions (for example, assuming only uniform continuity), the *rate* of convergence of the bias to 0 cannot be precisely characterized and thus a rate cannot properly be asserted in (1.4).

2. Some general results on strong uniform consistency rates. Our target results will follow from a basic theorem we establish on convergence of quantities of the form

$$(2.1) \quad \sup_{t \in I} \sup_{x \in J} |D_{tn}(x) - D_t(x)|,$$

with

$$(2.2a) \quad D_t(x) = \int_{\mathbb{R}} \gamma_t(y) f(x, y) dy, \quad x \in \mathbb{R},$$

$$(2.2b) \quad D_{tn}(x) = c_n^{-1} [G_{tn}(x + c'_n) - G_{tn}(x - c''_n)], \quad x \in \mathbb{R},$$

$\{c'_n\}$ and $\{c''_n\}$ nonnegative sequences tending to 0, $c_n = c'_n + c''_n$ and

$$(2.2c) \quad G_{tn}(x) = n^{-1} \sum_{i=1}^n \gamma_t(Y_i) I\{X_i \leq x\}, \quad x \in \mathbb{R}.$$

Here (X, Y) , (X_i, Y_i) , $1 \leq i \leq n$, $f(x, y)$, $f(y|x)$, $f_0(x)$, I and J will be as in Section 1, but the functions $\{\beta_t, t \in I\}$ there are replaced for the present by a family $\{\gamma_t, t \in I\}$ satisfying some specialized assumptions, and for the moment we do not concern ourselves with kernels $\{K_n\}$. The "randomly weighted" empirical df G_{tn} has mean function

$$(2.3) \quad G_t(x) = EG_{tn}(x) = \int_{-\infty}^x D_t(z) dz,$$

and we readily see by the classical SLLN that for each fixed pair t and x , $D_{tn}(x) \rightarrow D_t(x)$ a.s., $n \rightarrow \infty$. Our purpose here is to strengthen this by giving a rate for this convergence uniformly in $t \in I$ and $x \in J$.

The following assumptions come into play. We define a function g on \mathbb{R} to be *uniformly locally Lipschitz of order α* (uLL- α), where $0 < \alpha \leq 1$, if for some $\delta > 0$ and $M < \infty$, $\sup_{x \in \mathbb{R}} |g(x+z) - g(x)| \leq M|z|^\alpha$, for $|z| \leq \delta$.

ASSUMPTIONS.

$$(A.1) \quad \sup_{t \in I} \sup_{x \in J} \int_{\mathbb{R}} \gamma_t^2(y) f(y|x) dy = M_0 < \infty.$$

$$(A.2) \quad \sup_{x \in J} f_0(x) = M_1 < \infty.$$

$$(A.3) \quad 0 \leq \gamma_t(y) \leq \gamma_{t'}(y), \quad t < t' \in I, y \in \mathbb{R}.$$

(A.4) For some α , $0 < \alpha \leq 1$, $D_t(\cdot)$ is uLL- α on J , uniformly for $t \in I$; i.e., for some $\delta_\alpha > 0$ and $M^{(\alpha)} < \infty$,

$$\sup_{t \in I} \sup_{\substack{x, x' \in J \\ |x-x'| \leq \delta_\alpha}} |D_t(x) - D_t(x')| \leq M^{(\alpha)} |x - x'|^\alpha.$$

(A.5) $E\gamma_t(Y)$ is a continuous function of t in I .

(A.6) The limit functions $\gamma_{t_*} = \lim_{t \rightarrow t_*} \gamma_t$ and $\gamma_{t^*} = \lim_{t \rightarrow t^*} \gamma_t$ exist and are finite a.s. (w.r.t. the df of Y), where $t_* = \inf I (\geq -\infty)$ and $t^* = \sup I (\leq +\infty)$.

(A.7) $(E|\gamma_{t^*}(Y)|^\lambda)^{1/\lambda} = M_\lambda < \infty$ for some λ , $2 < \lambda \leq \infty$ [in the case $\lambda = \infty$, M_∞ denotes $\sup_{y \in \mathbb{R}} |\gamma_{t^*}(y)|$].

REMARKS 2.1. Consider the assumptions:

$$(B.1) \quad \inf_{x \in J} f_0(x) = m_1 > 0;$$

$$(B.2) \quad \sup_{t \in I} \sup_{x \in J} \int_{\mathbb{R}} \gamma_t^2(y) f(x, y) dy = M_0^* < \infty;$$

$$(B.3) \quad \sup_{t \in I} \sup_{x \in J} |D_t(x)| = M_2 < \infty.$$

By simple arguments, we obtain:

- (i) Under (A.1) and (A.2), (B.2) holds with $M_0^* \leq M_0 M_1$.
- (ii) Under (B.1) and (B.2), (A.1) holds with $M_0 \leq M_0^* / m_1$.
- (iii) Under (A.2) and (B.2), (B.3) holds with $M_2 \leq (M_1 M_0^*)^{1/2}$.
- (iv) If $\sup_{t \in I} \sup_{y \in \mathbb{R}} |\gamma_t(y)| < \infty$, then (A.1) holds with $J = \mathbb{R}$.

For the case $\gamma_t(y) = y$, (B.2) is a type of assumption used by Mack and Silverman (1982), who also assumed (B.1) and (A.2). Statements (i) and (ii) indicate that we are enabled to have (A.1), (A.2) and (B.2) while bypassing (B.1), thus providing our result on the quantity (2.1) with a broader scope of potential application. [However, in dealing with r_{tn} and establishing (1.4), we will need (B.1).] Assumption (A.1) may be interpreted as requiring the conditional second moments $E[\gamma_t^2(Y)|X = x]$ to be uniformly bounded for $t \in I$, $x \in J$. Statements (i) and (iii) will be used in the proof of Theorem 2.3.

THEOREM 2.1. *Assume (A.1)–(A.7). Let $\{c_n\}$ satisfy (i) $0 \leq c_n \rightarrow 0$, (ii) $\Delta_n = nc_n/\log n \rightarrow \infty$ and (iii) $1 \leq c_n^{-1} \leq (n/\log n)^{1-2/\lambda}$, for λ as in (A.7). Then, with α as in (A.4),*

$$(2.4a) \quad \sup_{t \in I} \sup_{x \in J} |D_{tn}(x) - D_t(x)| = O(\max\{\Delta_n^{-1/2}, c_n^\alpha\}) \quad \text{a.s., } n \rightarrow \infty.$$

Further, in the case $\lambda = \infty$ in (A.7), there exists a number n_0 and for each real $\kappa > 0$ there exists a constant A_κ , not depending on the sequence $\{c_n\}$, such that

$$(2.4b) \quad P\left\{\sup_{t \in I} \sup_{x \in J} |D_{tn}(x) - D_t(x)| > A_\kappa \Delta_n^{-1/2} + M^{(\alpha)} c_n^\alpha\right\} < n^{-\kappa},$$

all $\kappa > 0$ and $n \geq n_0$, with $M^{(\alpha)}$ as in (A.4).

REMARKS 2.2. (i) By the Borel–Cantelli lemma, if (2.4b) holds, then so does (2.4a).

(ii) From (2.7), Lemma 2.1 and the proof of Lemma 2.2, it will be seen that the constant A_κ in (2.4b) may be taken as $6A + 4$, where A is chosen (sufficiently large) to satisfy

$$\frac{(A - M_1)^2}{2M_0M_1 + \frac{2}{3}(A - M_1)M_\infty} \geq \kappa + 2,$$

with M_0, M_1, M_∞ as in (A.1), (A.2) and (A.7).

(iii) Note that for $2 < \lambda < \infty$ condition (iii) implies condition (ii). For $\lambda = \infty$ the right inequality of condition (iii) follows from condition (ii).

We prove the theorem by decomposing $D_{tn} - D_t$ into a stochastic component $A_{tn} = D_{tn} - ED_{tn}$ and a deterministic (bias) component $B_{tn} = ED_{tn} - D_t$, each to be treated separately. For the bias part, we readily obtain, using (2.3),

LEMMA 2.1. *Under (A.4) and for $c_n \rightarrow 0$,*

$$(2.5) \quad \sup_{t \in I} \sup_{x \in J} |B_{tn}(x)| \leq M^{(\alpha)} c_n^\alpha, \quad \text{for all large } n.$$

For the stochastic component, it is easily checked that $|A_{tn}(x)| \leq 2c_n^{-1}V_{tn}(x, c_n)$, where

$$(2.6) \quad V_{tn}(x, \delta) = \sup_{|z| \leq \delta} |G_{tn}(x+z) - G_{tn}(x) - [G_t(x+z) - G_t(x)]|.$$

Putting

$$V_n = \sup_{t \in I} \sup_{x \in J} V_{tn}(x, c_n),$$

we have

$$(2.7) \quad \sup_{t \in I} \sup_{x \in J} |A_{tn}(x)| \leq 2c_n^{-1} V_n.$$

Consequently, Theorem 2.1 follows from Lemma 2.1 with the following central result.

LEMMA 2.2. *Under the conditions of Theorem 2.1, excepting (A.4),*

$$(2.8a) \quad V_n = O(\Delta_n^{-1/2} c_n) \quad \text{a.s., } n \rightarrow \infty.$$

Further, in the case $\lambda = \infty$ in (A.7), there exists a number n_0 and for each real $\kappa > 0$ there exists a constant B_κ , not depending on the sequence $\{c_n\}$, such that

$$(2.8b) \quad P\{V_n > B_\kappa \Delta_n^{-1/2} c_n\} < n^{-\kappa}, \quad \text{all } \kappa > 0 \text{ and } n \geq n_0.$$

The proof of this lemma is given in Section 4.

From Theorem 2.1, we now can establish analogous results for the estimators d_{tn} and r_{tn} introduced in Section 1, in the case of $\{\beta_t, t \in I\}$.

We shall assume that the family $\{\beta_t, t \in I\}$ has a representation

$$(2.9) \quad \beta_t(y) = \sum_{i=1}^{i_0} q_i \gamma_{ti}(y), \quad y \in \mathbb{R}, t \in I,$$

with fixed and finite i_0, q_1, \dots, q_{i_0} and with the families $\{\gamma_{ti}, t \in I\}, 1 \leq i \leq i_0$, satisfying assumptions (A.1) and (A.3)–(A.7), with common α in (A.4) and common λ in (A.7).

We first consider kernel sequences of *step-function* form,

$$(2.10a) \quad K_n(u) = \sum_{j=1}^{j_n} a_{nj} I\{-b''_{nj} \leq u < b'_{nj}\}, \quad u \in \mathbb{R},$$

with $\{j_n\}, \{a_{nj}\}, \{b'_{nj}\}, \{b''_{nj}\}$ nonnegative constants such that, with $b_{nj} = b'_{nj} + b''_{nj}$,

$$(2.10b) \quad \sum_{j=1}^{j_n} a_{nj} b_{nj} = 1 \quad \left[\text{i.e., } \int K_n(u) du = 1 \right],$$

$$(2.10c) \quad \sup_n \sum_{j=1}^{j_n} a_{nj} b_{nj}^{1/2} < \infty$$

and

$$(2.10d) \quad \sup_n \sum_{j=1}^{j_n} a_{nj} b_{nj}^2 < \infty.$$

THEOREM 2.2. *Let $d_{tn}(\cdot)$ be defined by (1.3c), with $\{\beta_t, t \in I\}$ having representation (2.9), $\{K_n\}$ having form (2.10) and $\{h_n\}$ satisfying (i) $h_n B_n \rightarrow 0$,*

(ii) $nh_n b_n / \log n \rightarrow \infty$ and (iii) $B_n \leq h_n^{-1} \leq b_n (n / \log n)^{1-2/\lambda}$, where $b_n = \min_{j \leq j_n} b_{nj}$, $B_n = \max_{j \leq j_n} b_{nj}$ and λ is given in (A.7). Assume also (A.2) and either

$$(2.11a) \quad \lambda < \infty; \quad j_n \equiv j_0 < \infty$$

or

$$(2.11b) \quad \lambda = \infty; \quad j_n = O(n^s), \quad \text{some } s > 0.$$

Then

$$(2.12) \quad \sup_{t \in I} \sup_{x \in J} |d_{tn}(x) - d_t(x)| = O(\max\{\Delta_n^{-1/2}, h_n^\alpha\}) \quad \text{a.s., } n \rightarrow \infty,$$

with $\Delta_n = nh_n / \log n$ and α as in (A.4).

PROOF. With the assumed forms for $\{\beta_t, t \in I\}$ and $\{K_n\}$, the estimator d_{tn} has a decomposition into terms of the type treated in Theorem 2.1, and accordingly we obtain

$$(2.13) \quad \sup_{t \in I} \sup_{x \in J} |d_{tn}(x) - d_t(x)| \leq \sum_{i=1}^{i_0} |q_i| S_{ni},$$

where

$$(2.14) \quad S_{ni} = \sum_{j=1}^{j_n} a_{nj} b_{nj} \sup_{t \in I} \sup_{x \in J} |D_{tn}^{(i,j)}(x) - D_t^{(i)}(x)|,$$

with

$$\begin{aligned} D_t^{(i)}(x) &= \int_{\mathbf{R}} \gamma_{ti}(y) f(x, y) dy, \\ D_{tn}^{(i,j)}(x) &= c_{nj}^{-1} [G_{tn}^{(i)}(x + c'_{nj}) - G_{tn}^{(i)}(x - c''_{nj})], \\ G_{tn}^{(i)}(x) &= n^{-1} \sum_{k=1}^n \gamma_{ti}(Y_k) I\{X_k \leq x\} \end{aligned}$$

and

$$c'_{nj} = h_n b'_{nj}, \quad c''_{nj} = h_n b''_{nj}, \quad c_{nj} = h_n b_{nj}.$$

Fix i and j and put $\Delta_{nj} = \Delta_n b_{nj}$. Then conditions (i), (ii) and (iii) assumed in the present theorem yield their counterparts in Theorem 2.1 with $\{c_n\}$ replaced by $\{c_{nj}\}_{n \geq 1}$ and Δ_n replaced by Δ_{nj} , $n \geq 1$, and Theorem 2.1 thus yields

$$(2.15) \quad \sup_{t \in I} \sup_{x \in J} |D_{tn}^{(i,j)}(x) - D_t^{(i)}(x)| = O(\max\{\Delta_{nj}^{-1/2}, c_{nj}^\alpha\}) \quad \text{a.s., } n \rightarrow \infty.$$

Now suppose that (2.11a) holds. Then (2.15) yields

$$(2.16) \quad S_{ni} = O\left(\sum_{j=1}^{j_0} a_{nj} b_{nj} \max\{\Delta_{nj}^{-1/2}, c_{nj}^\alpha\}\right) \quad \text{a.s., } n \rightarrow \infty.$$

It is easily seen, using (2.10c) and (2.10d) that the right-hand side of (2.16) is $O(\Delta_n^{-1/2}, h_n^\alpha)$, and thus (2.12) follows, via (2.13).

Alternatively, assume (2.11b), choose real $\kappa > 0$ and put

$$\epsilon_n = A_\kappa \Delta_n^{-1/2} \sum_{j=1}^{j_n} a_{nj} b_{nj}^{1/2} + M^{(\alpha)} h_n^\alpha \sum_{j=1}^{j_n} a_{nj} b_{nj}^{1+\alpha},$$

with $A_\kappa, M^{(\alpha)}$ as in Theorem 2.1. Then

$$P\{S_{ni} > \epsilon_n\} \leq \sum_{j=1}^{j_n} P\left\{ \sup_{t \in I} \sup_{x \in J} |D_{tn}^{(i,j)}(x) - D_t^{(i)}(x)| > A_\kappa \Delta_n^{-1/2} + M^{(\alpha)} c_{nj}^\alpha \right\}$$

and by (2.4b) of Theorem 2.1 we obtain

$$P\{S_{ni} > \epsilon_n\} \leq j_n n^{-\kappa}, \quad \text{all } n \geq n_0.$$

Choosing $\kappa > s + 1$, with s as in (2.11b), and applying the Borel-Cantelli lemma, we obtain

$$(2.17) \quad S_{ni} = O(\epsilon_n) \quad \text{a.s., } n \rightarrow \infty.$$

Again using (2.10c) and (2.10d), we have $\epsilon_n = O(\max\{\Delta_n^{-1/2}, h_n^\alpha\})$, $n \rightarrow \infty$, and thus (2.12) follows, via (2.13). \square

As discussed in Example 4 of Section 1, Theorem 2.2 yields a result on density estimation with discrete kernels, as follows.

COROLLARY 2.1. *Let $f_n(\cdot)$ be defined by (1.3b) with $\{K_{0n}\}$ having form (2.10) with $j_n = O(n^s)$, some $s > 0$, and with $\{h_n\}$ satisfying (i), (ii) and (iii) of Theorem 2.2 with $\lambda = \infty$. Assume (A.2) and that f_0 is uLL- α on J for some α , $0 < \alpha \leq 1$. Then*

$$(2.18) \quad \sup_{x \in J} |f_n(x) - f_0(x)| = O(\max\{\Delta_n^{-1/2}, h_n^\alpha\}) \quad \text{a.s., } n \rightarrow \infty,$$

with $\Delta_n = nh_n/\log n$.

PROOF. With the family $\{\beta_t, t \in I\}$ reduced to the single function $\beta(y) \equiv 1$, d_{tn} given by (1.3c) reduces to f_n in form, and, under the present assumptions, the conditions of Theorem 2.2 are satisfied with the option (2.11b). Thus (2.12) holds, which is the same as (2.18). \square

This corollary not only extends the results of Serfling (1982) to a wider scope of kernels and thus is of independent interest, but also serves as a tool in developing our result for r_{tn} , as follows.

THEOREM 2.3. (i) *Discrete kernels.* Let $r_{tn}(\cdot)$ be defined by (1.3) with $\{\beta_t, t \in I\}$ having representation (2.9), $\{K_n\}$ having form (2.10) and $\{K_{0n}\}$ having form (2.10) with constants $\{\tilde{j}_n\}, \{\tilde{a}_{nj}\}, \{\tilde{b}'_{nj}\}, \{\tilde{b}''_{nj}\}$ and with $\tilde{j}_n = O(n^s)$ for some

$\tilde{s} > 0$. Let $\{h_n\}$ satisfy

- (a) $h_n \max\{B_n, \tilde{B}_n\} \rightarrow 0$,
- (b) $(\log n)^{-1} n h_n \min\{b_n, \tilde{b}_n\} \rightarrow \infty$,
- (c) $B_n \leq h_n^{-1} \leq b_n (n/\log n)^{1-2/\lambda}$ and
- (d) $\tilde{B}_n \leq h_n^{-1} \leq \tilde{b}_n (n/\log n)$,

with $b_n = \min_{j \leq j_n} b_{nj}$, $B_n = \max_{j \leq j_n} b_{nj}$, $\tilde{b}_n = \min_{j \leq \tilde{j}_n} \tilde{b}_{nj}$, $\tilde{B}_n = \max_{j \leq \tilde{j}_n} \tilde{b}_{nj}$ and λ as in (A.7). Assume (A.2), (B.1), f_0 ull- α_0 on J for some α_0 , $0 < \alpha_0 \leq 1$, and either (2.11a) or (2.11b). Then

$$(2.19) \quad \sup_{t \in I} \sup_{x \in J} |r_{tn}(x) - r_t(x)| = O(\max\{\Delta_n^{-1/2}, h_n^{\tilde{\alpha}}\}) \quad \text{a.s., } n \rightarrow \infty,$$

with $\Delta_n = nh_n/\log n$ and $\tilde{\alpha} = \min\{\alpha, \alpha_0\}$, for α as in (A.4).

(ii) Smooth kernels. Let $r_{tn}(\cdot)$ be defined by (1.3) with $\{\beta_t, t \in I\}$ having representation (2.9) and with $K_n(\cdot) \equiv K(\cdot)$, $K_{0n}(\cdot) \equiv K_0(\cdot)$, where K is symmetric, has bounded support and bounded first two derivatives and K_0 satisfies similar conditions. Assume (A.2), (B.1) and f_0 ull- α_0 on J for some α_0 , $0 < \alpha_0 \leq 1$. Assume that $\{h_n\}$ satisfies

- (a) $h_n \rightarrow 0$,
- (b) $nh_n/\log n \rightarrow \infty$ and
- (c) $B \leq h_n^{-1} \leq b(n/\log n)^{1-2\lambda}$,

for some constants b and B and for λ as in (A.7). Then (2.19) holds.

PROOF. (i) It is immediate that, under the assumptions of the present theorem, the conditions of Theorem 2.2 and Corollary 2.1 are satisfied, and we have (2.12) as well as

$$(2.20) \quad \sup_{x \in J} |f_n(x) - f_0(x)| = O(\max\{\Delta_n^{-1/2}, h_n^{\alpha_0}\}) \quad \text{a.s., } n \rightarrow \infty.$$

We now apply relation (1.5). By (B.1), (2.12) and (2.20), we have

$$(2.21) \quad \sup_{t \in I} \sup_{x \in J} |R_{tn}(x)| = O(\max\{\Delta_n^{-1/2}, h_n^{\alpha}\}) \quad \text{a.s., } n \rightarrow \infty.$$

Using (B.1) again, as well as (B.3) [see statements (i) and (iii) of Remarks 2.1] and (2.20), we have

$$(2.22) \quad \sup_{t \in I} \sup_{x \in J} |S_{tn}(x)| = O(\max\{\Delta_n^{-1/2}, h_n^{\alpha_0}\}) \quad \text{a.s., } n \rightarrow \infty.$$

Combining (2.21) and (2.22), we have (2.19).

(ii) Let K be symmetric with bounded support, say $\subset [-1, 1]$ and let us introduce an associated sequence $\{K_n\}$ of discrete kernels, defined by

$$K_n(u) = \sum_{i=1}^{j_n} I\{(i-1)\delta_n < u \leq i\delta_n\} K(i\delta_n), \quad u > 0,$$

where $j_n = [\delta_n^{-1}] + 1$, with $0 < \delta_n \rightarrow 0$ and $K_n(u) = K_n(-u)$, for $u < 0$, and

$K_n(0) = K(\delta_n)$. For this kernel the regularity conditions (2.10c) and (2.10d) reduce to

$$\sup_n \sum_{i=1}^{j_n} [K(i\delta_n) - K((i+1)\delta_n)](i\delta_n)^{1/2} < \infty$$

and

$$\sup_n \sum_{i=1}^{j_n} [K(i\delta_n) - K((i+1)\delta_n)](i\delta_n)^2 < \infty.$$

For K'' bounded, these reduce to

$$\sup_n \delta_n^{3/2} \sum_{i=1}^{j_n} i^{1/2} |K'(i\delta_n)| < \infty$$

and

$$\sup_n \delta_n^3 \sum_{i=1}^{j_n} i^2 |K'(i\delta_n)| < \infty,$$

which in turn are satisfied if we have $\int x^2 |K'(x)| dx < \infty$, which indeed follows from our restrictions on $K(\cdot)$. Similar considerations apply in connection with $K_0(\cdot)$.

Now note that for $\gamma_t(\cdot)$ and $K(\cdot)$ bounded, we have

$$\sup_{t \in I} \sup_{x \in J} |d_{tn}(x; K) - d_{tn}(x; K_n)| = O\left(\frac{\delta_n}{h_n}\right),$$

where $d_{tn}(x; L)$ denotes (1.3c) based on the kernel $L(\cdot)$. Therefore, we can take $\delta_n = O(h_n^2)$ so that $j_n = O(h_n^{-2})$, which is $O(n^s)$, for some $s > 0$ under our condition on the bandwidth. Thus we may now apply part (i) to obtain (2.19) again in the present case. \square

Useful corollaries of Theorem 2.3 are obtained by choosing h_n to make the rates $\Delta_n^{-1/2}$ and $h_n^{\tilde{\alpha}}$ agree. In particular, for the case of discrete kernels we have

COROLLARY 2.2. *Let $r_{tn}(\cdot)$ be defined by (1.3) with $\{\beta_t, t \in I\}$ having representation (2.9), $\{K_n\}$ having form (2.10) and $\{K_{0n}\}$ having form (2.10) with constants $\{\tilde{j}_n\}, \{\tilde{\alpha}_{nj}\}, \{\tilde{b}'_{nj}\}, \{\tilde{b}''_{nj}\}$ and with $\tilde{j}_n = O(n^{\tilde{s}})$, for some $\tilde{s} > 0$. Assume $\alpha = 1$, in (A.4) and $\lambda > 3$ in (A.7). Assume (A.2), (B.1), f_0 uLL-1 on J and either (2.11a) or (2.11b). With the notation of Theorem 2.3, assume*

- (i) $\max\{B_n, \tilde{B}_n\} = o((n/\log n)^{1/3}),$
- (ii) $(n/\log n)^{-2/3} = o(\min\{b_n, \tilde{b}_n\}) \leq c_0 \tilde{b}_n,$
- (iii) $(n/\log n)^{-2(1/3-1/\lambda)} \leq c_0 \tilde{b}_n,$

for some constant $c_0 > 0$. Let $h_n \sim c_0(n/\log n)^{-1/3}$ in (1.3). Then

$$(2.23) \quad \sup_{t \in I} \sup_{x \in J} |r_{tn}(x) - r_t(x)| = O((n/\log n)^{-1/3}) \quad a.s., n \rightarrow \infty.$$

PROOF. It is readily seen that this choice of h_n and (i), (ii) and (iii) in the preceding discussion yield (a)–(d) of Theorem 2.3(i). Also, the other assumptions of Theorem 2.3(i) are obviously fulfilled by the present assumptions. Thus (2.19) holds and yields (2.23). \square

REMARKS 2.3. (i) An analogue of (2.23) for the density estimator f_n may be easily derived.

(ii) It would be of interest, in the case $\lambda < \infty$, to relax the restriction (2.11a) on $\{j_n\}$ in Theorems 2.2 and 2.3(i) to a condition of form $j_n = O(n^s)$, for some $s > 0$. However, this would require a strengthened version of Lemma 2.2 with (2.8b) extended to the case $\lambda < \infty$. The present proof of Lemma 2.2, given in Section 4, does not appear to yield such a strengthening, due to the complication presented by the truncation step involving the random variable W_n in (4.9). A possible approach could be to control the rate at which the r.h.s. of (4.11) converges to 0.

(iii) From the proofs of Theorem 2.2, Corollary 2.1, Theorem 2.3(i) and Corollary 2.2, it is clear that the restrictions on $\{K_n\}$ and $\{K_{0n}\}$ may be dropped or relaxed, at the expense of introducing further factors (involving $b_n, \tilde{b}_n, B_n, \tilde{B}_n$, etc.) into the rates expressed in the relations (2.12), (2.18), (2.19) and (2.22).

(iv) In the case of single step-function kernels $K(\cdot)$ and $K_0(\cdot)$, with finitely many jumps in place of the sequences $\{K_n\}$ and $\{K_{0n}\}$, the restrictions on $\{h_n\}$ in Theorem 2.2, Corollary 2.1 and Theorem 2.3(i) reduce to those given by (a), (b) and (c) in Theorem 2.3(ii), with $b = \min\{b_1, \dots, b_{j_0}, \tilde{b}_1, \dots, \tilde{b}_{\tilde{j}_0}\}$, $B = \max\{b_1, \dots, b_{j_0}, \tilde{b}_1, \dots, \tilde{b}_{\tilde{j}_0}\}$ and λ as in (A.7)

(v) From the proof of Theorem 2.2 it is easily seen that in the case $\lambda = \infty$ we may express (2.12) in the form

$$\sup_{t \in I} \sup_{x \in J} |d_{t_n}(x) - d_t(x)| \leq A\Delta_n^{-1/2} + A'h_n^\alpha, \quad \text{all large } n, \text{ a.s.},$$

with A and A' constants not depending on n .

(vi) We may also consider smooth kernels with *unbounded* support, by restricting attention to a finite interval of increasing length. For example, in the case of the standard normal density, we restrict to $[-t_n, t_n]$ with $t_n = n^\alpha$ for some $\alpha > 0$, take $\delta_n = n^{-\beta} \geq \exp(-n^{2\alpha}/2)$ for some $\beta > 0$ and finally note that $j_n = O(t_n \delta_n^{-1}) = O(n^{\alpha+\beta})$.

3. Strong consistency rates in selected applications. Using Theorem 2.3(i) and Corollary 2.2, we develop strong consistency rates for the applications discussed in Section 1, except for density estimation (Example 4), which has been treated in Corollary 2.1.

For convenience and simplicity, we confine our attention to the case that the kernels in (1.3b) and (1.3c) are step-functions not depending on n and having finitely many jumps. Thus [see Remark 2.3(iv)] throughout this section we shall assume the following standard conditions and notation with respect to the

bandwidth sequence $\{h_n\}$ and kernels K and K_0 in (1.3b) and (1.3c):

$$(3.1a) \quad h_n \rightarrow 0,$$

$$(3.1b) \quad \Delta_n = nh_n/(\log n) \rightarrow \infty,$$

$$(3.1c) \quad B \leq h_n^{-1} \leq b(n/\log n)^{1-2/\lambda},$$

with b, B defined as in Remark 2.3(iv) and λ a constant to be specified in each particular application.

All of the results to be given have extensions to general kernel sequences of form (2.39), at the expense of complicating the formulation. We also could develop some analogous results for smooth kernels, but we omit this in the interest of brevity.

3.1. Nonparametric regression function estimation. As in Example 1, we take $\beta_1(y) \equiv \beta(y) = y$, in which case the representation (2.38) becomes $\beta(y) = \gamma_1(y) - \gamma_2(y)$, with $\gamma_1(y) = \max\{0, y\}$ and $\gamma_2(y) = -\min\{0, y\}$. Then the assumptions (A.1)–(A.7), (B.1) and f_0 uLL may be reduced to:

$$(3.2a) \quad \sup_{x \in J} \int_{\mathbf{R}} y^2 f(y|x) dy = M_0 < \infty;$$

$$(3.2b) \quad E|Y|^\lambda < \infty, \quad \text{with } 2 < \lambda \leq \infty;$$

$$(3.2c) \quad 0 < m_1 \leq f_0(x) \leq M_1 < \infty, \quad x \in J;$$

$$(3.2d) \quad \text{the functions } f_0(x) \text{ and } g_0(x) = \int y f(x, y) dy \text{ are uLL-}\alpha \text{ on } J, \\ \text{with } 0 < \alpha \leq 1.$$

Thus Theorem 2.3(i) and Corollary 2.2 yield the following result.

THEOREM 3.1. Assume (3.2) and let $\{h_n\}$ satisfy (3.1) with λ as in (3.2b). Then

$$(3.3) \quad \sup_{x \in J} |r_n(x) - r(x)| = O(\max\{\Delta_n^{-1/2}, h_n^\alpha\}) \quad \text{a.s., } n \rightarrow \infty.$$

In the case $\alpha = 1$ and $\lambda \geq 3$ and for $h_n \sim C_0(n/\log n)^{-1/3}$, we have $O((n/\log n)^{-1/3})$ in (3.3).

Let us compare, for example, with Theorem B of Mack and Silverman (1982). There the kernels $K(\cdot)$ and $K_0(\cdot)$ are taken to be equal, symmetric and subject to a set of smoothness conditions. Our (3.2b) and (3.2c) are also assumed, but in place of (3.2a) is the stronger requirement (see Remark 2.1) $\sup_{x \in J} \int_{\mathbf{R}} |y|^\lambda f(x, y) dy < \infty$, with λ as in (3.2b). Also, the functions f_0 and g_0 in our (3.2d) are assumed to have bounded second derivatives [thus implying (3.2d) with $\alpha = 1$]. As bandwidth assumptions, our (3.1a) and an equivalent of (3.1b) are assumed, and a slightly stronger version of (3.1c), namely that $n^\eta h_n \rightarrow \infty$ for some $\eta < 1 - 2/\lambda$, is assumed. Also, $\sum_n h_n^s < \infty$ for some $s > 0$, and (*) $h_n = O(\tilde{\Delta}_n^{-1/2})$, where $\tilde{\Delta}_n = nh_n/\log(1/h_n)$, are assumed. (Note that $\tilde{\Delta}_n \sim \Delta_n$ for all typical choices of $\{h_n\}$.) Their theorem asserts for the quantity in (3.3) the a.s.

rate $O(\tilde{\Delta}_n^{-1/2})$, which is compatible with our rate under their additional assumption (*). In summary, our theorem considers step-function kernels instead of smooth kernels, requires weaker moment assumptions, weaker regularity assumptions and weaker bandwidth restrictions and provides a rate in (3.3) more sensitive to the regularity assumptions.

In particular, given (3.2d) with $\alpha = 1$ (implied by Mack and Silverman's conditions), the optimal rate in (3.3) is $n^{-1/3}$ (ignoring log factors). This, in view of (*), is also the optimal rate attainable in Mack and Silverman's Theorem B. For such a rate, our theorem requires $\lambda \geq 3$ in (3.2b), whereas their theorem requires $\lambda > 3$ and regularity stronger than (3.2d) with $\alpha = 1$.

3.2. Nonparametric scale curve estimation. This may be handled very much like the previous application (see discussion of Example 2 in Section 1), and we shall leave the details implicit.

3.3. The conditional distribution function. With the family $\{\beta_t, t \in I\}$ given by $\beta_t(y) = I\{y \leq t\}$, $y \in \mathbb{R}$, $t \in I = \mathbb{R}$, the quantity $r_t(x)$ defined by (1.1) becomes the conditional df $F(t|x)$. Let us take $K(\cdot)$ and $K_0(\cdot)$ in (1.3b) and (1.3c) to be equal, in which case the quantity $r_{tn}(x)$ in (1.3a) becomes a df (in the variable t), which we shall denote by $F_n(t|x)$, $t \in \mathbb{R}$, for each x . For the present choice of $\beta_t(\cdot)$, representation (2.38) holds trivially and assumptions (A.1)–(A.7), (B.1) and f_0 uLL may be reduced to:

$$(3.4a) \quad F_y, \text{ the marginal df of } Y, \text{ is continuous;}$$

$$(3.4b) \quad 0 < m_1 \leq f_0(x) \leq M_1 < \infty, \quad x \in J;$$

$$(3.4c) \quad f_0(\cdot) \text{ is uLL-}\alpha \text{ on } J, \text{ and the functions } F(t|\cdot), t \in \mathbb{R}, \text{ are uLL-}\alpha \text{ on } J, \text{ uniformly in } t \in \mathbb{R}, \text{ with } 0 < \alpha \leq 1.$$

Thus Theorem 2.3(i) and Corollary 2.2 yield

THEOREM 3.2. Assume (3.4) and let $\{h_n\}$ satisfy (3.1) with $\lambda = \infty$. Then

$$(3.5) \quad \sup_{x \in J} \sup_{t \in \mathbb{R}} |F_n(t|x) - F(t|x)| = O(\max\{\Delta_n^{-1/2}, h_n^\alpha\}) \quad \text{a.s., } n \rightarrow \infty.$$

3.4. L-smoothing. For L -smoothers with trimmed weight functions, uniform strong consistency rates may be obtained by reduction to an application of rates established for conditional df estimators, for example as given in the preceding section. As discussed in Example 5 of Section 1, we consider a conditional L -functional of the form

$$(3.6) \quad l(x) = \int_0^1 J_0(v) F^{-1}(v|x) dv, \quad x \in J.$$

Let $\hat{l}_n(x)$ be a corresponding estimator defined by replacing $F^{-1}(v|x)$ in (3.6) by $\hat{F}_n^{-1}(v|x)$, where $\hat{F}_n(\cdot|x)$ is a df for each x and is uniformly strongly consistent for estimation of $F(\cdot|x)$, in the sense that

$$(3.7) \quad Z_n = \sup_{x \in J} \sup_{t \in \mathbb{R}} |\hat{F}_n(t|x) - F(t|x)| \rightarrow 0 \quad \text{a.s., } n \rightarrow \infty.$$

Assume that $J_0(\cdot)$ satisfies

$$(3.8) \quad J_0(\cdot) \text{ is bounded on } [v_0, v_1] \text{ and vanishes elsewhere, with } 0 < v_0 < v_1 < 1.$$

We shall utilize the following elementary reduction lemma. For any function $g(\cdot)$, let $\|g\|_\infty$ denote $\sup|g(\cdot)|$.

LEMMA 3.1. *Let $J_0(\cdot)$ satisfy (3.8) and let F and G be arbitrary df's. Then*

$$(3.9) \quad \left| \int_0^1 J_0(G^{-1} - F^{-1}) \right| \leq \|J_0\|_\infty [F^{-1}(v_1 + \delta) - F^{-1}(v_0 - \delta)] \delta,$$

where $\delta = \|G - F\|_\infty$ and v_0, v_1 are as in (3.8).

PROOF. Put $H_0(u) = \int_0^u J_0(v) dv$, $0 < u < 1$, and $y_0 = F^{-1}(v_0 - \delta)$, $y_1 = F^{-1}(v_1 + \delta)$. Then, using integration by parts, (3.8) and the inequalities $\max\{F(y), G(y)\} < v_0$ for $y < y_0$, $\min\{F(y), G(y)\} > v_1$ for $y > y_1$, we have

$$\begin{aligned} \int_0^1 J_0(v)[G^{-1}(v) - F^{-1}(v)] dv &= - \int_{-\infty}^{\infty} [H_0(G(y)) - H_0(F(y))] dy \\ &= - \int_{y_0}^{y_1} [H_0(G(y)) - H_0(F(y))] dy. \end{aligned}$$

Now, using $|H_0(u) - H_0(u')| \leq \|J_0\|_\infty |u - u'|$, we obtain (3.9). \square

We now give a general uniform strong convergence result for estimators $\hat{l}_n(x)$ formulated as before. We shall suppose that the given family of conditional df's, $\{F(\cdot|x), x \in J\}$, satisfies

$$(3.10) \quad a_0 < F^{-1}(v_0 - \epsilon_0|x) < F^{-1}(v_1 + \epsilon_0|x) < a_1, \quad \text{all } x \in J,$$

with $-\infty < a_0 < a_1 < \infty$, $\epsilon_0 < \min\{v_0, 1 - v_1\}$ and v_0, v_1 as in (3.8).

THEOREM 3.3. *Let $l(\cdot)$ be defined by (3.6), with $J_0(\cdot)$ satisfying (3.8) and $\{F(\cdot|x), x \in J\}$ satisfying (3.10). Let $\hat{l}_n(\cdot)$ be based on a family $\{\hat{F}_n(\cdot|x), x \in J\}$ satisfying (3.7). Then*

$$(3.11) \quad \sup_{x \in J} |\hat{l}_n(x) - l(x)| = O(Z_n) \quad \text{a.s., } n \rightarrow \infty.$$

PROOF. For each $x \in J$, we apply Lemma 3.1 with F and G given by $F(\cdot|x)$ and $\hat{F}_n(\cdot|x)$, respectively. Combining these results, we obtain

$$(3.12) \quad \sup_{x \in J} |\hat{l}_n(x) - l(x)| \leq \|J_0\|_\infty Z_n \sup_{x \in J} [F^{-1}(v_1 + Z_n|x) - F^{-1}(v_0 - Z_n|x)].$$

By (3.7) and (3.10), the third factor on the right-hand side of (3.12) is a.s. bounded above by $(a_1 - a_0)$ for all large n . Thus (3.11) follows. \square

Let us now consider the special case that $l(x)$ is estimated by $\hat{l}_n(x)$ based on the family $\{F_n(\cdot|x), x \in J\}$ considered in Section 3.3. We have in this case the following explicit rate.

COROLLARY 3.1. Let $l(\cdot)$ be defined by (3.6), with $J_0(\cdot)$ satisfying (3.8) and $\{F(\cdot|x), x \in J\}$ satisfying (3.10). Let $l_n(x)$ be based on the family $\{F_n(\cdot|x), x \in J\}$ considered in Theorem 3.2 and assume the conditions of that theorem. Then

$$\sup_{x \in J} |l_n(x) - l(x)| = O(\max\{\Delta_n^{-1/2}, h_n^\alpha\}) \quad a.s., n \rightarrow \infty.$$

3.5. M-smoothing. Continuing the discussion in Example 6 of Section 1, we establish here, for a fixed $\psi(\cdot)$ function, a uniform strong convergence rate for $r_{\psi n}$. We apply the results of Section 2 by taking $\beta_t(y) = \psi(y - t)$, $y \in \mathbb{R}$, for $t \in I = \mathbb{R}$. In this case (1.2) becomes

$$(3.13) \quad d_t(x) = \int_{\mathbb{R}} \psi(y - t) f(x, y) dy$$

and (1.3c) becomes, for a fixed kernel $K(\cdot)$,

$$(3.14) \quad d_{tn}(x) = (nh_n)^{-1} \sum_{i=1}^n \psi(Y_i - t) K\left(\frac{x - X_i}{h_n}\right).$$

Clearly, we may characterize $r(x)$ and $r_{\psi n}(x)$ as the solutions, with respect to t , of the equations

$$(3.15a) \quad d_t(x) = 0,$$

$$(3.15b) \quad d_{tn}(x) = 0,$$

respectively. Accordingly, we shall reduce the problem of strong convergence of $r_{\psi n}(x)$ to $r(x)$, uniformly in x , to an application of the strong convergence of $d_{tn}(x)$ to $d_t(x)$, uniformly in x and t , as given by Theorem 2.2.

To apply Theorem 2.2, we satisfy the representation (2.38) for $\{\beta_t, t \in \mathbb{R}\}$ by taking $q_1 = q_2 = -1$, $\gamma_{t1}(y) = \max\{0, -\psi(y - t)\}$ and $\gamma_{t2}(y) = \min\{0, -\psi(y - t)\}$ and adopting the following assumptions:

$$(3.16a) \quad \psi(\cdot) \text{ is bounded, antisymmetric, monotone (incr.) and continuous;}$$

$$(3.16b) \quad 0 < m_1 \leq f_0(x) \leq M_1 < \infty, \quad x \in J;$$

$$(3.16c) \quad \text{the conditional densities } f(\cdot|y), y \in \mathbb{R}, \text{ are uLL-}\alpha \text{ on } J, \text{ uniformly in } y \in \mathbb{R}, \text{ with } 0 < \alpha \leq 1.$$

It is readily seen that these yield (A.1)–(A.7), with $\lambda = \infty$ in (A.7), and thus from Theorem 2.2 and Remark 2.3(v) we immediately have

LEMMA 3.2. Let $d_t(\cdot)$ and $d_{tn}(\cdot)$ be given by (3.13) and (3.14). Assume (3.16) and let $\{h_n\}$ satisfy (3.1) with $\lambda = \infty$. Then, for some constant A^* , we have a.s.

$$(3.17) \quad \sup_{t \in \mathbb{R}} \sup_{x \in J} |d_{tn}(x) - d_t(x)| \leq A^* \max\{\Delta_n^{-1/2}, h_n^\alpha\}, \quad \text{all large } n.$$

For our result on $r_{\psi n}(\cdot)$, we shall also require

$$(3.18) \quad \inf_{x \in J} \left| \int \psi(y - r(x) + \varepsilon) dF(y|x) \right| \geq q_0 |\varepsilon|, \quad \text{for } |\varepsilon| \leq \delta,$$

where δ and q_0 are some positive constants. [This assumption is also used by Härdle and Luckhaus (1984); see their discussion.]

THEOREM 3.4. Under the conditions of Lemma 3.2 and also assuming (3.18), we have a.s.

$$(3.19) \quad \sup_{x \in J} |r_{\psi_n}(x) - r(x)| \leq B^* \max\{\Delta_n^{-1/2}, h_n^\alpha\}, \quad \text{all large } n,$$

with $B^* = 2A^*/m_1q_0$.

PROOF. By the monotonicity of ψ and the definition of $r_{\psi_n}(x)$ as solution of (3.15b), we have, for $\varepsilon > 0$,

$$(3.20) \quad r_{\psi_n}(x) > r(x) + \varepsilon \Rightarrow d_{r(x)+\varepsilon, n}(x) > 0.$$

Now

$$(3.21) \quad d_{r(x)+\varepsilon, n}(x) \leq d_{r(x)+\varepsilon}(x) + \sup_{t \in \mathbf{R}} |d_{tn}(x) - d_t(x)|.$$

Also, by monotonicity of $\psi(\cdot)$ and the identity $d_{r(x)}(x) = 0$, the function $d_{r(x)+\varepsilon}(x)$ is nonpositive and by (3.16b) and (3.18) has magnitude $\geq m_1q_0\varepsilon$, for $0 < \varepsilon < \delta$. That is, for $0 < \varepsilon < \delta$,

$$(3.22) \quad d_{r(x)+\varepsilon}(x) \leq -m_1q_0\varepsilon.$$

Combining (3.20), (3.21) and (3.22), we have, for $0 < \varepsilon < \delta$,

$$r_{\psi_n}(x) > r(x) + \varepsilon \Rightarrow \sup_{t \in \mathbf{R}} |d_{tn}(x) - d_t(x)| > m_1q_0\varepsilon.$$

With a similar inequality proved for the case $r_{\psi_n}(x) < r(x) - \varepsilon$, we obtain, for $0 < \varepsilon < \delta$,

$$(3.23) \quad \sup_{x \in J} |r_{\psi_n}(x) - r(x)| > \varepsilon \Rightarrow \sup_{r \in \mathbf{R}} \sup_{x \in J} |d_{tn}(x) - d_t(x)| > m_1q_0\varepsilon.$$

It readily follows that (3.23) and (3.17) imply (3.19). \square

4. Proof of Lemma 2.2. Put

$$(4.1) \quad a_n = \Delta_n^{-1/2} c_n = n^{-1/2} (c_n \log n)^{1/2}.$$

We first reduce $\sup_{t \in I}$ in (2.7) to a maximum over a finite set. By (A.3), (A.5)–(A.7) and the monotone convergence theorem, the function $g(t) = E\gamma_t(Y)$ is nondecreasing and continuous in t with finite limits $g(t_*)$ and $g(t^*)$ as $t \rightarrow t_*$ and t^* . Let us partition I by finite points $t_1 < t_2 < \dots < t_{N_n}$ such that $g(t_1) - g(t_*) \leq a_n$, $g(t^*) - g(t_{N_n}) \leq a_n$ and $g(t_j) - g(t_{j-1}) \leq a_n$ for $2 \leq j \leq N_n$. Clearly, we may arrange that

$$(4.2) \quad N_n \leq 2(g(t^*) - g(t_*))/a_n.$$

Also, for fixed x and z , the functions $G_{tn}(x+z) - G_{tn}(x)$ and $G_t(x+z) - G_t(x)$ are monotone in t and, by (A.3) and (A.6) a.s., these functions for all x and z have finite limits as $t \rightarrow t_*$, t^* .

Letting I_n denote the set $\{t_*, t_1, \dots, t_{N_n}, t^*\}$ and I_n^* the set $\{(t_*, t_1), (t_1, t_2), \dots, (t_{N_n}, t^*)\}$, we therefore have, for arbitrary $t \in I$,

$$(4.3) \quad \begin{aligned} & |G_{tn}(x+z) - G_{tn}(x) - [G_t(x+z) - G_t(x)]| \\ & \leq \max_{t \in I_n} |G_{tn}(x+z) - G_{tn}(x) - [G_t(x+z) - G_t(x)]| \\ & \quad + \max_{(s, t) \in I_n^*} |G_t(x+z) - G_t(x) - [G_s(x+z) - G_s(x)]|. \end{aligned}$$

Now, by nonnegativity of $\{\gamma_t, t \in I\}$ [by (A.3)], for $s < t$, the function $G_t(x) - G_s(x)$ is nonnegative and nondecreasing in x , so that for $(s, t) \in I_n^*$ we have

$$(4.4) \quad |G_t(x) - G_s(x)| \leq G_t(\infty) - G_s(\infty) = g(t) - g(s) \leq a_n, \quad \text{all } x.$$

It follows from (4.3) and (4.4) that

$$(4.5) \quad V_n \leq \max_{t \in I_n} \sup_{x \in J} V_{tn}(x, c_n) + 2a_n.$$

Next we reduce $\sup_{x \in J}$ to a maximum over a finite set. In this case, we first transform to a supremum over a finite interval, as follows. Define

$$\tilde{G}_{tn}(v) = n^{-1} \sum_{i=1}^n \gamma_t(Y_i) I\{F_0(X_i) \leq v\}$$

and $\tilde{G}_t(v) = E\tilde{G}_{tn}(v)$, $0 \leq v \leq 1$, where F_0 denotes the (continuous) df of X . Then $G_{tn}(x) = \tilde{G}_{tn}(F_0(x))$ a.s., $G_t(x) = \tilde{G}_t(F_0(x))$ and, by (A.2), $|F_0(x+z) - F_0(x)| \leq M_1|z|$. Define

$$\tilde{V}_{tn}(v, \delta) = \sup_{|u| \leq \delta} |\tilde{G}_{tn}(v+u) - \tilde{G}_{tn}(v) - [\tilde{G}_t(v+u) - \tilde{G}_t(v)]|.$$

Then $V_{tn}(x, c_n) \leq \tilde{V}_{tn}(F_0(x), M_1c_n)$ a.s. and hence (4.5) yields

$$(4.6) \quad V_n \leq \max_{t \in I_n} \sup_{v \in F_0(J)} \tilde{V}_{tn}(v, M_1c_n) + 2a_n \quad \text{a.s.}$$

We now partition the interval $[0, 1]$ by $v_0 = 0$ and $v_k = k[2/M_1c_n]^{-1}$ for $1 \leq k \leq [2/M_1c_n]$, where $[\cdot]$ denotes greatest integer part. Let $|u| \leq M_1c_n$. For v and $v+u$ in the same subinterval $[v_k, v_{k+1}]$, we easily find

$$|\tilde{G}_{tn}(v+u) - \tilde{G}_{tn}(v) - [\tilde{G}_t(v+u) - \tilde{G}_t(v)]| \leq 2\tilde{V}_{tn}(v_k, M_1c_n),$$

and for v and $v+u$ in $[v_k, v_{k+1}]$ and $[v_j, v_{j+1}]$, respectively, with $k < j$, we find

$$\begin{aligned} & |\tilde{G}_{tn}(v+u) - \tilde{G}_{tn}(v) - [\tilde{G}_t(v+u) - \tilde{G}_t(v)]| \\ & \leq \tilde{V}_{tn}(v_j, M_1c_n) + 2\tilde{V}_{tn}(v_{k+1}, M_1c_n). \end{aligned}$$

It follows that

$$(4.7) \quad V_n \leq 3 \max_{t \in I_n} \max_{v \in \tilde{J}_n} \tilde{V}_{tn}(v, M_1c_n) + 2a_n \quad \text{a.s.,}$$

with $\tilde{J}_n = \{0, [2/M_1c_n]^{-1}, 2[2/M_1c_n]^{-1}, \dots, 1\}$.

In order to set the stage for an application of Bernstein's inequality, we now replace $\tilde{V}_{tn}(v, M_1c_n)$ by an analogue given by replacing \tilde{G}_{tn} and \tilde{G}_t by analogues based on truncation of $\{\gamma_t(Y_i)\}$. Put

$$(4.8) \quad Q_n = M_\lambda a_n^{-1/(\lambda-1)}, \quad n \geq 1,$$

and

$$H_{tn}(v) = n^{-1} \sum_{i=1}^n \gamma_t(Y_i) I\{\gamma_t(Y_i) \leq Q_n\} I\{F_0(X_i) \leq v\},$$

define $V_{tn}^*(v, \delta)$ by substitution of H_{tn} for \tilde{G}_{tn} and EH_{tn} for \tilde{G}_t in the definition of $\tilde{V}_{tn}(v, \delta)$ and define

$$V_n^* = \max_{t \in I_n} \max_{v \in \tilde{J}_n} V_{tn}^*(v, M_1c_n).$$

Then (4.7) yields

$$(4.9) \quad V_n \leq 3V_n^* + 3a_n(2/3 + W_n + \theta_n) \quad \text{a.s.},$$

where

$$W_n = a_n^{-1} \sup_{t \in I_n} \sup_{v \in J_n} \sup_{|u| \leq M_1 c_n} |\tilde{G}_{tn}(v+u) - \tilde{G}_{tn}(v) - [H_{tn}(v+u) - H_{tn}(v)]|$$

and

$$\theta_n = a_n^{-1} \sup_{t \in I_n} \sup_{v \in J_n} \sup_{|u| \leq M_1 c_n} |\tilde{G}_t(v+u) - \tilde{G}_t(v) - [EH_{tn}(v+u) - EH_{tn}(v)]|.$$

Note that $W_n \equiv 0$ and $\theta_n \equiv 0$ in the case $\lambda = \infty$.

Using monotonicity of γ_t in t [by (A.3)] and (A.6) and noting that $a_n^{-1} = (Q_n/M_\lambda)^{\lambda-1}$, we readily obtain

$$(4.10) \quad \begin{aligned} M_\lambda^{\lambda-1} W_n &\leq Q_n^{\lambda-1} n^{-1} \sum_{i=1}^n \gamma_{t^*}(Y_i) I\{\gamma_{t^*}(Y_i) > Q_n\} \\ &\leq n^{-1} \sum_{i=1}^n [\gamma_{t^*}(Y_i)]^\lambda I\{\gamma_{t^*}(Y_i) > Q_n\}. \end{aligned}$$

For fixed Q , we have by (A.7) and the classical SLLN,

$$(4.11) \quad n^{-1} \sum_{i=1}^n [\gamma_{t^*}(Y_i)]^\lambda I\{\gamma_{t^*}(Y_i) > Q\} \rightarrow E\{[\gamma_{t^*}(Y)]^\lambda I\{\gamma_{t^*}(Y) > Q\}\} \quad \text{a.s.}$$

Thus, since the right-hand side of (4.11) a.s. dominates $\limsup_n W_n$ and $\rightarrow 0$ as $Q \rightarrow \infty$, we have

$$(4.12) \quad W_n \rightarrow 0 \quad \text{a.s., } n \rightarrow \infty.$$

Also, we see via (4.10) that

$$(4.13) \quad \theta_n \leq EW_n \rightarrow 0, \quad n \rightarrow \infty.$$

By (4.9), (4.12) and (4.13), it suffices for (2.8a) to show

$$(4.14) \quad V_n^* = O(a_n) \quad \text{a.s., } n \rightarrow \infty.$$

We shall establish this and (2.8b) as well, by developing a suitable upper bound for $P\{V_n^* > B_0 a_n\}$, for appropriate choice of B_0 . We write

$$(4.15) \quad P\{V_n^* \geq B_0 a_n\} \leq \sum_{t \in I_n} \sum_{v \in J_n} P\{V_{tn}^*(v, M_1 c_n) \geq B_0 a_n\},$$

with B_0 to be specified later, and we estimate the terms of this summation by an adaptation of the proof of Lemma 2.2 of Serfling (1982).

By (4.1) it is seen that

$$(4.16) \quad a_n/c_n \rightarrow 0, \quad n \rightarrow \infty.$$

Now define

$$w_n = \left[\frac{2Q_n c_n}{a_n} + 1 \right],$$

with $[\cdot]$ denoting greatest integer part. Fix v and put

$$\eta_{n,r} = v + \frac{rM_1c_n}{w_n}, \text{ for } r = -w_n, -w_n + 1, \dots, w_n.$$

Note that $\eta_{n,r+1} - \eta_{n,r} = M_1c_n/w_n$. Defining

$$\xi_{tnr} = |H_{tn}(\eta_{n,r}) - H_{tn}(v) - [EH_{tn}(\eta_{n,r}) - EH_{tn}(v)]|,$$

we have, by monotonicity of $H_{tn}(v)$ and $EH_{tn}(v)$ as functions of v , that

$$V_{tn}^*(v, M_1c_n) \leq \max_{-w_n \leq r \leq w_n} \xi_{tnr} + \max_{-w_n \leq r \leq w_n-1} |EH_{tn}(\eta_{n,r+1}) - EH_{tn}(\eta_{n,r})|.$$

Now

$$\begin{aligned} E [H_{tn}(\eta_{n,r+1}) - H_{tn}(\eta_{n,r})] &\leq Q_n P\{\eta_{n,r} < F_0(X) \leq \eta_{n,r+1}\} \\ &= Q_n(\eta_{n,r+1} - \eta_{n,r}) \\ &\leq M_1Q_n c_n/w_n \\ &\leq M_1a_n/2 \leq M_1a_n, \end{aligned}$$

so that

$$\begin{aligned} P\{V_{tn}^*(v, M_1c_n) \geq B_0a_n\} &\leq P\left\{\max_{-w_n \leq r \leq w_n} \xi_{tnr} \geq (B_0 - M_1)a_n\right\} \\ &\leq \sum_{r=-w_n}^{w_n} P\{\xi_{tnr} \geq (B_0 - M_1)a_n\}. \end{aligned}$$

By Bernstein's inequality [Uspensky (1937)],

$$P\{\xi_{tnr} \geq (B_0 - M_1)a_n\} \leq 2 \exp(-\delta_{n,r}),$$

where

$$\delta_{n,r} = \frac{(B_0 - M_1)^2 n^2 a_n^2}{2n\sigma_{tnr}^2 + \frac{2}{3}(B_0 - M_1)Q_n n a_n}$$

and $\sigma_{tnr}^2 = \text{Var}\{Z_{tnr}\}$, with

$$Z_{tnr} = \gamma_t(Y)I\{\gamma_t(Y) \leq Q_n\}I\{v < F_0(X) \leq \eta_{n,r}\}.$$

Applying (A.1), we obtain

$$\begin{aligned} \sigma_{tnr}^2 &\leq EZ_{tnr}^2 \\ (4.17) \quad &\leq \int \int \gamma_t^2(y)I\{\gamma_t(y) \leq Q_n\}I\{v < F_0(x) \leq v + M_1c_n\}f(x, y) dx dy \\ &\leq M_0M_1c_n. \end{aligned}$$

By (4.8) and restriction (iii) on $\{c_n\}$ in the hypothesis of the lemma, we obtain

$$(4.18) \quad Q_n a_n = M_\lambda a_n^{(\lambda-2)/(\lambda-1)} = M_\lambda \left(\frac{c_n \log n}{n}\right)^{(\lambda-2)/2(\lambda-1)} \leq M_\lambda c_n.$$

By (4.1), (4.17) and (4.18), we thus have

$$(4.19) \quad \delta_{n,r} \geq B_0^* \log n,$$

with

$$(4.20) \quad B_0^* = \frac{(B_0 - M_1)^2}{2M_0M_1 + \frac{2}{3}(B_0 - M_1)M_\lambda}.$$

Since (4.19) holds uniformly in r , $r = -w_n, -w_n + 1, \dots, w_n$, we have

$$(4.21) \quad P\{V_{tn}^*(v, M_1c_n) \geq B_0a_n\} \leq 6w_n n^{-B_0^*}.$$

And since (4.21) holds uniformly in $t \in I_n$ and $v \in \tilde{J}_n$, (4.15) yields

$$(4.22) \quad P\{V_n^* \geq B_0a_n\} \leq 6(N_n + 2)\tilde{N}_n w_n n^{-B_0^*},$$

where \tilde{N}_n denotes the cardinality of the set \tilde{J}_n . By (4.1) and (4.2) we find

$$(4.23a) \quad N_n + 2 \leq 2(M_\lambda + 1) \left(\frac{n}{c_n \log n} \right)^{1/2}.$$

Also, using the restriction $c_n \leq 1$,

$$(4.23b) \quad \tilde{N}_n \leq [2/M_1 c_n] + 1 \leq (2/M_1 + 1)c_n^{-1}.$$

By (4.1) and (4.8), we have

$$w_n \leq \frac{2Q_n c_n}{a_n} + 1 = 2M_\lambda c_n \left(\frac{n}{c_n \log n} \right)^{\lambda/2(\lambda-1)} + 1.$$

Using the restriction (iii) on $\{c_n\}$, we easily obtain

$$c_n \left(\frac{n}{c_n \log n} \right)^{\lambda/2(\lambda-1)} \geq c_n^{-2/(\lambda-2)} \geq 1.$$

Thus

$$(4.23c) \quad w_n \leq (2M_\lambda + 1)c_n \left(\frac{n}{c_n \log n} \right)^{\lambda/2(\lambda-1)}.$$

Putting

$$(4.24) \quad L_\lambda = 12(M_\lambda + 1)(2/M_1 + 1)(2M_\lambda + 1)$$

and combining (4.22) and (4.23), we obtain

$$(4.25) \quad P\{V_n^* \geq B_0a_n\} \leq L_\lambda \left(\frac{n}{c_n \log n} \right)^{(2\lambda-1)/2(\lambda-1)} n^{-B_0^*}.$$

Again using the restriction (iii) on $\{c_n\}$, we find $c_n^{-1} \leq (n/\log n)^{\lambda/(\lambda-2)}$, whence (4.25) yields

$$(4.26) \quad P\{V_n^* \geq B_0a_n\} \leq L_\lambda \left(\frac{n}{\log n} \right)^{(2\lambda-1)/(\lambda-2)} n^{-B_0^*}.$$

Now, for given λ and for given real $\kappa > 0$, the constant B_0^* can be made to satisfy

$$B_0^* \geq \kappa + \frac{2\lambda - 1}{\lambda - 2}$$

by taking B_0 sufficiently large in (4.20). Let $B_{\kappa, \lambda}$ denote such a determination of B_0 . Then (4.26) yields [using $(2\lambda - 1)/(\lambda - 2) \geq 2$ for $\lambda > 2$],

$$(4.27) \quad P\{V_n^* \geq B_{\kappa, \lambda} a_n\} \leq L_\lambda (\log n)^{-2} n^{-\kappa}.$$

In particular, taking $\kappa = 2$ in (4.27) and applying the Borel-Cantelli lemma, we obtain (4.14), thus establishing (2.8a).

To obtain (2.8b), we take $\lambda = \infty$ and apply (4.9) with $W_n = 0$ and $\theta_n = 0$ to write, for any $B \geq 2$,

$$(4.28) \quad P\{V_n \geq B a_n\} \leq P\{V_n^* \geq \frac{1}{3}(B - 2)a_n\}.$$

For each real $\kappa > 0$, define $B_\kappa = 3B_{\kappa, \infty} + 2$. Then (4.27) and (4.28) yield

$$(4.29) \quad P\{V_n \geq B_\kappa a_n\} \leq L_\infty (\log n)^{-2} n^{-\kappa},$$

which, recalling the definition (4.1) of a_n , yields (2.8b). \square

Acknowledgment. The authors thank one of the referees for the suggestions leading to Theorem 2.3(ii) and all referees for careful reading of the manuscript.

REFERENCES

- CARROLL, R. (1982). Adapting for heteroscedasticity in linear models. *Ann. Statist.* **10** 1224-1233.
- COLLOMB, G. (1980). Estimation non paramétrique de probabilités conditionnelles. *C. R. Acad. Sci. Paris Sér. A-B* **291** 427-430.
- COLLOMB, G. (1981). Estimation non-paramétrique de la régression: Reveu bibliographique. *Internat. Statist. Rev.* **49** 75-93.
- HÄRDLE, W. (1984). Robust regression function estimation. *J. Multivariate Anal.* **14** 169-180.
- HÄRDLE, W. and LUCKHAUS, S. (1984). Uniform consistency of a class of regression function estimators. *Ann. Statist.* **12** 612-623.
- MACK, Y. P. and SILVERMAN, B. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. verw. Gebiete* **61** 405-415.
- NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141-142.
- SERFLING, R. J. (1982). Properties and applications of metrics on nonparametric density estimators. In *Nonparametric Statistical Inference* (B. V. Gnedenko, M. L. Puri and I. Vincze, eds.) 859-873. North-Holland, Amsterdam.
- STONE, C. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5** 595-645.
- STONE, C. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040-1053.
- STUTE, W. (1986). On almost sure convergence of conditional empirical distribution functions. *Ann. Statist.* **14** 891-901.
- TSYBAKOV, E. A. (1983). Robust estimates of a function. *Problems Inform. Transmission* **18** 39-52.
- USPENSKY, J. V. (1937). *Introduction to Mathematical Probability*. McGraw-Hill, New York.
- WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 359-372.

INSTITUT FÜR GESELLSCHAFTS- UND
WIRTSCHAFTSWISSENSCHAFTEN
UNIVERSITÄT BONN
ADENAUERALLEE 24-26
D-5300 BONN 1
WEST GERMANY

LIMBURGS UNIVERSITAIR CENTRUM
UNIVERSITAIRE CAMPUS
B-3610 DIEPENBEEK
BELGIUM

DEPARTMENT OF MATHEMATICAL SCIENCES
THE JOHNS HOPKINS UNIVERSITY
BALTIMORE, MARYLAND 21218

STATISTICAL METHODS FOR DEVELOPING AND DISTINGUISHING
MULTINOMINAL RESPONSE MODELS IN THE TRAUMATOLOGICAL
ANALYSIS OF SIMULATED AUTOMOBILE IMPACTS

W. Härdle

Rheinische Friedrich-Wilhelms-Universität Bonn
D-5300 Bonn, Federal Republic of Germany

D. Kallieris

Ruprecht-Karls-Universität Heidelberg
D-6900 Heidelberg, Federal Republik of Germany

R. Mattern

Johannes Gutenberg-Universität Mainz
D-6500 Mainz, Federal Republic of Germany

ABSTRACT

Simulated car-to-car side impacts, designed for the analysis of traumatological aspects, involve two sets of variables. Predictors include exogenous biomechanical factors as well as anthropometric variables, such as age. The response is measured a scale of injury scores and is thus multinomial.

It is the aim of a statistical analysis of such data to devise a multinomial response model that explains possible patterns of injury as a function of a suitable set of predictor variables. Several approaches for modelling such a multinomial response relationship have been proposed in the literature, among them the Logistic and the Weibull regression models. Two major questions in applying such models are as follows: What model is appropriate and how should different models be compared. Another concern is how the quality of a given model should be presented for varying sets of predictors.

In this paper we discuss the first question by constructing a goodness-of-fit test based on bootstrapping flexible, non-parametric alternatives to a given parametric candidate model. Secondly, we present several graphical techniques that allow relatively simple comparisons of different models.

1. Modelling the influence of anthropometric and mechanical parameters on trauma indices:

The aim of the statistical analysis of simulated car impacts is to develop models that allow one to understand how the severity of impacts depend on observable input variables. Typically such input variables can be divided into two types. The first set of variables is describing

the test subject's physical characteristics, such as height or age. A second set is concerned with the actual experimental setting, and contains such parameters as velocity of the impact and acceleration measured at various places. These input variables determine jointly the response variable. The observed response variable is a trauma index usually scaled according to some injury scale, e.g. AIS (1980). The AIS trauma index, for example, is a discrete variable in $\{0,1,2,3,4,5,6\}$, with the lightest (or non) injury indexed by "0" and the severest injury indexed by "6". The input variables are mostly of continuous nature, i.e. they can possibly take each value in a certain interval.

Phrased in terms of statistical methodology we are given a discrete regression problem, i.e. discrete response variables (trauma index) are regressed on various kinds of predictor variables (possibly continuous or also discrete). (See Bickel and Doksum (1977), Neter and Wasserman (1974, Chapter 9)). The aim of this statistical problem is to construct suitable models for explaining the probability of a certain level of trauma index as a function of the given covariables. In this paper we denote by (X_i, Y_i) , $i = 1, \dots, n$, the data points from such an experiment; X standing for the vector of predictor variables (input) and Y denoting the discrete response variable (output vector). Since the response variable is multinomial (i.e. takes values in a discrete ordered set) it is reasonable to define the regression function as the probability that Y is bigger than some value c . Hence, we are dealing with a set of regression functions

$$p_c(x) = P(Y \geq c | X=x).$$

where c runs through the discrete set of possible response values (trauma indices). In determining such functions p one would like to have some basis requirements fulfilled that are direct consequences of the experimental setup. These are

(1.1) Monotonicity, i.e. if the input variables are ordered in some natural way then increasing the strength of impact or increasing age, the probability of having a trauma index greater than or equal to c should also increase.

(1.2) Consistency, i.e. $p_{c_1} \geq p_{c_2}$ for $c_1 \leq c_2$

Consistency means that the curves p_c should be so that the probability of having trauma index greater than c increases if c decreases.

In the next section we discuss several multinomial response models. In section 3 we show how nonparametric smoothing techniques help in selecting a suitable response model. In section 4 we discuss some graphical methods for enhancing the summary statistics of a given fit when the set of predictor variables is varied. In section 5 the application of these methods to the Heidelberg side impact data is presented. Section 6 is devoted to conclusions.

2. Multinomial Response Models

There are two different approaches to model the dependence of the conditional probability $p_c(x) = P(Y_c | X=x)$ as a function of the covariables x . The first approach is to assume that this function p_c is a member of a specific class of parameterized functions. The second approach is called non-parametric since the form of p_c is not restricted by any requirement except those of (1.1) and (1.2) above. The parametric approach has the advantage of easier interpretation of coefficients and also of numerical computations, whereas the non-parametric approach has the advantage of not being bound to any functional form. Both should serve each other as an alternative and should not be seen as mutually exclusive models. Well-known parametric models include the Logistic and the Probit regression models. The basic structural assumption for both approaches is the same; both are models based on linear combinations (projections) of the predictor variable x , i.e. the function p_c is modelled as

$$p_c(x) = G_c(\beta^T x).$$

with a link function G_c and parameter β . The parametric approach consists of fixing the function $G_c(\cdot) = G_c(\alpha_c + \cdot)$ to a certain shape whereas the non-parametric approach does not prescribe the form of G_c . In the following we just write G to describe the general form of G_c .

In a Logit analysis one assumes that G is of the form of a logistic distribution function, i.e.

$$G(z) = \exp(z)/(1+\exp(z)).$$

The functions p_c are determined by the maximum likelihood method, i.e. one maximizes for each c

$$\begin{aligned} & \prod_{i=1}^n P(Y_i \geq c | X_i = x_i) \\ &= \prod_{i=1}^n G(\alpha_c + \beta^T x_i)^{Y_i^c} (1 - G(\alpha_c + \beta^T x_i))^{(1 - Y_i^c)}. \\ & \quad Y_i^c = I(Y_i \geq c). \end{aligned}$$

subject to the consistency condition. In the same way other models like the Probit model with G equal to the standard normal distribution function can be adapted. Yet another shape function is the Weibull distribution function.

The non-parametric approach does not fix the shape function G , but rather lets it be any smooth function following the requirements (1.1) and (1.2). Given the parameter vector β the link function G is determined by a non-parametric smoothing technique, such as spline or kernel, see Härdle (1988). The kernel smoother $\hat{G}_h(z)$ at the point

$$z = \beta^T x \text{ for data } (Z_i = \beta^T X_i, Y_i)$$

is defined by

$$\hat{G}_h(z) = n^{-1} \sum_{i=1}^n K_h(z - Z_i) Y_i / n^{-1} \sum_{i=1}^n K_h(z - Z_i)$$

where $K_h(u) = h^{-1}K(u/h)$ is a delta function sequence with bandwidth h and kernel K , where K is a continuous probability density. The kernel smoother is a consistent estimate of G if $h \rightarrow 0$ as the sample size n tends to infinity. The parameter β can be determined in various numerical ways, since the function G is not determined up to scale. One of the possibilities is to determine G and β jointly by minimizing the Residual Sum of Squares (RSS) or other measures of accuracy. This amounts to finding G and β such that

$$n^{-1} \sum_{i=1}^n (Y_i - G(\beta^T X_i))^2$$

is minimal. This minimization is done iteratively by searching over all possible directions β , that is why this method is called Projection Pursuit Regression (PPR), see Friedman and Stuetzle (1981). Another method is called Average Derivative Estimation (ADE). In ADE estimates of β are obtained in a direct way without involving the link function G . This estimate of β is defined as

$$\hat{\beta} = n^{-1} \sum_{i=1}^n Y_i \hat{f}'(X_i) / \hat{f}(X_i)$$

where \hat{f} denotes an estimate of the partial derivatives of f , the density of X . For details see Härdle and Stoker (1988).

3. Selecting a suitable model

The task finding a suitable model among the many possible parametric and non-parametric alternatives involves the statistical precision of the model as well as the numerical applicability. It is widely known that the Logistic regression model can be quite easily fitted numerically, SAS Supplementary User's Guide (1985). Other link functions G, for example the Probit curve have a similar shape (see Berkson, 1951) but require more computational effort. Also the non-parametric smoothing method requires a lot more on computations but has the advantage of not being restricted in its functional form. In particular the symmetry of the link function that is inherent to the Logit model is no restriction for the non-parametric approach. Indeed the response of the side impact experiments is somewhat asymmetric, as was pointed out by several people who tried a skewed Weibull distribution as a link function G. The price one has to pay though for this additional feature is that the number of parameters, and thus the numerical cost and precision of the algorithm, increase.

Since the non-parametric alternative allows fitting in a much wider class of functions it seems reasonable that it can be used in a formal test of goodness of fit of low dimensional parametric models. To simplify matters let us consider only a binominal response model of one dimensional X variables, i.e. Y takes the values 0 or 1. the proposed test is based on smoothing the response variables of a given parametric fit $p(x; \hat{\beta})$. One defines the kernel smoother on data (X_i, Y_i) as

$$\hat{p}(X_j) = n^{-1} \sum_{i=1}^n K_h(X_j - X_i) Y_i / n^{-1} \sum_{i=1}^n K_h(X_j - X_i).$$

The smoothing parameter h can be determined by crossvalidation, see Härdle (1988). The test is described formally as follows.

1. Fit a candidate parametric model $p(x; \hat{\beta})$
2. Simulate new observations (X_i^*, Y_i^*) from this model by using a pseudo random number generator¹ based on $p(x; \hat{\beta})$ (bootstrapping).
3. Determine for each X_i that has been observed the empirical 5 % quantiles of a kernel smoother of the simulated data.
4. Center these 5 % bands around the assumed parametric candidate model.
5. Check whether the kernel smoother based on the original data lies in between these bands.

Figure 3.1

Another method is based on comparing the likelihood for different models with a bias correction for different number of parameters. This is related to ideas of Akaike (1977) and works as follows. One compares the Log-Likelihoods under both models, i.e.

$$n L_1(\hat{\beta}_1) - n L_2(\hat{\beta}_2) - (\dim(\text{model } 1) - \dim(\text{model } 2)).$$

Based on the limiting chi-square distribution of twice the likelihood ratio statistic one cannot distinguish the two models if the magnitude of the above difference is less than 0.5.

4. Comparing similar models

If the above models are run for several types and sets of input variables it is important to compare the output of the different fits. In the study of the Heidelberg data we found the following, mostly graphically oriented tools very convenient.

Concomitant pairs

Concomitant pairs are defined through all pairs of observations with different response values. Now count all pairs of observations where the current model fit predicted a higher probability for the higher Y-value. Then compute the share of these pairs among all pairs with different Y-values. Certainly if this share of concomitant pairs is close to 1 the model fits quite well. The procedure LOGIST of the SAS system computes this number on request.

Prediction Table

The prediction table is simply a frequency table of the observed trauma indices versus the predicted trauma indices. The number of correctly predicted response variables is the classification rate. This number lies between 0 and 1. Certainly a number close to one is desirable. It is quite intuitive that the empirically determined classification rates are over optimistic since the data is used to determine the model as well as to judge it. An unbiased estimate of the classification rate can be obtained by, for example, cross validation. In this method the whole analysis is performed n times on n subsamples each of size $n-1$ (leave one out method). The left out observation is predicted by the model constructed from the rest of the observations. This leads to an unbiased estimate of the prediction error, as was shown by Stone (1974).

NONPARAMETRIC LOGISTIC REGRESSION BOOTSTRAP
NSIM = 500

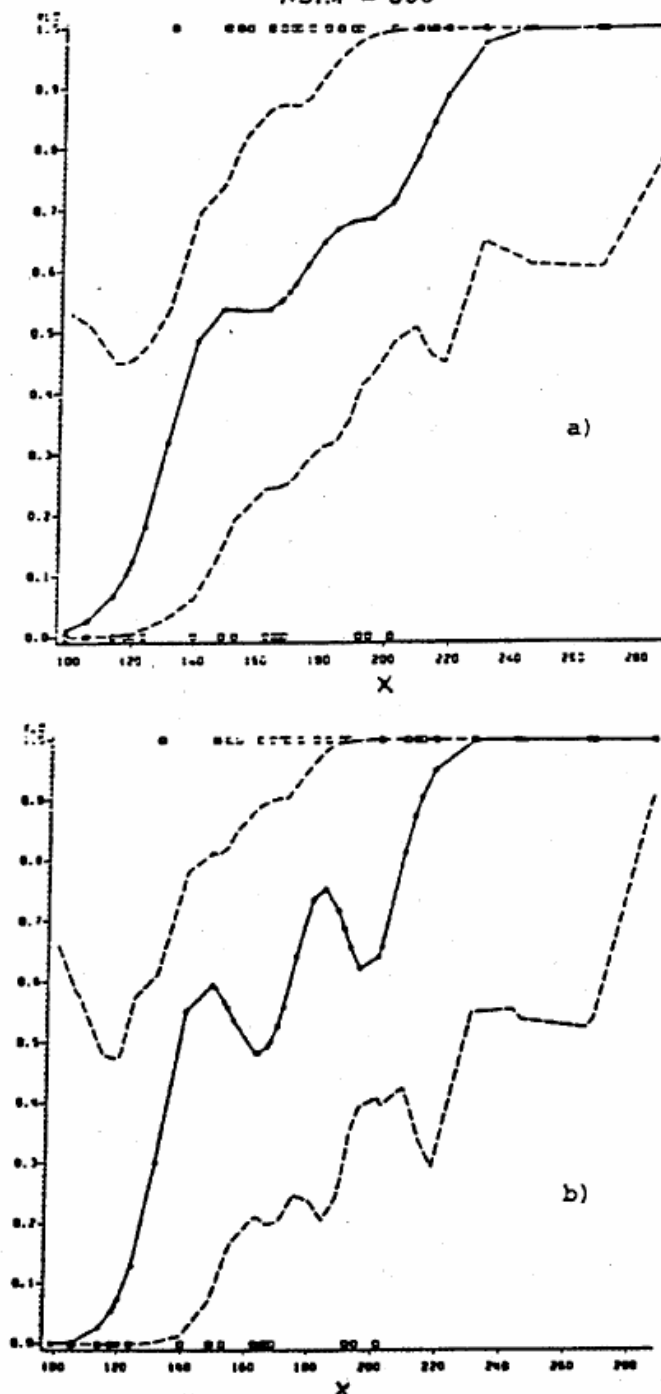


Fig. 3.1 Nonparametric logistic distribution of the injury severity ($y = 1$ for $AIS > 3$ and $Y = 0$ for $AIS \leq 3$) over the TTI with 5% confidencebands for 500 simulations according to the bootstrap method.
a) bandwidth $h = 13$
b) bandwidth $h = 9$

The enhanced histogram of prediction errors

This is a histogram of the observed differences between the observed trauma index and the predicted index where large indices are marked in a special way. The procedure is as follows. 1. Compute all the differences predicted response - observed response. 2. Index all large trauma values (for the AIS values (predicted or observed) greater than 4. 3. Draw a histogram of these differences where the big trauma indices get marked by using special symbol.

In figure 4.1 we show an enhanced histogram for the TTI (Eppinger et al., 1984) as a predictor variable for the TOAIS (thorax AIS).

Figure 4.1

This Thoracic Trauma Index is defined through

$$TTI = 1.4 \text{ AGE} + 0.5 \text{ FORCE.}$$

One sees from this enhanced histogram of prediction errors that the TTI leans toward over estimating the true responses. Indeed, the histogram is skewed to the right. There are 11 observations involving the thorax AIS value of 4. Two of these eleven observations have prediction error zero. One observation has been predicted to have AIS value 4, but really had value 3 (prediction error 2 to the right in the histogram), and eight observations had AIS value 4 but were wrongly classified as 3. One should therefore search for a model that more faithfully predicts the high AIS values.

A distortion measure

As a measure of distortion of current fit we would like to propose two subintegrals of the above histogram. This pair of numbers tells first whether the fit is skew, i.e. has a bias towards over- or underestimating the true response value. Secondly the size of the subintegrals relative to the sample size immediately gives a goodness of fit criterion. The first subintegral just counts the number of positive exceedances (to the right of the column zero in figure 4.1). The second subintegral counts the number of negative exceedances, in this case -8. This together gives the distortion measure (-8, 35) which describes in a very condensed form the skewness of the prediction and how much the true values are missed by the above model.

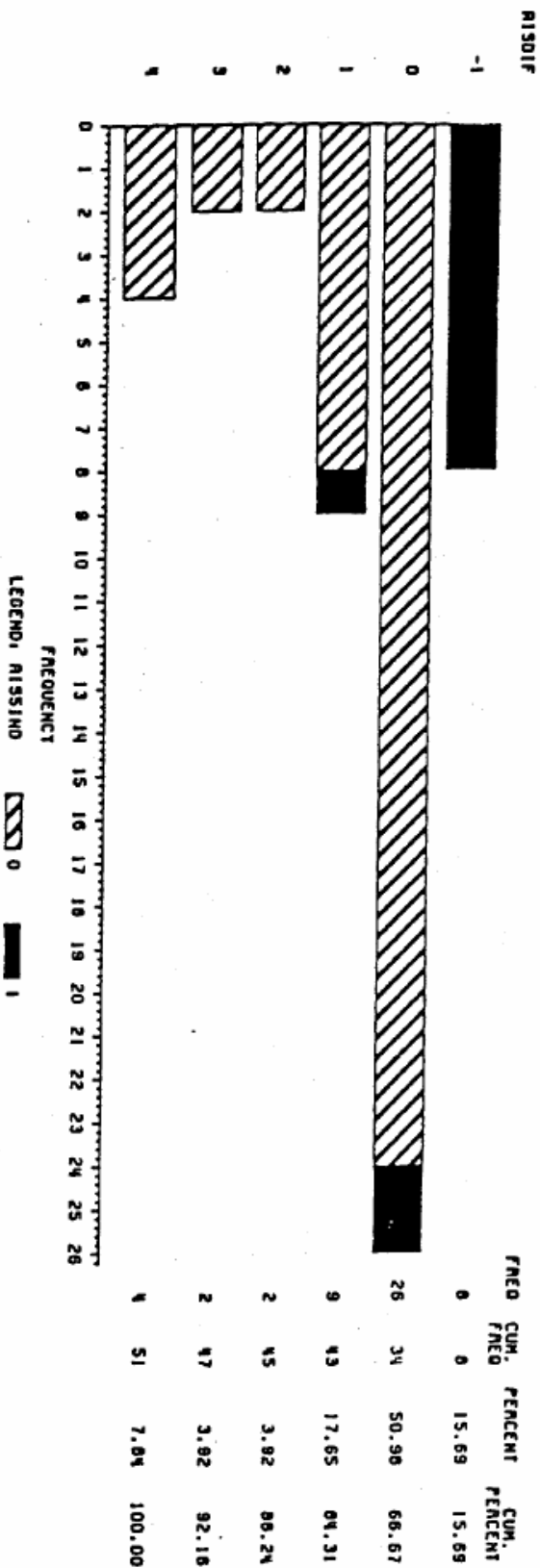


Fig. 4.1 Logistic prediction of the body-AIS (TMAIS) from TTI. Enhanced AIS-difference-histogram between the observed predicted TMAIS. Marked black: body-AIS-predictions, in which one TMAIS 5 is involved.

The Isoquants

The plot of isoquants is designed for two dimensional predictor variables and shows in a graphical way what trauma indices are to be expected given all possible combinations of covariables. In figure 4.2 we show the predicted thorax AIS classes as a function of AGE and FORCE, as defined in Kallieris, Mattern and Härdle (1986).

Figure 4.2

The region indicated by the letter A would be the region of (AGE, FORCE) combinations where AIS = 0 would be predicted. The region with AIS = 3 is shown by D and the highest AIS value of 4 is marked by an E. Overlaid in this plane are the original data values (0,1,2,3,4). This plot allows simple comparison of different fits by simply studying the regions that determine the AIS values. Given for instance the age of 30 one can easily determine by raising the values of FORCE at what points of FORCE the prediction to higher AIS classes would happen. (FORCE level 140 jump to predicted AIS 3, FORCE level 250 jump to predicted AIS 4).

5. Application to the Heidelberg data

Only a few research onsets are suited to determine the connection between mechanical influence and injury severity when measured in AIS degrees. There are real accident analyses on one hand and crash tests with post mortem human subjects (PMHS) on the other hand. Both research onsets are not ideal. The advantage of crash tests with PMHS is, e.g., that by defined conditions of the accident severity, loads acting on the body can be measured in physical magnitudes like acceleration at ribs, sternum, vertebral bodies and head. This is not possible in the real accident analyses. Differences of the injury limits against the living human beings are criticized as a disadvantage of the crash tests with PMHS. The load values measured on the bodies of the PMHS however, are indispensable basis data for the construction of dummies, if these dummies should be qualified for the injury prediction in crash tests.

At the Institute for Legal Medicine of the University of Heidelberg crash tests were conducted with PMHS and dummies for many years to investigate this research concept. As follows, the investigation of lateral collisions should represent which connections exist between loading parameters at the body of the PMHS, anthropometric data and injury severity and how these connections can be used for injury prediction by utilization of the statistical methods described above. Basis of the connection analyses are 58 90-degree lateral collisions. In these collisions PMHS have been loaded in near side position in the impacted/standing vehicle.

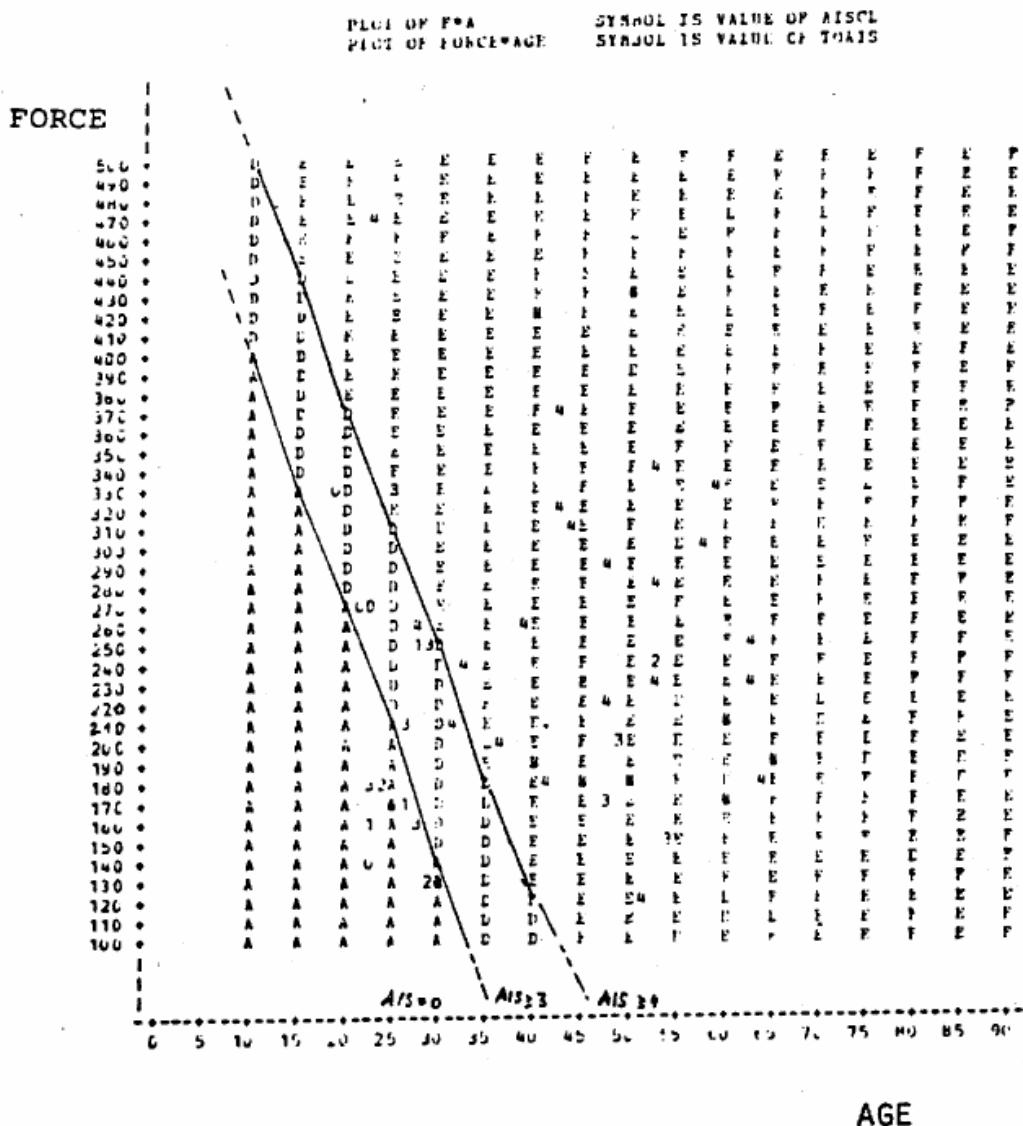


Fig. 4.2 Isoquantplot for the illustration of the prediction results of the logistic regression from AGE and FORCE.
 Zone A: prediction of TOAIS = 0
 Zone B: prediction of TOAIS = 3
 Zone E: prediction of TOAIS = 4
 Numbers in the zones: observed thorax-injury degrees
 FORCE = 1/2 (accel.max. 4th rib impacted side + max. result. accel. Th 12) x bodymass / 75

The crash tests have been conducted at impact velocities of 40, 45, 50 and 60 km/h (Kallieris et al., 1987). In the PMHS 22 acceleration values at head, thorax, spinal column and pelvis have been recorded for each test. The injuries of the PMHS have been scaled according to AIS 80. It was seen in the statistical analyses that the injury levels could be most effectively predicted by the method of logistic regression. In the 90 degree lateral collisions the body injury severity (TAAIS) was generally leading and determined the maximum injury severity (MAIS). Therefore, the prediction of the body injury severity for right side lateral collisions is presented here as an example. Among the 22 as maximum and 3 ms values recorded accelerations the following proved to be the best predictors:

1. Acceleration (3 ms value) in x-direction at lower sternum (BUX3) (g);
2. acceleration (3 ms value) at the 12th thoracic vertebra in y-direction (T12Y3) (g);

The further improvement of the injury prediction has been reached in considering the Body Mass (BMASS) (kg) as covariable. With these covariable combination, the logistic model estimated the following parameters for the injury index Z:

$$Z = 0.15 \text{ BMASS} + 0.08 \text{ T12Y3} + 0.06 \text{ BUX3}.$$

The probability curves for TAAIS rankings 0,4 and 5 are shown in figure 5.1, for impacts from the right. The three tests with TAAIS 2 and 3 in the test series were not considered.

Figure 5.1

Below a Z value of 18.3, the envelope of the AIS probability curves indicates a high probability to be uninjured (the highest probability is below $Z = 18$). Between $Z = 18.3$ and $Z = 20$, a TAAIS of 4 is largely to be expected and above $Z = 20$ the probability for TAAIS 5 of about 45 % increases continuously to 100 % (at $Z = 25$). The enhanced TAAIS difference histogram (see section 4) in figure 5.2 shows that the above mentioned covariable combination as correctly predicts 59 % of the cases. The model predicts the TAAIS in 19 % too high and in 15 % a level too low; each one time, the model underestimated the observed injury for two and 4 AIS degrees.

Figure 5.2

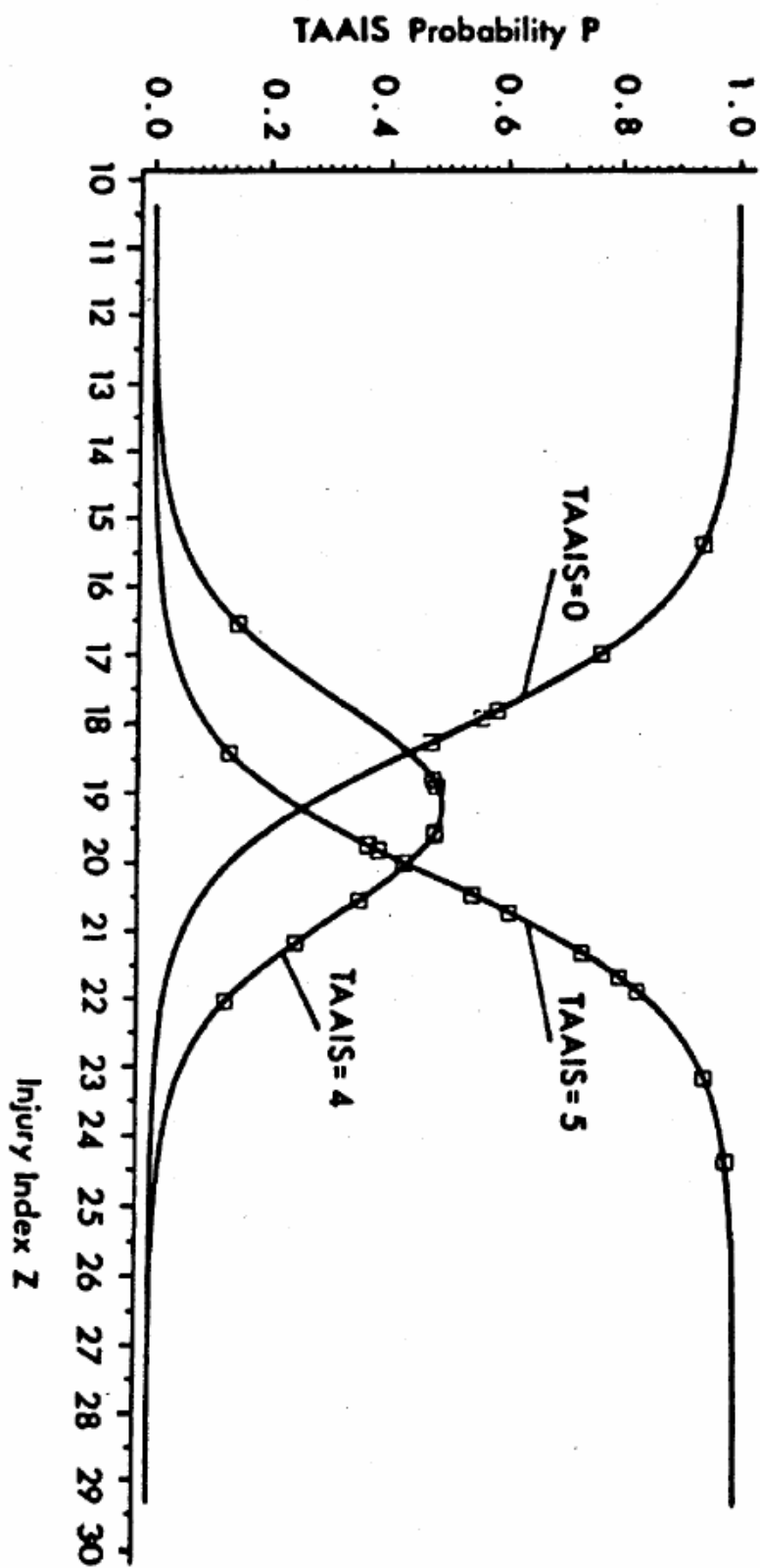
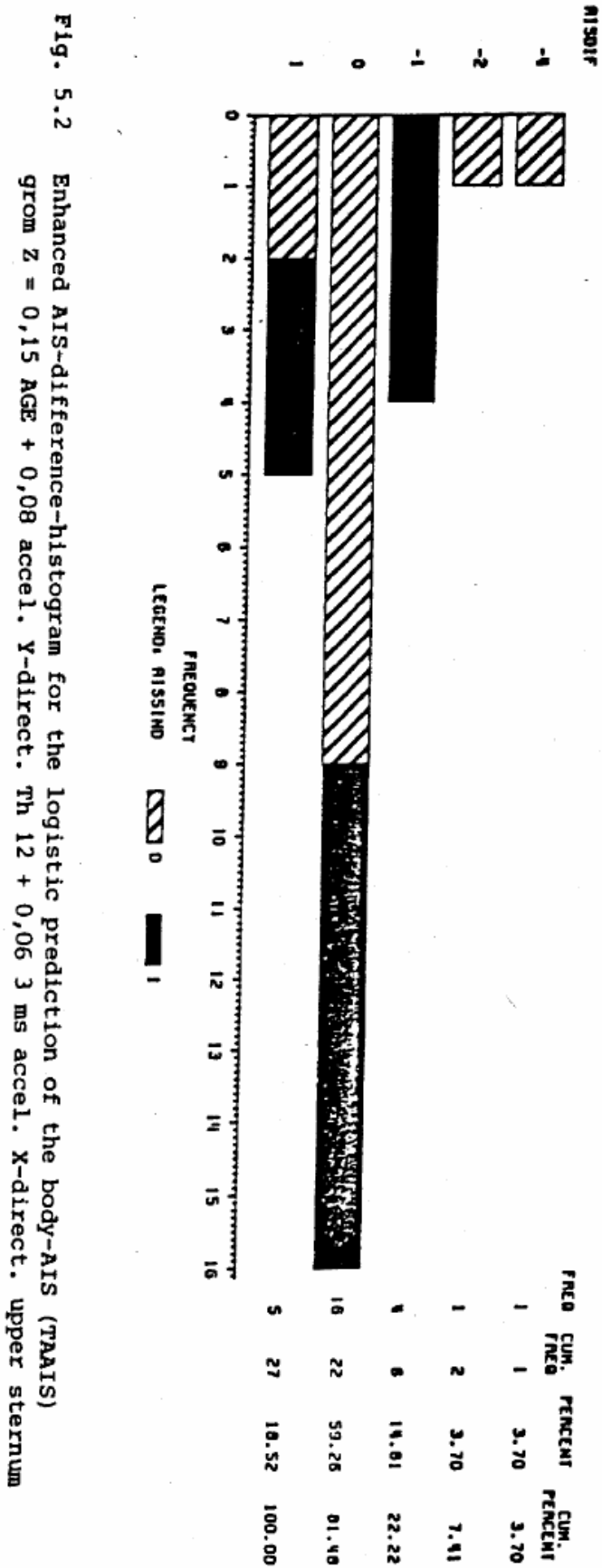


Fig. 5.1 Predicted probability of torso injury (TAAIS) for impacts from the right
□ Observed degree of injury



6. Conclusions

We have presented several multinomial response models of parametric and non-parametric nature. A way of comparing these models and deciding which one is more appropriate than others is given by considering non-parametric alternatives in the construction of a simulation band. This simulation band technique (section 3) lead for the Heidelberg data to the conclusion tht the Logistic response model is appropriate for the analysis of car-to-car side impacts. Comparing the Likelihoods of the Logistic and the Weibull link functions we found no better fit for the Weibull model, see Kallieris, Mattern and Härdle (1986). We furthermore presented a variety of graphical techniques which are of great assistance when looking for suitable predictor variables X, see section 4. Using these techniques we found for example that the Logistic model using the trauma index

$$Z = 0.15 \text{ BMASS} + 0.08 \text{ T12Y3} + 0.06 \text{ BUX3}$$

had good prediction properties for the TAAIS, see section 5.

REFERENCES

- AIS (1980) States JD, Huelke DF, Baker SP, Bryant RW et. al. The Abbreviated Injury Scale, 1980 Revision.
- Akaike H (1977) On Entropy Maximization Principle. In Applications of Statistics, Ed.P.R. Krishaniah. Amsterdam, North Holland
- Berkson J (1951) Why I prefer Logits to Probits. Biometrics 7: 327-339
- Bickel P, Doksum K (1977) Mathematical Statistics. Holden-Day Inc., San Fransisco
- Eppinger RH, Marcus JH, Morgan RM (1984) Development of Dummy and Injury Index for NHTSA's thoracic side impact protection research program, SAE technical paper series 840885, Government/Industry Meeting and Exposition Washington D.C.
- Friedman J, Stuetzle W (1981) Projection Pursuit Regression. J. Amer. Statist. Assoc. 76: 817-823
- Härdle W, Stoker T (1988) Investigating smooth multiple regression models by the method of Average Derivatives. J. Amer. Statist. Assoc., to appear
- Härdle W (1988) Applied Nonparametric Regression. Book to appear

Kallieris D, Mattern R, Härdle W (1986) Belastbarkeitsgrenzen und Verletzungsmechanik des angegurteten Pkw-Insassen beim Seitenaufprall. Phase II: Ansätze für Verletzungsprädiktionen. Schriftenreihe der Forschungsvereinigung Automobiltechnik e.V. (FAT) Nr. 60, Frankfurt/Main 17

Kallieris D, Schmidt Gg, Mattern R (1987) Vertebral Column Injuries in 90 degree Collisions - A study with Post Mortem Human Subjects. Proc. of Intern. IRCOBI Conf. on the Biomechanics of Impacts, Birmingham, 189-192

Neter J, Wasserman W (1974) Applied Linear Statistical Models. Richard D. Irwin, Inc. Homewood Illinois

SAS - Statistical Analysis System, Cary North Carolina

SAS - Supplementary User's Guide, Cary North Carolina

Stone M (1974) Crossvalidatory choice and assessment of statistical predictions (with discussion). Journal of the Royal Statistical Society, Series B, 36, 111-147

This paper is kindly dedicated to Prof. Dr. med. Georg Schmidt, Heidelberg, on the occasion of his 65th birthday.

XploRe

A COMPUTING ENVIRONMENT FOR EXPLORATORY REGRESSION AND DENSITY SMOOTHING

Wolfgang HÄRDLE

Rechts- und Sozialwissenschaftliche Fakultät, Wirtschaftstheoretische Abteilung II, Universität Bonn
Adenauerallee 40-42, D-5300 Bonn 1, FRG
Faculty of Science and Technology, Keio University, Yokohama, Japan

Abstract

XploRe is a graphically oriented interactive system for eXploratory regression and density smoothing. Various nonparametric smoothing techniques for low and high dimensions are implemented. Higher dimensional response surfaces can be approximated by means of additive models: Alternating Conditional Expectations (ACE); Projection Pursuit Regression (PPR); Recursive Partitioning Regression Trees (RPR). XploRe uses the object oriented approach and makes extensive use of the inheritance principle. It is written in TURBO PASCAL and runs on IBM PC/AT, XT or compatibles with MS-DOS.

Zusammenfassung

XploRe ist ein graphisch ausgerichtetes System für eXplorative Regression und Dichteschätzung. Verschiedene nichtparametrische Dichteschätzungen für niedrige und hohe Dimensionen sind implementiert. Höher dimensionale Regressionsoberflächen kann man mit Hilfe folgender Modelle approximieren: Alternating Conditional Expectations (ACE); Projection Pursuit Regression (PPR); Recursive Partitioning Regression Trees (RPR). XploRe bedient sich des objekt-orientierten Ansatzes und macht ausführlichen Gebrauch vom "inheritance principle". Geschrieben ist es in TURBO PASCAL und ist mit MS-DOS auf IBM PC/AT, XT oder kompatiblen Geräten zu benutzen.

*How we think about data analysis
is strongly influenced by the computing
environment in which the analysis is done.*

McDONALD and PEDERSON (1986):

I. WHY AN INTERACTIVE COMPUTING ENVIRONMENT?

XploRe is an interactive system for analyzing various kinds of data smoothing operations. More precisely, XploRe is a graphically oriented computing environment for exploratory regression and density smoothing techniques with sophisticated data management tools. Data can be *rotated, brushed, masked, labeled, transformed and smoothed*. Higher dimensional data clouds can be analyzed by means of *additive models*: Projection Pursuit Regression; Recursive Partitioning Regression Trees; Alternating Conditional Expectations or Average Derivative Estimation. A personal computer, like an IBM PC/AT, XT or compatibles (under MS-DOS) is sufficient for the use of XploRe.

A personal computer or a workstation provides the need of a statistical analysis to improvize alternative ways of interpretation on the spot. A typical scenario in nonparametric regression smoothing is the determination of the best fitting polynomial to a given two-dimensional data set. There are methods which determine the order of a polynomial in an asymptotic sense (SHIBATA (1981)) but it is interesting to see how the fit changes, when the order of the polynomial varies in a small neighborhood around the "best fit". In order to see qualitative changes even for "small variations" of the polynomial order it is necessary to have an interactive computing device.

McDONALD and PEDERSON (1986) point out that the computing environment strongly influences the analysis: If a statistician performs an exploratory or experimental *data mining* in low or high dimensions, he does in fact a special kind of programming work. An interactive computing environment that is designed for the special needs of *experimental programming* of data smoothing is therefore most appropriate. To see why this experimental programming cannot be performed with batch oriented systems consider the following *analysis cycle* (Figure 1.1). A typical round through this cycle is the following. First, a *smoothing* operation (e.g. response surface estimation) is performed based on a specific method and smoothing parameter. Second, the fit and residuals are *examined* for certain features (e.g. remaining structure in the residual pattern). In a third step one *evaluates* the effect and impact of detected features on the fitted curve (e.g. how seriously an outlier influences the smooth). The last step in a round might be to *compare* the current smooth with other fits, possibly stemming from alternative, parametric models. Such a round through the analysis cycle may be repeated many more times. It seems to be impossible to perform effectively this analysis cycle in a batch oriented computing environment. Another szenario inside such an analysis cycle is the *masking* operation on some data points (e.g. outliers). We might want to put aside some of the points and run a certain manipulation with the remaining data in order to study the effect of the left-out points. Batch oriented systems most badly serve this need for interactive decision making since one would basically have to write an additional program for identifying the points which are to be left out. In an interactive computing environment one would mark those points by mouse clicks for instance.

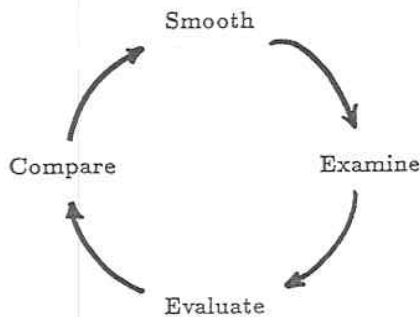


Figure 1.1: Typical analytic cycle

The design of XploRe meets the desiderata for *improvisational programming* by extensive use of interactive graphical methods (mouse oriented selection and identification; pull down menus). Moreover, it supports the user with a set of utilities for masking, brushing, labeling and even rotating of data. XploRe is an *open system* which is written in TURBO PASCAL. It is basically a framework awaiting more "soft work" that enhances the capabilities. Its construction has been influenced by similar systems like S (BECKER and CHAMBERS (1984)) or DINDE (OLDFORD and PETERS (1985)): XploRe uses the object oriented approach and makes extensive use of the inheritance principle to be described below. A detailed description of the functions and procedure to install user written code is given in AERTS and HOLTSBERG (1987).

This paper is organized as follows. Section 2 describes the objects, structure and the basic primitives of XploRe, in particular the workunit objects and the inheritance of attributes. Section 3 is devoted to the description of the display functions. In section 4 the user interface is explained via a construction of a *running median* primitive. Section 5 gives an overview over additive models for fitting high dimensional data. Section 6 gives details about the availability of the software.

Inheritance avoids redundant specification of information and simplifies modification, since information that is common is defined in, and need be changed in, only one place.

OLDFORD and PETERS (1985)

II. OBJECTS AND INHERITANCE

XploRe uses the *object oriented* approach, i.e. the basic elements that are dealt with are structures of simpler variable types and manipulations of data is made solely by reference to those structures (objects). For the purposes of data smoothing we found the following four objects sufficient: *vector*, *workunit*, *picture*, *text*. *Vectors* are the simplest objects, they contain a real data array of variable length. *Workunits* are collections of pointers to vectors and may include display and mask attributes. *Picture* objects are viewports, defining the location and tic marks of the axes in 2D or 3D views. *Texts* are sequences of text lines. The above objects can be *created/deleted*, *activated/deactivated*, *read/written*, *manipulated*, *displayed*.

Moreover, objects can *inherit* certain properties. Workunits can inherit display attributes, such as linestyle or symbols. They can also inherit a *mask*. A mask is a vector of integer classification numbers, including the option to

show points as "invisible". Picture objects inherit the location of the axes and the ticmarks on the screen. Suppose, for example, that a workunit is displayed in a certain picture object. The picture object may then be manipulated by rotation of the pointcloud or by clipping certain parts of the data. This viewport information is inherited by the picture object. If another projection of the same workunit or a different workunit is shown in the same picture object, we would obtain (even after clearing the screen) the same viewport aspect as for the first pointcloud. The inheritance principle thus simplifies overlaying and comparing several curves into the same viewport and hence the same scale. Since display attributes or masks are part of the workunit object, different objects can be distinguished quite easily without using an extra scrapbook aside the computer.

The notion of workunits seems to allow a flexible analysis of several data vectors at a time. Suppose that one wants to analyse a three dimensional data set consisting of vectors X,Y,Z. Workunit wu-one could consist of the vectors X,Y, another wu-two could point to all three vectors. When displaying wu-one one could have detected some interesting points, which one interactively has marked with the classification number "7". Other observations might have been given the mask "invisible". Earlier one might have decided to see then points as stars (except those that leave mask "7"). If wu-two wants to be shown with squares and needles pointing into the (X,Z) plane one can think of the following graphical presentation of the two workunits (Figure 2.1).

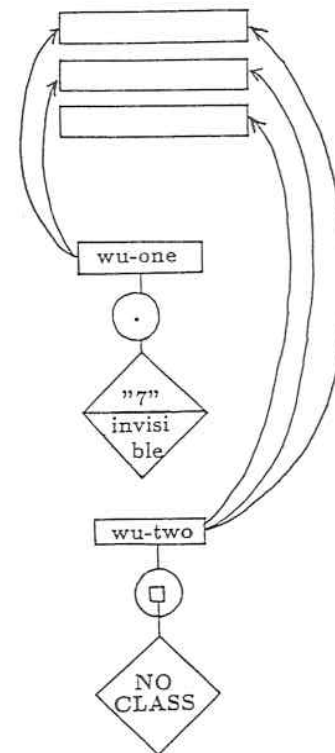


Figure 2.1: Graphical presentation of the two workunits

In a similar way a picture object can be represented as shown in Figure 2.2. The picture object inherited this specific constellation and viewpoint of the axis. It is also indicated above, that the ticmarks may be different along all axis.

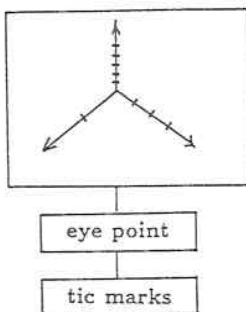


Figure 2.2: Representation of a picture object

The possibility of *activating* objects allows a fast way through command sequences, since as default arguments for object handling always the active object will be assumed. The computation of several smoothing operations of the same (active) workunit does therefore not need the repeated explicit statement of the workunits name.

Different workunits may be displayed in different picture objects. Figure 2.3 shows a workunit (pointing to the raw data) as a pointcloud together with another workunit showing the smooth regression curve both in one picture object. A density estimate of the marginal density of X is displayed in another picture object (viewport "picture 2") at the upper right corner of the screen.

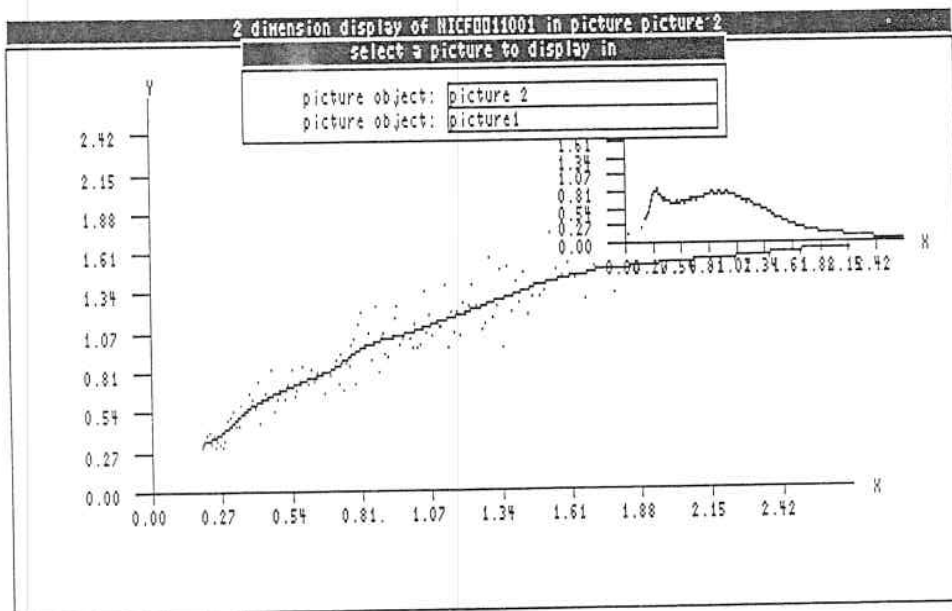


Figure 2.3: Workunits displayed in different picture objects

```

help window level: 4

+ace.hlp)
GENERAL INFORMATION
The ACE algorithm determines the best fitting functions phi[j] in the
following ADDITIVE MODEL

      psi(Y) = sum_{j=1}^p phi[j]( X[j] ) + error,

Where X[j] denotes the j-th coordinate of the p-dimensional predictor
variable X=(X[1], ..., X[p]).
XploRe expects as input for this manipulation a workunit of the form :

      workunit = (X[1], ..., X[p], Y),

Where X and Y denote column vectors. XploRe will create a new workunit
consisting of the fitted functions phi[j], j=1, ..., p and of the fitted
transformation psi.
    
```

Figure 2.4: Example of a help window

Help files can be attached by the system programmer through a stack of "help windows". The designer of the computing environment determines at which analysis stage which "help windows" should appear. The help information is simply obtained by pressing F1. Subsequent pressing of the help key guides through the stack of currently attached help windows. The help windows are in fact internally handled as temporary text objects which are displayed as in Figure 2.4. As more procedures are added to XploRe help-files can be added also. Through a stack mechanism the user can call such help files.

The help windows (and also text objects) can be scrolled backwards and forward by using the PgeDown and PgeUp key. All pull-down menus can be folded and unfolded by successive pressing of the function key F10.

The manipulation of workunits contains currently the following operations:

Regression smoothing

- regressogram
- k-Nearest Neighbour estimation
- super smoothing
- kernel estimation
- weighted averaging using rounded points
- isotonic regression
- running median
- polynomial fitting
- bootstrapping for confidence bounds
- choice of squared error optimal smoothing parameter

Density smoothing

- histogramm
- k-Nearest neighbour estimation
- kernel smoothing
- (log)normal fitting
- choice of smoothing parameter

For details on these operations see HÄRDLE (1988).

Additive Models

- Alternating Conditional Expectations (ACE)
BREIMAN and FRIEDMAN (1985)
- Projection Pursuit Regression (PPR)
FRIEDMAN and STUETZLE (1981)
- Recursive Partitioning Regression Trees (RPR)
BREIMAN, FRIEDMAN, OLSHEN, STONE (1984)
- Average Derivative Estimation (ADE)
HÄRDLE and STOKER (1988)

Other manipulations include the possibility to remove missing observations (or ties) or to define new workunits from an existing one according to certain mask attributes.

III. THE INTERACTIVE DISPLAY

Experimental programming techniques rely very much on an interactive display system. Removal, identification and classification of points should be done in an interactive way by just pointing with a cursor to a group of points. This technique is incorporated in XploRe by the *label* and *mask* option of the graphics command menu, see Figure 3.1.

By clicking the "label" field the cursor can be moved to any point on the screen. After pressing ENTER a window pops up that shows the index of the observation (closest in Eukledian distance) together with the coordinate of the workunit. This feature enables the user to see all coordinates of a high dimensional workunit although he might be looking only at one "interesting" point in a two or three dimensional projection. The "mask" field allows the user to interactively define a rectangle of points which he would like to classify into groups 1-9 or invisible. The "unmask" option reverses this action. The *edit* field allows to change the tickmarks and the scaling of the axis and also the display style of the workunit currently shown. The *movoff* is a switch to *movon* which means that all screen information is stored in a movie fashion to disk. By pressing *movie* the saved screens will be shown, this feature allows

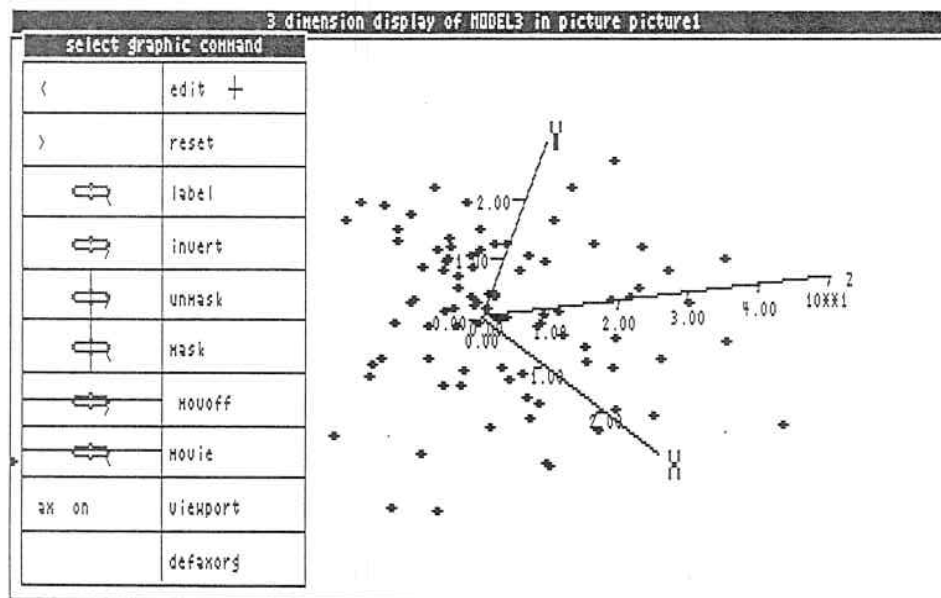


Figure 3.1: Demonstration of label and mask option

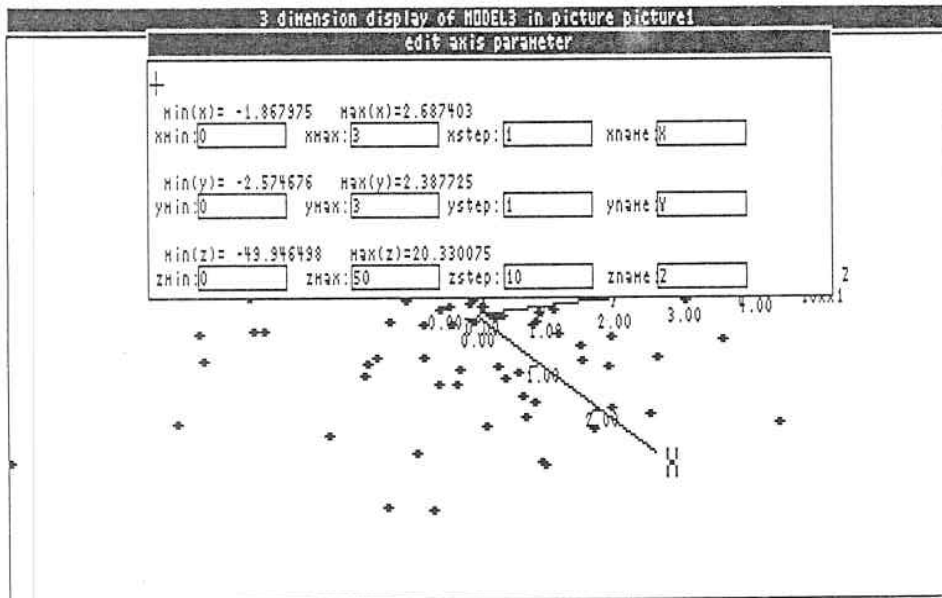


Figure 3.2: Edit command possibilities

tracking of past actions as well as dynamic 3D views of rotating point clouds.

The *viewport* option allows the user to map certain sub-rectangles of the screen to the whole screen. By zooming into a point cloud one may get better understanding of local structures. The *defaxorg* field is for interactive definition of the axis origin. Clicking *ax on* switches to *ax off* which has the effect to display the data without the axis. The six fields above refer to rotations clock- and counterclockwise around each of the three axis in 3D space. The two fields in the upper left corner define the distance of the eyepoint relative to the pointcloud. Clicking successively ">" gives the impression to come closer to the data, whereas "<" makes the distance bigger.

The *edit* field is for locally changing the display style and for inheriting the current picture object ticmarks and axis labelling. Figure 3.2 shows the screen just after clicking "edit" in the situation of Figure 3.1.

The sensitive fields, shown by rectangles, show the current tics. By overwriting in these fields one changes the layout of the axis. The *reset* option gives the standard axis in the cube $[0, \max(x,y,z)]^3$.

IV. INSTALLING NEW PROCEDURES

As an example of how to install own routines I describe how the *running median* primitive was implemented into XploRe. I assume that there is already a procedure *runmed* (y, n, k, s) with input array y , length n , smoothing parameter k and output array s (containing the running median sequence). The user chooses the running median manipulation basically by some mouseclicks and the manipulation refers then to the active workunit object. This workunit has to be sorted by the first column (interpreted as the predictor variable x), then the response variable y has to be stripped off to determine the running median smooth s . It is convenient to build a vector object for this output array s and to create a workunit containing links to the predictor variable x . Inside XploRe these operations would read as follows:

```

procedure dorunmed (wu);
var
  x, y, s: workarray;
  n,k: integer;
  xvec, yvec, svec, newwuobj:objectid;
begin
  quicksort(wu);
  getvector(wu, xvec, x, n, 1);
  getvector(wu, yvec, y, n, 2);
  getparameter(k);
  runmed(y, n, k, s);
  createobj(svec, s, n);
  incvector(svec, s, n);
  createobj(newwu, wuobjpartyp);
  inclink(newwu, xvec, 1);
  inclink(newwu, svec, 2);
end.
    
```

The *getvector* procedure extracts from workunit wu the x and y array. The *createobj* procedure creates an object of the specified type (*vectorpartyp*, *wuobjpartyp*). The *incvector* (*inclink*) procedure includes an array (a link) into vector objects (Workunit objects).

V. HIGHER DIMENSIONAL SMOOTHING TECHNIQUES

Nonparametric regression models with more than one predictor variable are handled in XploRe by means of fitting *additive* models. Currently the following models can be fit for a d -dimensional predictor variable (X_1, \dots, X_d)

$$\Psi(Y) = \sum_{j=1}^d \Phi(X_j) + \text{error}$$

and

$$Y = g(\sum_{j=1}^d \alpha_j X_j) + \text{error}.$$

XploRe uses the ACE-algorithm to find the nonparametric transformations Ψ and $(\Phi_j)_{j=1}^d$, see BREIMAN and FRIED-

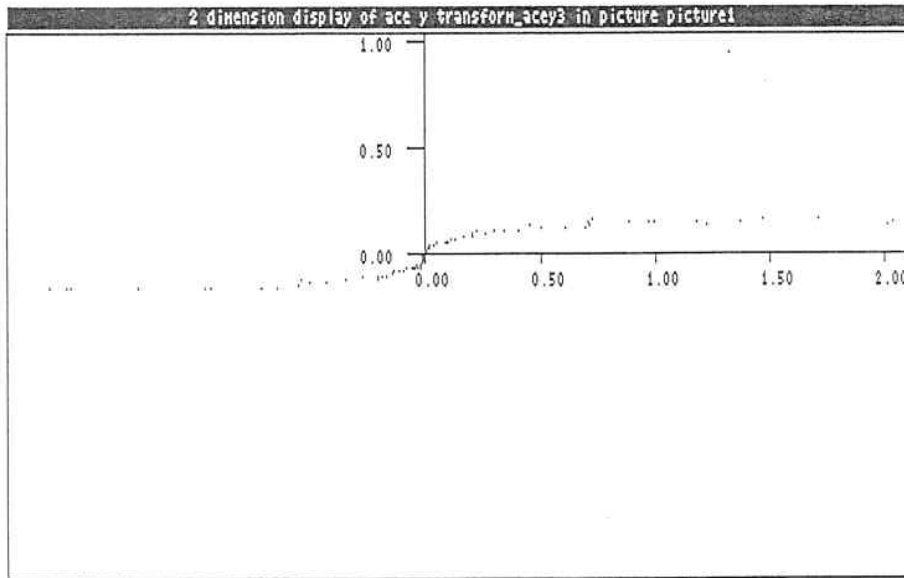


Figure 5.1: Application of the ACE algorithm

MAN (1985). The model exhibiting the "additivity inside", and a nonparametric univariate function g is handled either by Projection Pursuit Regression (PPR), see FRIEDMAN and STUETZLE (1982), or by Average Derivative Estimation (ADE), see HÄRDLE and STOKER (1988). A discrete approximation of the regression curve can be computed using recursive partitioning regression trees (RPR), see BREIMAN et al. (1984). Figure 5.1 shows the transformation $\Psi(y)$ versus y after application of the ACE-algorithm.

The simulated model for this example was

$$Y = (X_1 + X_2)^3 + \text{error.}$$

Clearly the Ψ -transformation recovered the cubic root structure of the data set (as displayed in Figure 3.1). After

optimization over projections we find essentially the same structure by the PPR-technique, see Figure 5.2.

A typical output of the RPR-tree algorithm is shown in Figure 5.3. It gives a good graphical expression of the splits (occurring always parallel to some coordinate axis). In a protocol shows XploRe the corresponding mean and the reduction in sample variance.

VI. AVAILABILITY

The program XploRe is available from the author. It fits on a 1.2 MB disk and runs under MS-DOS with almost all video systems (Hercules, CGA, EGA, Olivetti, etc.). The technical report by AERTS and HOLTSBERG (1987) describing the systems programmer level of XploRe can be obtained by the author, too.

Acknowledgement

The financial support of the Deutsche Forschungsgemeinschaft and the Koizumi Foundation is gratefully acknowledged. The presentation of the paper improved substantially through discussion with A. Hörmann and R. Shibata.

References

- AERTS, M. and HOLTSBERG, A. (1987): Getting Started with XploRe - A Computing Environment for Exploratory Regression and Density Estimation Methods. Technical Report No. A-126, University of Bonn
- BECKER, R.A. and CHAMBERS, J.M. (1984): An Interactive Environment for Data Analysis. Belmont: Wadsworth Press
- BREIMAN, L. and FRIEDMAN, J.H. (1985): Estimating Optimal Transformations for multiple Regression and Correlation (with Discussion). JASA 80, 580-619
- BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R. and STONE, C.J. (1984): Classification and regression trees. Belmont: Wadsworth Press
- FRIEDMAN, J. and STUETZLE, W. (1981): Projection pursuit regression. JASA 76, 817-823

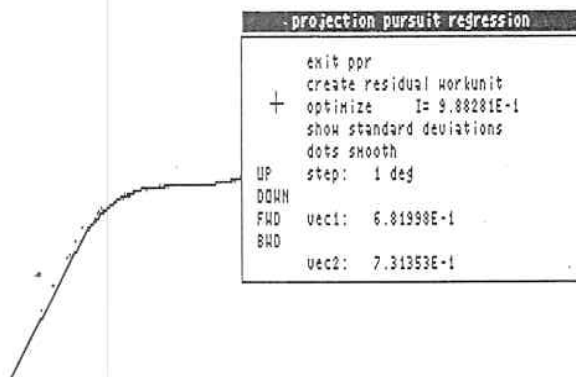


Figure 5.2: Application of the PPR-technique

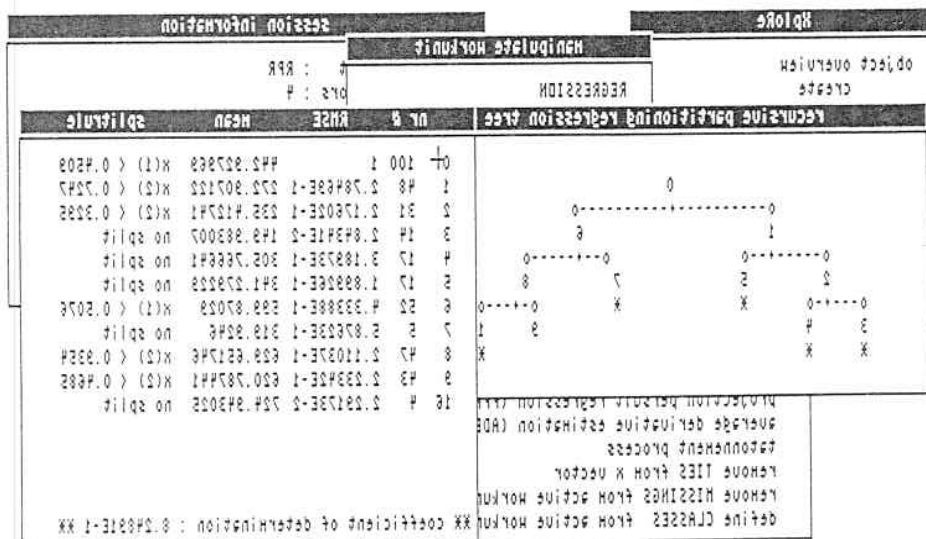


Figure 5.3: Output of the RPR-tree algorithm

HÄRDLE, W. (1988): Applied Nonparametric Regression Book (to appear)

HÄRDLE, W. and STOKER, T. (1988): Investigating multiple regression by the method of averaged derivatives. JASA (to appear)

MCDONALD, J. and PEDERSON, J. (1986): Computing environments for data analysis: part 3: programming environments. Laboratory for Computational Statistics. Stanford University, Technical Report 24

OLDFORD, R.W. and PETERS, S.C. (1985): DINDE: Towards more statistically sophisticated software. Massachusetts Institute of Technology, Technical Report Tr-55

SHIBATA, R. (1981): An optimal selection of regression variables. Biometrika 68, 45-54

SILVERMAN, B.W. (1985): Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). Journal of the Royal Statistical Society (B) 47, 1-45

ROBUST NONPARAMETRIC REGRESSION WITH SIMULTANEOUS SCALE CURVE ESTIMATION¹

BY W. HÄRDLE AND A. B. TSYBAKOV

Universität Bonn and Academy of Sciences of the USSR

Let $\{X_i, Y_i\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ be independent identically distributed random variables. If the conditional distribution $F(y|x)$ can be parametrized by $F(y|x) = F_0((y - m(x))/\sigma(x))$ with a fixed and known distribution F_0 , the regression curve $m(x)$ and scale curve $\sigma(x)$ could be estimated by some parametric method. More generally, we assume that F is unknown and consider nonparametric simultaneous M -type estimates of the unknown functions $m(x)$ and $\sigma(x)$, using kernel estimators for the conditional distribution function $F(y|x)$. We show pointwise consistency and asymptotic normality of these estimates. The rate of convergence is optimal in the sense of Stone (1980). The asymptotic bias term of this robust estimate turns out to be the same as for the linear Nadaraya-Watson kernel estimate.

1. Introduction. Let $(X_1, Y_1), (X_2, Y_2), \dots$ be a sequence of independent identically distributed $(d + 1)$ -dimensional random vectors. Assume that the conditional distribution $P\{Y_1 \leq y | X_1 = x\} = F(y|x)$ has the form $F(y|x) = F_0((y - m(x))/(\sigma(x)))$ with a fixed (but unknown) distribution function F_0 . Call $m(\cdot)$ the regression curve and $\sigma(\cdot)$ the scale curve and assume that they are continuous functions on a set $\Xi \subset \mathbb{R}^d$. Our goal is the simultaneous and nonparametric estimation of the regression curve $m(\cdot)$ and the scale curve $\sigma(\cdot)$ from a random sample $(X_1, Y_1), \dots, (X_n, Y_n)$.

There exists a tradition of nonparametric regression [Nadaraya (1964), Watson (1964)], where $m(x)$ is viewed as an expression for the conditional expectation $E(Y|X = x)$ and this $m(x)$ is estimated by a weighted average of the response variables Y . Mild conditions on the distribution of the Y -variables and on the weights ensure convergence of the estimators to the conditional expectation $E(Y|X = x)$, as Stone (1977) has shown. In the discussion to Stone's paper, Brillinger raised the point that a nonlinear M -type estimate of the regression curve might be worthwhile to study in order to achieve desirable robustness properties.

In this paper we consider more generally simultaneous nonparametric estimation of $m(x)$ and $\sigma(x)$ by M -type smoothers. Our approach is closely related to simultaneous M -estimation of location and scale; see Huber [(1981), Chapter 6.4]. Our approach differs in that we have to consider additional bias terms, due to the fact that $m(\cdot)$ and $\sigma(\cdot)$ are unspecified functions and $F(y|x)$ is estimated by the nonparametric kernel method. The simultaneous M -type smoothers of the

Received April 1985; revised May 1987.

¹Research supported by Deutsche Forschungsgemeinschaft, Sonderforschungsbereiche 303 and 123.

AMS 1980 subject classification. 62G05.

Key words and phrases. Robust curve estimation, M -estimation, nonparametric regression, joint estimation of regression and scale curve, optimal rate of convergence.

regression curve and of the scale curve are determined by a system of nonlinear equations. Define for $s \in \mathbb{R}^+$, $t \in \mathbb{R}$, $x \in \Xi$,

$$(1.1) \quad T_1(s, t) = \int \psi\left(\frac{y-t}{s}\right) dF(y|x)$$

and

$$(1.2) \quad T_2(s, t) = \int \chi\left(\frac{y-t}{s}\right) dF(y|x),$$

with ψ and χ some bounded real functions satisfying additional properties to be stated later. We generalize the preceding notion about $m(x)$ and $\sigma(x)$ by assuming that the curves $m(x)$ and $\sigma(x)$ can be represented as simultaneous zeros of T_1 and T_2 , i.e., $T_1(\sigma(x), m(x)) = T_2(\sigma(x), m(x)) = 0$.

The unknown conditional distribution $F(y|x)$ is estimated by the kernel method,

$$F_n(y|x) = \sum_{i=1}^n W_{ni}(x; X_1, \dots, X_n) I(Y_i \leq y).$$

Here $\{W_{ni}\}_{i=1}^n$ denotes a sequence of weights

$$W_{ni}(x; X_1, \dots, X_n) = \frac{K((X_i - x)/h)}{\sum_{j=1}^n K((X_j - x)/h)}$$

with kernel $K: \mathbb{R}^d \rightarrow \mathbb{R}$ and bandwidth sequence $h = h_n \in \mathbb{R}^+$. In analogy to (1.1) and (1.2) the nonparametric estimates $(m_n(x), \sigma_n(x))$ are defined as simultaneous zeros of

$$(1.3) \quad T_{1n}(s, t) = \int \psi\left(\frac{y-t}{s}\right) dF_n(y|x)$$

and

$$(1.4) \quad T_{2n}(s, t) = \int \chi\left(\frac{y-t}{s}\right) dF_n(y|x).$$

Such simultaneous zeros exist as is shown in Theorem 1. Under regularity conditions on the kernel and the functions ψ and χ , we prove strong consistency of $(m_n(x), \sigma_n(x))$ as well as the asymptotic normality of

$$\sqrt{nh^d} \left[\begin{pmatrix} m_n(x) \\ \sigma_n(x) \end{pmatrix} - \begin{pmatrix} m(x) \\ \sigma(x) \end{pmatrix} \right].$$

Numerous examples of functions ψ and χ for the simultaneous M -estimation of location and scatter can be found in the literature on robust estimation. For instance, the well known

$$\begin{aligned} \psi(u) &= -k \vee (k \wedge u), & k > 0, \\ \chi(u) &= c^2 \wedge u^2 - \beta, & 0 < \beta < c^2, \end{aligned}$$

fulfill our assumptions for suitable β [Assumption (A1)]; see Huber [(1981), page

137]. Note that in the case $c = k = \infty$ and $\beta = \int u^2 dF_0(u)$, this class of functions ψ and χ give the Nadaraya–Watson kernel estimate and the natural estimate

$$\sigma(x) = \left[n^{-1} \sum_{i=1}^n W_{ni}(x) (Y_i - m_n(x))^2 \right]^{1/2}$$

for the conditional scale $\sigma(x)$.

The estimation of $m(x)$ alone by M -type estimators has been investigated by several authors. Tsybakov (1982a, b) and Härdle (1984) showed consistency and asymptotic normality. Some Monte Carlo results for kernel “ M -smoothers” are presented in Härdle and Gasser (1984). A recursive M -type regression function estimator was considered by Tsybakov (1982a, b). An M -type smoothing spline was considered by Huber (1979), Cox (1983) and Silverman (1985). An M -type estimation on functional classes was investigated by Nemirovskii, Polyak and Tsybakov (1983).

The results of this paper are relevant for several applications. For instance, in physical chemistry the Raman spectra estimation instrumental noise is considerably reduced by the robust estimator $m_n(x)$; see Bussian and Härdle (1984). In image processing Justusson (1981) applied two-dimensional running medians to image restoration from noisy signals. Hildenbrand and Hildenbrand [(1986), Figure 7] report aberrant observations in an analysis of expenditure curves for potatoes as a function of (normalized) income and use a robust two stage estimation technique. Also in the a posteriori construction of parametric models following a previous nonparametric analysis, a robust nonparametric estimator seems to be desirable. Outliers might mimic nonexistent structure resulting in a biased parametric model.

It has been conjectured that robust smoothers are inclining to oversmooth the data by chopping off existing peaks of the regression curve which finally would result in an increased bias. It turns out (Theorem 2) that this conjecture is not true: The “ M -smoothers” considered here have the same asymptotic bias as their linear relatives such as the Nadaraya–Watson estimator $\int y dF_n(y|x)$. Our representation of $(m(x), \sigma(x))$ as zeros of certain functionals of the conditional distribution $F(y|x)$ introduces a slightly more general class of regression curves than the conditional expectation curve of Y or X . We may also note that even when outliers are absent it is reasonable to complement the nonparametric regression estimate m_n by a suitable estimate of its accuracy σ_n . This was not commonly realized in earlier work on nonparametric regression. In the setting of parametric linear regression, however, robust estimation from heteroscedastic data has been considered by Carroll (1982) via construction of a (linear) nonparametric estimate of the scale curve $\sigma(x)$. There are some open questions. In this paper we do not consider the choice of the bandwidth $h = h_n$ that has to be made in practice. A cross-validatory choice for the Nadaraya–Watson estimator has been proposed by Härdle and Marron (1985). In a forthcoming paper we will present an adaptive bandwidth selection rule that minimizes the maximal risk over specific classes of regression curves. Also the functions ψ and χ have to be

chosen in practice. Our result on the asymptotic normality of (m_n, σ_n) suggests that, as in the classical M -estimation of location and scale, there are estimators that minimize the maximal asymptotic variance over a certain class of distributions. Is it possible to adapt ψ and χ to the underlying F_0 in order to achieve asymptotic efficiency?

2. Simultaneous M -smoothing of regression and scale curve. The following regularity conditions on ψ and χ are needed to ensure consistency of the estimates.

- (A1) The distribution function F_0 is continuous and symmetric. Further every nonempty neighborhood of zero has nonnull F_0 -measure, and $\int \chi(u) dF_0(u) = 0$.
- (A2) The function $\psi(t)$ is continuous, nondecreasing, bounded and odd.
- (A3) The function $\chi(t)$ is continuous, bounded and even, nondecreasing for $t \geq 0$ and strictly increasing in the interval, where $\chi(t) < \chi(\infty)$.
- (A4) The functions $t^{-1}\psi(t)$ and $t^{-2}(\chi(t) - \chi(0))$ are continuous and nonincreasing for $t \geq 0$.
- (A5) There exists a constant $t_0 > 0$ such that $\chi(t_0) > 0$ and $t^{-1}\psi(t) > 0$ for $t \leq t_0$.

The next two conditions specify the class of kernels K and regulate the speed of the bandwidth sequence.

- (A6) The kernel $K: \mathbf{R}^d \rightarrow \mathbf{R}$ is bounded and nonnegative with bounded support and $\int K(u) du \neq 0$.
- (A7) The sequence of bandwidths $h = h_n$ tends to zero such that

$$(a) \quad nh^d \rightarrow \infty$$

or

$$(b) \quad nh^d / \log n \rightarrow \infty.$$

Assumption (A7a) is necessary to obtain convergence in probability whereas (A7b) is used to show almost sure convergence. Such conditions on the rate of convergence of h_n are compatible to other smoothing techniques; see the survey article of Collomb (1981). Finally we postulate continuity of the marginal density $f(\cdot)$ of X , the regression curve and the scale curve.

- (A8) The density $f(\cdot)$ of X is continuous and positive in some neighborhood of x .
- (A9) The functions $m(\cdot)$ and $\sigma(\cdot)$ are continuous in some neighborhood of x , and $\sigma(x) > 0$.

THEOREM 1. *Let (A1)–(A9) be satisfied. Then*

- (i) $(m(x), \sigma(x))$ are unique simultaneous zeros of (1.1) and (1.2);
- (ii) there exist simultaneous zeros $(m_n(x), \sigma_n(x))$ of (1.3) and (1.4) with probability tending to 1 as $n \rightarrow \infty$ (a.s. for n sufficiently large) if (A7a) [respectively, (A7b)] holds;

(iii) for any simultaneous zeros $(m_n(x), \sigma_n(x))$ of (1.3) and (1.4),
 $(m_n(x), \sigma_n(x)) \rightarrow (m(x), \sigma(x)), \quad n \rightarrow \infty,$
 in probability (almost surely) if (A7a) [respectively, (A7b)] holds.

The next conditions are refinements of the preceding assumptions and are used to show the asymptotic normality of (m_n, σ_n) .

(A10) The functions ψ and χ are continuously differentiable with bounded derivative and $t\psi'(t)$ and $t\chi'(t)$ are continuous and bounded.

$$(A11) \quad 0 < \varphi_0 = \int \psi'(u) dF_0(u),$$

$$0 < \kappa_0 = \int u\chi'(u) dF_0(u).$$

Note that (A10) implies that the preceding two integrals are finite.

(A12) The functions m and σ are Lipschitz continuous with Lipschitz constants L, L' , respectively. The directional derivatives

$$m'(x; u) = \lim_{\varepsilon \rightarrow 0} \varepsilon^{-1} (m(x + \varepsilon u) - m(x))$$

[and similarly for $\sigma(x)$] exist for all $u \in \mathbb{R}^d$.

Assumption (A12) appears to be the minimal smoothness assumption under which the asymptotic normality may yet be expected. Using the argument of Stone (1980) one can show that under (A12) the squared error optimal pointwise rate of convergence of (m_n, σ_n) to (m, σ) is attained for $h_n \sim n^{-1/(d+2)}$. This is the bandwidth rate for which the squared bias and the variance of the estimate are asymptotically of the same order. Therefore it is reasonable to assume:

(A13) There is a constant $0 \leq \beta < \infty$ such that

$$\lim_{n \rightarrow \infty} h_n n^{1/(d+2)} = \beta.$$

Note that $\beta \neq 0$ corresponds to the optimal rate $\{n^{-1/(d+2)}\}$; see Stone (1980). In the case $\beta = 0$ the bias is of smaller order than the variance. The case $\beta = \infty$ is not considered. In this case the asymptotic variance of (m_n, σ_n) is negligible compared to the bias. The convergence rate of (m_n, σ_n) could be improved by so-called higher order kernels at the expense of assuming higher differentiability of (m, σ) [Härdle and Marron (1985)]. For instance, if m and σ are twice continuously differentiable and a smooth symmetric kernel is used, the rate $\{n^{2/(4+d)}\}$ can be achieved. Indeed, a second order Taylor expansion in (5.9) would result in the rate h_n^{2d} for the bias. Setting $h_n \sim n^{-1/(4+d)}$ yields the faster rate $\{n^{-2/(4+d)}\}$ for (m_n, σ_n) to (m, σ) .

THEOREM 2. Let (A1)–(A13) be satisfied and define $\varphi_2 = \int \psi^2(u) dF_0(u)$ and $\kappa_2 = \int \chi^2(u) dF_0(u)$. Then, as $n \rightarrow \infty$,

$$\sqrt{nh^d} \left[\begin{pmatrix} m_n(x) \\ \sigma_n(x) \end{pmatrix} - \begin{pmatrix} m(x) \\ \sigma(x) \end{pmatrix} \right]$$

is asymptotically normally distributed with mean

$$\beta^{d/2+1} \left(\frac{\int m'(x; u) K(u) du}{\int \sigma'(x; u) K(u) du} \right) / \int K(u) du$$

and covariance matrix

$$\frac{\sigma^2(x) \int K^2(u) du}{f(x) (\int K(u) du)^2} \begin{pmatrix} \varphi_2/\varphi_0^2 & 0 \\ 0 & \kappa_2/\kappa_0^2 \end{pmatrix}.$$

COROLLARY 1. If $\beta \neq 0$, then as $n \rightarrow \infty$,

$$n^{1/(d+2)}(m_n(x) - m(x)) \rightarrow_{\mathcal{D}} N(b_m, V_m),$$

where

$$b_m = \beta \int m'(x; u) K(u) du / \int K(u) du,$$

$$V_m = \frac{\sigma^2(x)}{\beta^d f(x)} \frac{\varphi_2}{\varphi_0^2} \int K^2(u) du / \left(\int K(u) du \right)^2.$$

Also,

$$n^{1/(d+2)}(\sigma_n(x) - \sigma(x)) \rightarrow_{\mathcal{D}} N(b_\sigma, V_\sigma),$$

where

$$b_\sigma = \beta \int \sigma'(x; u) K(u) du / \int K(u) du,$$

$$V_\sigma = \frac{\sigma^2(x)}{\beta^d f(x)} \frac{\kappa_2}{\kappa_0^2} \int K^2(u) du / \left(\int K(u) du \right)^2.$$

3. Preliminary lemmas.

LEMMA 1. Let $\{Q_n(t)\}$ be a sequence of bounded nondecreasing random functions defined on the closed interval $U \subseteq \mathbb{R}$. Suppose that $Q(t)$ is a continuous nondecreasing bounded function on U . Assume:

1. $Q_n(t) \rightarrow Q(t)$, $n \rightarrow \infty$, a.s. (in probability) $\forall t \in U$.
2. If the right endpoint of U is $+\infty$, then

$$\lim_{t \rightarrow \infty} Q_n(t) = \lim_{t \rightarrow \infty} Q(t), \quad \forall n,$$

and if the left endpoint of U is $-\infty$, then

$$\lim_{t \rightarrow -\infty} Q_n(t) = \lim_{t \rightarrow -\infty} Q(t), \quad \forall n.$$

Then

$$\sup_{t \in U} |Q_n(t) - Q(t)| \rightarrow 0, \quad n \rightarrow \infty,$$

a.s. (in probability, respectively).

The proof of Lemma 1 is obtained by the same argument as for the Glivenko–Cantelli theorem.

LEMMA 2. Let F_0 be continuous and let conditions (A6)–(A9) be satisfied. Then

$$\sup_{y \in \mathbb{R}} |F_n(y|x) - F(y|x)| \rightarrow 0, \quad n \rightarrow \infty,$$

in probability (almost surely) if (A7a) [respectively, (A7b)] holds.

PROOF. From Collomb [(1980), Proposition 1 (2)], it follows that $F_n(y|x) \rightarrow F(y|x)$, $n \rightarrow \infty$, $\forall y \in \mathbb{R}$, in probability (almost surely) if (A7a) [respectively, (A7b)] holds. \square

Now Lemma 1 is applied with $Q_n(t) = F_n(t|x)$ and $Q(t) = F(t|x)$ to yield uniform convergence of conditional functions.

LEMMA 3. Let $Q(y, t)$ be continuous in (y, t) and a bounded function of $y \in \mathbb{R}$, $t \in T$, T a compact set in \mathbb{R}^d . If

$$(3.1) \quad \int \varphi(y) F_n(dy|x) \rightarrow \int \varphi(y) F(dy|x), \quad n \rightarrow \infty,$$

a.s. (in probability) for any bounded continuous function φ , then

$$\sup_{t \in T} \left| \int Q(y, t) F_n(dy|x) - \int Q(y, t) F(dy|x) \right| \rightarrow 0,$$

$n \rightarrow \infty$, a.s. (in probability).

PROOF. Consider for brevity only the a.s. case. Let N be a minimal ε -net on T in the Euclidean metric. Let

$$V_n(t) = \int Q(y, t) F_n(dy|x), \quad V(t) = \int Q(y, t) F(dy|x).$$

Then,

$$(3.2) \quad \begin{aligned} \sup_{t \in T} |V_n(t) - V(t)| &\leq \max_{\tilde{t} \in N} |V_n(\tilde{t}) - V(\tilde{t})| \\ &+ \max_{\tilde{t} \in N} \sup_{t: |t-\tilde{t}| \leq \varepsilon} |V_n(t) - V_n(\tilde{t})| \\ &+ \max_{\tilde{t} \in N} \sup_{t: |t-\tilde{t}| \leq \varepsilon} |V(t) - V(\tilde{t})|. \end{aligned}$$

In (3.2), let $n \rightarrow \infty$ and then $\varepsilon \rightarrow 0$. The first summand in (3.2) tends to 0 a.s. as $n \rightarrow \infty$ since (3.1) holds and since $\text{card } N = N(\varepsilon) < \infty$. The third summand tends to 0 as $\varepsilon \rightarrow 0$ by continuity of $V(t)$ on T .

It remains to prove that the second summand tends to 0.

Let

$$\varphi_\varepsilon(y) = \sup_{t, \tilde{t} \in T: |t-\tilde{t}| \leq \varepsilon} |Q(y, t) - Q(y, \tilde{t})|.$$

Then

$$\begin{aligned} \max_{\tilde{t} \in N} \sup_{t: |t-\tilde{t}| \leq \varepsilon} |V_n(t) - V_n(\tilde{t})| &\leq \sup_{t, \tilde{t} \in T: |t-\tilde{t}| \leq \varepsilon} \int |Q(y, t) - Q(y, \tilde{t})| F_n(dy|x) \\ &\leq \int \varphi_\varepsilon(y) F_n(dy|x). \end{aligned}$$

Since Q is continuous in (y, t) then φ_ε is continuous in y . Therefore (3.1) yields

$$(3.3) \quad \limsup_n \max_{\tilde{t} \in N} \sup_{t: |t-\tilde{t}| \leq \varepsilon} |V_n(t) - V_n(\tilde{t})| \leq \int \varphi_\varepsilon(y) F(dy|x).$$

But $\lim_{\varepsilon \rightarrow 0} \varphi_\varepsilon(y) = 0, \forall y$, because Q is continuous in (y, t) . In view of boundedness of φ_ε the right side of (3.3) tends to 0 as $\varepsilon \rightarrow 0$. This completes the proof. \square

4. Proof of Theorem 1. Without loss of generality assume that $m(x) = 0$ and $\sigma(x) = 1$. The assertion (i) of Theorem 1 is deduced from the following lemma.

LEMMA 4.

(4.1) For each t there exists a unique solution $s^*(t)$ of $T_2(s^*(t), t) = 0$.

(4.2) $s^*(t)$ is a continuous function and $\inf_t s^*(t) > 0$.

(4.3) For each s , $T_1(s, t) = 0$ if and only if $t = 0$.

PROOF. The assertions (4.1) and (4.2) are contained in Theorem 1 and Lemma 2 of Maronna (1976). The "if" part of (4.3) follows from the fact that F_0 is symmetric and ψ is odd. The "only if" part of (4.3) follows from monotonicity of ψ and (4.10). (Set $t = \pm \varepsilon$ there to prove by contradiction.) \square

We shall prove the assertions (ii) and (iii) of Theorem 1 in the case when (A7b) holds [the case (A7a) is considered in a similar way].

By (A2) and (A5) the function ψ is monotone and $\psi(\infty) > 0$ and $\psi(-\infty) < 0$. Hence there exists a solution $t_n(s)$ of

$$(4.4) \quad T_{1n}(s, t_n(s)) = 0, \quad \forall s > 0.$$

From Lemma 2 and continuity of F_0 it follows that F_n satisfies condition (E) of Maronna (1976), a.s. for large n . It is easy to verify that conditions (A2)–(A5) coincide with the univariate version of conditions (A)–(D) of Maronna (1976). Therefore we can apply Theorem 2 of Maronna (1976), which yields the assertion (ii) of Theorem 1. In addition there exist some constants $a, A, 0 < a \leq A < \infty$ such that

$$(4.5) \quad a \leq \sigma_n \leq A, \text{ a.s., for } n \text{ sufficiently large.}$$

This follows in the same manner as (5.1) in Maronna [(1976), page 59] (use Lemma 2 instead of the Glivenko–Cantelli theorem there).

LEMMA 5. For any sequence of functions $\{t_n\}$ satisfying (4.4),

$$(4.6) \quad \sup_{a \leq s \leq A} |t_n(s)| \rightarrow 0, \quad \text{a.s., } n \rightarrow \infty.$$

PROOF. Note that for fixed $s_0 > 0$ the function $T_1(s_0, t)$ is nonincreasing in t . Therefore, if for some constants a, A and arbitrarily small $\varepsilon > 0$,

$$(4.7) \quad \inf_{a \leq s \leq A} T_1(s, -\varepsilon) > 0,$$

$$(4.8) \quad \sup_{a \leq s \leq A} T_1(s, +\varepsilon) < 0$$

and

$$(4.9) \quad \sup_{a \leq s \leq A} |T_1(s, \pm\varepsilon) - T_{1n}(s, \pm\varepsilon)| \rightarrow 0, \quad \text{a.s., } n \rightarrow \infty,$$

then

$$\liminf_n \inf_{a \leq s \leq A} T_{1n}(s, -\varepsilon) > 0, \quad \text{a.s.,}$$

$$\limsup_n \sup_{a \leq s \leq A} T_{1n}(s, +\varepsilon) < 0, \quad \text{a.s.,}$$

which entails (4.6).

It remains to show (4.7)–(4.9). We first show (4.9) only for one case; the other cases follow by symmetry. Let $U = [a, A]$ and let

$$Q_n(s) = \int g(y, s) F_n(dy|x), \quad Q(s) = \int g(y, s) F(dy|x),$$

with

$$g(y, s) = \psi\left(\frac{y - \varepsilon}{s}\right) I(y - \varepsilon \leq 0).$$

Note that Q_n and Q are nondecreasing functions; therefore, by Lemmas 1 and 2 and by Billingsley [(1968), Theorem (5.2(iii))], we have that

$$\sup_{s \in U} |Q_n(s) - Q(s)| \rightarrow 0, \quad \text{a.s., } n \rightarrow \infty,$$

which entails (4.9).

It remains to show (4.7) because (4.8) will follow by a symmetry argument. Note that by conditions (A1) and (A2) for all $s \in \mathbb{R}^+$,

$$T_1(s, -\varepsilon) = \int \psi\left(\frac{u + \varepsilon}{s}\right) dF_0(u) \geq T_1(s, 0) = 0.$$

Hence, by continuity of $T_1(s, -\varepsilon)$, it suffices to show

$$(4.10) \quad T_1(s, -\varepsilon) - T_1(s, 0) \neq 0,$$

for all $s \in U$. Assume that (4.10) is not true; then there is an $\tilde{s} \in U$ such that the set $\{u: \psi((u + \varepsilon)/\tilde{s}) \neq \psi(u/\tilde{s})\}$ has F_0 -measure zero. By (A1) it is open and does not contain any neighborhood of zero; therefore, $\psi(\varepsilon/\tilde{s}) = \psi(0) = 0$. This contradicts (A5) and shows (4.10). \square

In order to prove Theorem 1(iii) we first show that for any small $\delta > 0$ there exists a compact interval $I = I_\delta$ centered by 0 such that

$$(4.11) \quad \liminf_n \inf_{t \in I} T_{2n}(s^*(t) - \delta, t) > 0, \quad \text{a.s.},$$

$$(4.12) \quad \limsup_n \sup_{t \in I} T_{2n}(s^*(t) + \delta, t) < 0, \quad \text{a.s.}$$

We show (4.11). The proof of (4.12) is similar. Fix some $\delta \in (0, \inf_t s^*(t))$. Using (4.2) and continuity of ψ one obtains that the function

$$Q(y, t) = \psi((y - t)/(s^*(t) - \delta))$$

is continuous in (y, t) . Therefore $T_2(s^*(t) - \delta, t)$ is continuous in t . In addition, monotonicity of $T_2(s, 0)$ and (4.1) entail that $T_2(s^*(0) - \delta, 0) > 0$. Hence there exists a compact interval I centered by 0 such that

$$(4.13) \quad \inf_{t \in I} T_2(s^*(t) - \delta, t) > 0.$$

By Lemma 3

$$(4.14) \quad \sup_{t \in I} |T_2(s^*(t) - \delta, t) - T_{2n}(s^*(t) - \delta, t)| \rightarrow 0, \quad \text{a.s., } n \rightarrow \infty,$$

and (4.11) follows from (4.13) and (4.14).

Now observe that $m_n = t_n(\sigma_n)$ by definition. Pulling (4.5) and (4.6) together yields

$$(4.15) \quad m_n \rightarrow 0, \quad \text{a.s., } n \rightarrow \infty.$$

In particular, $m_n \in I$, a.s. for n sufficiently large, and hence by (4.11) and (4.12)

$$\liminf_n T_{2n}(s^*(m_n) - \delta, m_n) > 0, \quad \text{a.s.},$$

$$\limsup_n T_{2n}(s^*(m_n) + \delta, m_n) < 0, \quad \text{a.s.}$$

These inequalities imply that $\sigma_n - s^*(m_n) \rightarrow 0$, a.s., $n \rightarrow \infty$, since $T_{2n}(s, m_n)$ is monotone in s and $T_{2n}(\sigma_n, m_n) = 0$. Applying (4.12) and (4.15) we finally obtain

$$|\sigma_n - s^*(0)| \leq |\sigma_n - s^*(m_n)| + |s^*(m_n) - s^*(0)| \rightarrow 0, \quad \text{a.s., } n \rightarrow \infty.$$

Since $s^*(0) = 1 = \sigma(x)$ this completes the proof of Theorem 1(iii).

5. Proof of Theorem 2. To simplify our notation we introduce the parameter $\vartheta = (t, s)$ and the function

$$\Psi(y, \vartheta) = \begin{pmatrix} \psi((y - t)/s) \\ \chi((y - t)/s) \end{pmatrix}.$$

Recall that the point x was fixed. We will write ϑ_n for $(m_n(x), \sigma_n(x))$ and ϑ^* for $(m(x), \sigma(x))$.

Introduce the matrix of derivatives

$$\Psi'(y, \vartheta) = \begin{pmatrix} -1/s\psi'((y-t)/s) & (t-y)/s^2\psi'((y-t)/s) \\ -1/s\chi'((y-t)/s) & (t-y)/s^2\chi'((y-t)/s) \end{pmatrix}.$$

The existence of this matrix in some neighborhood of ϑ^* is guaranteed by condition (A10) and positiveness of $\sigma(x)$. Now

$$\begin{aligned} & \sqrt{nh_n^d} \int \Psi(y, \vartheta^*) F_n(dy|x) \\ (5.1) \quad & = \sqrt{nh_n^d} \int (\Psi(y, \vartheta^*) - \Psi(y, \vartheta_n)) F_n(dy|x) \\ & = \left(\int_0^1 \left\{ \int \Psi'(y, \tau\vartheta^* + (1-\tau)\vartheta_n) F_n(dy|x) \right\} d\tau \right) \sqrt{nh_n^d} (\vartheta^* - \vartheta_n), \end{aligned}$$

if $|\vartheta^* - \vartheta_n|$ is small enough for the existence of $\Psi'(y, \vartheta)$ for $\vartheta: |\vartheta - \vartheta^*| \leq |\vartheta_n - \vartheta^*|$.

Next we shall prove

$$(5.2) \quad \sup_{\{\vartheta: |\vartheta - \vartheta^*| \leq |\vartheta_n - \vartheta^*|\}} \left\| \int \Psi'(y, \vartheta) F_n(dy|x) - \int \Psi'(y, \vartheta^*) F(dy|x) \right\| \rightarrow_p 0, \quad n \rightarrow \infty,$$

where $\|\cdot\|$ is any norm in the space of 2×2 matrices. It suffices to prove (5.2) for all components of matrices separately. Condition (A10) and positiveness of $\sigma(x)$ imply that the components of $\Psi'(y, \vartheta)$ are continuous and bounded in (y, ϑ) for $y \in \mathbb{R}$ and ϑ belonging to some neighborhood of ϑ^* . Hence by Lemma 3,

$$\sup_{|\vartheta - \vartheta^*| \leq \delta} \left\| \int \Psi'(y, \vartheta) F_n(dy|x) - \int \Psi'(y, \vartheta) F(dy|x) \right\| \rightarrow_p 0, \quad n \rightarrow \infty,$$

for sufficiently small $\delta > 0$. This gives

$$(5.3) \quad \sup_{\{\vartheta: |\vartheta - \vartheta^*| \leq |\vartheta_n - \vartheta^*|\}} \left\| \int \Psi'(y, \vartheta) F_n(dy|x) - \int \Psi'(y, \vartheta) F(dy|x) \right\| \rightarrow_p 0, \quad n \rightarrow \infty,$$

since by Theorem 1(iii), $\vartheta_n \rightarrow_p \vartheta^*$ as $n \rightarrow \infty$. In addition,

$$(5.4) \quad \sup_{\{\vartheta: |\vartheta - \vartheta^*| \leq |\vartheta_n - \vartheta^*|\}} \left\| \int \Psi'(y, \vartheta) F(dy|x) - \int \Psi'(y, \vartheta^*) F(dy|x) \right\| \rightarrow_p 0, \quad n \rightarrow \infty,$$

by uniform continuity of $\int \Psi'(y, \vartheta) F(dy|x)$ in some neighborhood of ϑ^* . We see that (5.2) follows from (5.3) and (5.4).

Using (5.2) one obtains

$$\begin{aligned} (5.5) \quad & \int_0^1 \left\{ \int \Psi'(y, \tau\vartheta^* + (1-\tau)\vartheta_n) F_n(dy|x) \right\} d\tau \\ & \rightarrow_p \int \Psi'(y, \vartheta^*) F(dy|x) = \frac{1}{\sigma(x)} \begin{pmatrix} \varphi_0 & 0 \\ 0 & \kappa_0 \end{pmatrix}, \quad n \rightarrow \infty. \end{aligned}$$

We now study the asymptotic distribution of the left-hand side of (5.1). Write

$$(5.6) \quad \sqrt{nh_n^d} \int \Psi(y, \vartheta^*) F_n(dy|x) = \left(\frac{1}{nh_n^d} \sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right) \right)^{-1} G_n,$$

where

$$G_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{pmatrix} \eta_{in} \\ \zeta_{in} \end{pmatrix},$$

$$\eta_{in} = \frac{1}{\sqrt{h_n^d}} \Psi\left(\frac{Y_i - m(x)}{\sigma(x)}\right) K\left(\frac{X_i - x}{h_n}\right),$$

$$\zeta_{in} = \frac{1}{\sqrt{h_n^d}} \chi\left(\frac{Y_i - m(x)}{\sigma(x)}\right) K\left(\frac{X_i - x}{h_n}\right).$$

By Cacoullos (1966), under (A6)–(A8),

$$(5.7) \quad \frac{1}{nh_n^d} \sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right) \rightarrow_P f(x) \int K(u) du, \quad n \rightarrow \infty.$$

We shall show now that G_n is asymptotically normal. First consider the asymptotics of $E\{G_n\}$. We have

$$E\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n \eta_{in} \right\} = \sqrt{n} E\eta_{1n}$$

$$= \frac{\sqrt{n}}{\sqrt{h_n^d}} \int \int \psi\left(\frac{m(z) - m(x) + u}{\sigma(x)}\right) K\left(\frac{z - x}{h_n}\right) dF_0\left(\frac{u}{\sigma(z)}\right) f(z) dz$$

$$= \sqrt{nh_n^d} \frac{1}{h_n^d} \int K\left(\frac{z - x}{h_n}\right) f(z) \varphi\left(\frac{m(z) - m(x)}{\sigma(x)}, \frac{\sigma(z)}{\sigma(x)}\right) dz,$$

where $\varphi(a, b) = \int \psi(a + bu) dF_0(u)$.

Let D be diameter of the set $\{z: K(z) \neq 0\}$. It suffices to consider only such z that $|z - x| \leq Dh_n$. For such z it is obvious that $|m(z) - m(x)| \leq LDh_n$ and $|\sigma(z) - \sigma(x)| \leq L'Dh_n$. Thus by continuity of $\varphi'_a(a, b)$ and $\varphi'_b(a, b)$ one obtains

$$\sup_{|z-x| \leq Dh_n} \left| \varphi\left(\frac{m(z) - m(x)}{\sigma(x)}, \frac{\sigma(z)}{\sigma(x)}\right) - \varphi(0, 1) - \varphi'_a(0, 1) \left(\frac{m(z) - m(x)}{\sigma(x)}\right) - \varphi'_b(0, 1) \left(\frac{\sigma(z)}{\sigma(x)} - 1\right) \right|$$

$$= o(h_n), \quad n \rightarrow \infty,$$

where

$$\varphi(0,1) = \int \psi(u) dF_0(u) = 0,$$

$$\varphi'_a(0,1) = \int \psi'(u) dF_0(u) = \varphi_0,$$

$$\varphi'_b(0,1) = \int u\psi'(u) dF_0(u) = 0.$$

This implies

$$(5.8) \quad \left| \sqrt{n} E\eta_{1n} - \sqrt{nh_n^d} \frac{\varphi_0}{h_n^d} \int \left(\frac{m(z) - m(x)}{\sigma(x)} \right) K\left(\frac{z-x}{h_n} \right) f(z) dz \right| \\ = o\left(h_n \sqrt{nh_n^d} \right) = o(1), \quad n \rightarrow \infty,$$

$$(5.9) \quad \lim_n \sqrt{nh_n^d} \frac{1}{h_n^d} \int (m(z) - m(x)) K\left(\frac{z-x}{h_n} \right) \mu(z) dz \\ = \beta^{d/2+1} \int m'(x; u) K(u) du \mu(x).$$

Together (5.8) and (5.9) yield

$$(5.10) \quad \lim_n \sqrt{n} E\eta_{1n} = \frac{\varphi_0}{\sigma(x)} \beta^{d/2+1} f(x) \int m'(x; u) K(u) du = b_1.$$

Let $\kappa(a, b) = \int \chi(a + bu) dF_0(u)$. Then

$$E \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n \zeta_{in} \right\} = \sqrt{n} E\zeta_{1n} \\ = \frac{\sqrt{n}}{\sqrt{h_n^d}} \int \kappa \left(\frac{m(z) - m(x)}{\sigma(x)}, \frac{\sigma(z)}{\sigma(x)} \right) K \left(\frac{z-x}{h_n} \right) f(z) dz, \\ \sup_{\{z: |z-x| \leq Dh_n\}} \left| \kappa \left(\frac{m(z) - m(x)}{\sigma(x)}, \frac{\sigma(z)}{\sigma(x)} \right) - \kappa(0,1) - \kappa'_a(0,1) \left(\frac{m(z) - m(x)}{\sigma(x)} \right) \right. \\ \left. - \kappa'_b(0,1) \left(\frac{\sigma(z)}{\sigma(x)} - 1 \right) \right| = o(h_n), \quad n \rightarrow \infty,$$

where

$$\kappa(0,1) = \int \chi(u) dF_0(u) = 0,$$

$$\kappa'_a(0,1) = \int \chi'(u) dF_0(u) = 0,$$

$$\kappa'_b(0,1) = \int u\chi'(u) dF_0(u) = \kappa_0.$$

Similarly to (5.10) one proves

$$(5.11) \quad \lim_n \sqrt{n} E \zeta_{1n} = \frac{\kappa_0}{\sigma(x)} \beta^{d/2+1} \int \sigma'(x; u) K(u) du f(x) = b_2.$$

Note that by (5.10) and (5.11),

$$(5.12) \quad E \eta_{1n} = O\left(\frac{1}{\sqrt{n}}\right), \quad E \zeta_{1n} = O\left(\frac{1}{\sqrt{n}}\right), \quad n \rightarrow \infty.$$

Consider the asymptotics of the covariance matrix of G_n . In view of (5.12),

$$(5.13) \quad \begin{aligned} \lim_n \text{var} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \eta_{in} \right) &= \lim_n \left[E \eta_{1n}^2 - (E \eta_{1n})^2 \right] = \lim_n E \eta_{1n}^2 \\ &= \lim_n \frac{1}{h_n^d} \int \int \psi^2 \left(\frac{m(z) - m(x) + u}{\sigma(x)} \right) dF_0 \left(\frac{u}{\sigma(z)} \right) \\ &\quad \times K^2 \left(\frac{z-x}{h_n} \right) f(z) dz \\ &= f(x) \varphi_2 \int K^2(u) du = \sigma_1^2. \end{aligned}$$

Similarly to (5.13),

$$(5.14) \quad \lim_n \text{var} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \zeta_{in} \right) = f(x) \kappa_2 \int K^2(u) du = \sigma_2^2.$$

The components of G_n are asymptotically uncorrelated because

$$(5.15) \quad \begin{aligned} &\lim_n \frac{1}{n} \sum_{i=1}^n E \{ (\eta_{in} - E \eta_{in}) (\zeta_{in} - E \zeta_{in}) \} \\ &= \lim_n \frac{1}{n} \sum_{i=1}^n E \{ \eta_{in} \zeta_{in} \} \\ &= \lim_n \frac{1}{h_n^d} \int \int \psi \left(\frac{m(z) - m(x) + u}{\sigma(x)} \right) \chi \left(\frac{m(z) - m(x) + u}{\sigma(x)} \right) dF_0 \left(\frac{u}{\sigma(z)} \right) \\ &\quad \times K^2 \left(\frac{z-x}{h_n} \right) f(z) dz \\ &= f(x) \int K^2(u) du \int \chi(u) \psi(u) dF_0(u) = 0 \end{aligned}$$

[recall that $\psi(u)$ is odd and $\chi(u)$ is even].

We now prove

$$(5.16) \quad G_n \rightarrow_{\mathcal{D}} \begin{pmatrix} \eta \\ \zeta \end{pmatrix}, \quad n \rightarrow \infty,$$

where $\eta \sim \mathcal{N}(b_1, \sigma_1^2)$, $\zeta \sim \mathcal{N}(b_2, \sigma_2^2)$ and $\text{cov}\{\eta, \zeta\} = 0$. In view of (5.10), (5.11),

(5.13)–(5.15) and Theorem 7.7 of Billingsley (1968) (Cramér–Wold device), it is sufficient to prove that linear combinations of components of G_n satisfy the Lyapunov condition of the central limit theorem.

Since ψ and χ are bounded we obtain

$$(5.17) \quad |\eta_{1n}|, |\zeta_{1n}| \leq \frac{C}{\sqrt{h_n^d}} K\left(\frac{X_i - x}{h_n}\right),$$

where $C > 0$ is some constant. Let $a_1, a_2 \in \mathbb{R}$ be arbitrary. The Lyapunov condition for linear combinations follows from

$$\begin{aligned} & \sum_{i=1}^n E \left(\frac{a_1}{\sqrt{n}} (\eta_{in} - E\eta_{in}) + \frac{a_2}{\sqrt{n}} (\zeta_{in} - E\zeta_{in}) \right)^4 \\ & \leq \frac{8}{n} (a_1^4 E(\eta_{1n} - E\eta_{1n})^4 + a_2^4 E(\zeta_{1n} - E\zeta_{1n})^4) \\ & \leq \frac{64}{n} (a_1^4 E\eta_{1n}^4 + a_2^4 E\zeta_{1n}^4) + O\left(\frac{1}{n^3}\right) = O\left(\frac{1}{nh_n^d}\right) = o(1), \quad n \rightarrow \infty, \end{aligned}$$

where we used (5.12), (5.17) and the elementary inequality $(a + b)^4 \leq 8(a^4 + b^4)$. This proves (5.16). The assertion of Theorem 2 follows from (5.1), (5.5)–(5.7) and (5.16).

REFERENCES

- BILLINGSLEY, R. (1968). *Convergence of Probability Measures*. Wiley, New York.
- BUSSIAN, B. and HÄRDLE, W. (1984). Robust smoothing applied to white noise and outlier contaminated Raman spectra. *Appl. Spectroscopy* **38** 309–313.
- CACOULOS, T. (1966). Estimation of a multivariate density. *Ann. Inst. Statist. Math.* **18** 178–189.
- CARROLL, R. (1982). Adapting for heteroscedasticity in linear models. *Ann. Statist.* **10** 1224–1233.
- COLLOMB, G. (1980). Estimation non paramétrique des probabilités conditionnelles. *C. R. Acad. Sci. Paris Ser. A* **291** 427–430.
- COLLOMB, G. (1981). Estimation non paramétrique de la régression: Revue bibliographique. *Internat. Statist. Rev.* **49** 75–93.
- COX, D. D. (1983). Asymptotics for M -type smoothing splines. *Ann. Statist.* **11** 530–551.
- HÄRDLE, W. (1984). Robust regression function estimation. *J. Multivariate Anal.* **14** 169–180.
- HÄRDLE, W. and GASSER, T. (1984). Robust nonparametric function fitting. *J. Roy. Statist. Soc. Ser. B* **46** 42–51.
- HÄRDLE, W. and MARRON, S. (1985). Optimal bandwidth selection for nonparametric kernel regression. *Ann. Statist.* **13** 1465–1481.
- HILDENBRAND, K. and HILDENBRAND, W. (1986). On the mean income effect: A data analysis of the UK Family Expenditure Survey. In *Contributions to Mathematical Economics* (W. Hildenbrand and A. Mas-Colell, eds.). North-Holland, Amsterdam.
- HUBER, P. (1979). Robust smoothing. In *Robustness in Statistics* (R. L. Launer and G. N. Wilkinson, eds.) 33–47. Academic, New York.
- HUBER, P. (1981). *Robust Statistics*. Wiley, New York.
- JUSTUSSON, B. I. (1981). Median filtering: Statistical properties. In *Two-Dimensional Digital Signal Processing. Transforms and Median Filters*. **2**. Springer, Berlin.
- MARONNA, R. A. (1976). Robust M -estimators of multivariate location and scatter. *Ann. Statist.* **4** 51–67.
- NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.

- NEMIROVSKIĬ, A. S., POLYAK, B. T. and TSYBAKOV, A. B. (1983). Estimators of maximum likelihood type for nonparametric regression. *Soviet Math. Dokl.* **28** 788-792.
- SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B* **47** 1-52.
- STONE, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5** 595-645.
- STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348-1360.
- TSYBAKOV, A. B. (1982a). Nonparametric signal estimation when there is incomplete information on the noise distribution. *Problemy Peredači Informacii* **18**(2) 44-60. English translation in *Problems Inform. Transmission* **18** 116-130.
- TSYBAKOV, A. B. (1982b). Robust estimates of a function. *Problemy Peredači Informacii* **18**(3) 39-52. English translation in *Problems Inform. Transmission* **18** 190-201.
- TSYBAKOV, A. B. (1983). Convergence of nonparametric robust algorithms of reconstruction of functions. *Avtomat. i Telemekh.* (12) 66-76. English translation in *Avtomat. Remote Control* **44** 1582-1591.
- WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā* **26** 359-372.

DEPARTMENT OF ECONOMICS
UNIVERSITY OF BONN
ADENAUERALLEE 24-26
D-5300 BONN 1
WEST GERMANY

INSTITUTE FOR PROBLEMS OF
INFORMATION TRANSMISSION
ACADEMY OF SCIENCES OF THE USSR
ERMOLOVOYSTR. 19
101 447 MOSCOW GSP-4
USSR

A LAW OF THE ITERATED LOGARITHM FOR NONPARAMETRIC REGRESSION FUNCTION ESTIMATORS¹

BY WOLFGANG HÄRDLE

Universität Heidelberg

We study the estimation of a regression function by two classes of estimators, the Nadaraya-Watson Kernel type estimators and the orthogonal polynomial estimators. We obtain sharp pointwise rates of strong consistency by establishing laws of the iterated logarithm for the two classes of estimators. These results parallel those of Hall (1981) on density estimation and extend those of Noda (1976) on strong consistency of kernel regression estimators.

1. Introduction and background. Let $(X, \dot{Y}), (X_i, Y_i), i = 1, 2, \dots$ be i.i.d. bivariate random variables with common joint distribution $F(x, y)$ and joint density $f(x, y)$. Let $f_X(x)$ be the marginal density of X and let $m(x) = E(Y|X = x) = \int yf(x, y) dy/f_X(x)$ be the regression of Y on X . In the present paper we obtain sharp pointwise rates of strong consistency for the following type of regression estimator

$$(1.1) \quad m_n(x) = n^{-1} \sum_{i=1}^n K_{r(n)}(x; X_i) Y_i$$

where $\{K_r; r \in I\}$ denotes a sequence of "delta functions" (or kernel sequence).

Many nonparametric estimators of $m(x)$ have this form, for instance, the Nadaraya-Watson kernel estimator (more generally estimators based on delta function sequences, as introduced by Watson and Leadbetter, 1964) or orthogonal polynomial estimators.

Nadaraya (1964) and Watson (1964) independently introduced a kernel type variant of (1.1) and demonstrated weak pointwise consistency. Rosenblatt (1969) obtained the bias, variance and asymptotic distribution of kernel type regression estimators. Schuster (1972) and Johnston (1979) demonstrated the multivariate normality at a finite number of distinct points. The strong pointwise consistency (without rates) of the Nadaraya-Watson estimator was shown by Noda (1976). For this particular kernel type estimator Collomb (1979) gave necessary and sufficient conditions on the sequence $\{K_{r(n)}\}$ for strong consistency of m_n . Stone (1977) gave general conditions on the weights $K_r(x; X_i)$ for $m_n(x)$ to be consistent in L^r , i.e. for $E|m_n(X) - m(X)|^r \rightarrow 0$. From his conditions, however, it is not clear when the Nadaraya-Watson kernel sequence is consistent (Stone, 1977, page 607).

Recently, Schuster and Yakowitz (1979) derived uniform consistency on a finite interval for a kernel type estimator. Wandl (1980) and Johnston (1982) studied the global deviation and Revesz (1979) obtained analogous results includ-

Received February 1983; revised November 1983.

¹ Research supported by the Deutsche Forschungsgemeinschaft, SFB 123 and by the Scientific Research Contract AFOSR-F49620-82-C-0009.

AMS 1980 subject classification. Primary 60F10; secondary 60G15, 62G05.

Key words and phrases. Nonparametric regression function estimation, law of the iterated logarithm, kernel estimation, orthogonal polynomial estimation.

ing nearest neighbor regression estimators. In addition, Wandl (1980) obtained rates of uniform consistency, but under the rather restrictive assumption that the marginal distribution of Y has bounded support. The assumptions in Mack and Silverman (1982), who show weak and strong uniform consistency on a bounded interval of the Nadaraya-Watson kernel estimator, are less restrictive than in Wandl (1980); the difficulties with an unbounded support of Y are overcome by a truncation argument. A similar technique, together with strong approximations of the two dimensional empirical process, will be used in the present paper. Different criteria measuring the closeness of m_n to m , including the L_1 -distance, for kernel type estimators were considered by Devroye (1978, 1981) and by Devroye and Wagner (1980a, b).

The method of orthogonal polynomial estimation was originally introduced by Čencov (1962) for density estimation. Rutkowski (1982a, b) defined a regression estimator based on orthogonal polynomials in the case of fixed design variables X . He also presented conditions for (weak) consistency and discussed the applications of such estimators to a broad class of system identification problems. For more work and related problems concerning both kernel type and orthogonal polynomial type estimators, we refer to the review article of Collomb (1981).

In the present paper we show a law of the iterated logarithm for the centered estimate

$$(1.2) \quad m_n(x) - Em_n(x).$$

This result thus gives the exact order of convergence of $m_n(x) - Em_n(x)$. For statistical interpretations it is desirable to have exact pointwise strong convergence rates for $m_n(x) - m(x)$, but since the bias is purely analytically handled, it suffices to consider (1.2). The handling with the bias terms using different smoothness assumptions on m and K_r , is delayed to the sections where we apply the general result of Section 2. In Section 4 we show a law of the iterated logarithm for the Nadaraya-Watson kernel type estimator and for a related estimator that is useful if we know the marginal density f_X of X . In Section 5 we derive an analogous result for estimators based on orthogonal polynomials.

As a footnote, we would like to mention some related works on density estimation. These include among others Wegman and Davies (1979), Hall (1981), Stute (1982).

2. A law of the iterated logarithm for a special triangular array. Let $(X_1, Y_1), (X_2, Y_2), \dots$ be a sequence of independent and identically distributed random variables with probability density function $f(x, y)$ and cumulative distribution function $F(x, y)$ and $EY^2 < \infty$. As in (1.1), let $\{K_r: r \in I\}$ be a sequence of real valued functions each of bounded variation and define

$$S_n(r) = \sum_{i=1}^n \{K_r(X_i)Y_i - E[K_r(X_i)Y_i]\}.$$

Note that this sum is a multiple of (1.2), and that we omitted the dependence on the design point x for convenience. Define also

$$\sigma(r, s) = \text{cov}\{K_r(X)Y, K_s(X)Y\} \quad \text{and} \quad \sigma^2(r) = \sigma(r, r).$$

We will now establish conditions under which $S_n(r) = n[m_n(x) - Em_n(x)]$ follows

the law of the iterated logarithm. We demonstrate that

$$\limsup_{n \rightarrow \infty} \pm [\phi(n)]^{-1} S_n(r(n)) = 1 \quad \text{a.s.},$$

where $\phi(n) = (2n\sigma^2(r)\log \log n)^{1/2}$. An application of this result to the two classes of nonparametric regression function estimators, to be defined below, provides thus a precise description of the order of strong consistency of $m_n(x)$.

The set $\{S_n(r), n \geq 1\}$ is a triangular sequence and in this section it is seen that S_n may be strongly approximated by a Gaussian sequence with the same covariance structure. The law of the iterated logarithm will then be shown using parallel results on density estimation by Hall (1981) and Csörgő and Hall (1982). We shall also make use of the Rosenblatt transformation (Rosenblatt, 1952)

$$T(x, y) = (F_{Y|X}, F_X)(x, y),$$

transforming the original data points $\{(X_i, Y_i)\}_{i=1}^n$ into a sequence of mutually independent uniformly distributed over $[0, 1]^2$ random variables $\{(X'_i, Y'_i)\}_{i=1}^n$. This transformation was also employed by Johnston (1982) as an intermediate tool; also by Mack and Silverman (1982) to obtain (strong) uniform consistency of the Nadaraya-Watson kernel type regression function estimates. It will be convenient to define

$$v_n(u_n) = \int_{|x| \leq u_n} |dK_{r(n)}(x)| + |K_{r(n)}(-u_n -)|, \quad n \geq 1$$

with a sequence of constants $\{u_n\}$, $0 < u_n \leq \infty$.

THEOREM 1. *Suppose that the sequence of kernels $K_{r(n)}$ and $\{u_n\}$ satisfy*

$$(2.1) \quad a_n v_n(u_n) = o(n^{1/2} \sigma(r) (\log \log n)^{1/2} / (\log n)^2),$$

where $\{a_n\}$ is a sequence of positive constants tending to infinity. In addition, assume that the following holds.

$$(2.2a) \quad \sum_{n=3}^{\infty} E\{K_r^2(X)I(|X| > u_n)\} / (\sigma^2(r)\log \log n) < \infty$$

$$(2.2b) \quad \sum_{n=3}^{\infty} E\{K_r^2(X)I(|X| \leq u_n)Y^2I(|Y| > a_n)\} / (\sigma^2(r)\log \log n) < \infty.$$

Then on a rich enough probability space there exists a Gaussian sequence $\{T_n\}$ with zero means and the same covariance structure as $\{S_n(r)\}$, such that

$$S_n(r) - T_n = o(n^{1/2} \sigma(r) (\log \log n)^{1/2}) \quad \text{a.s.}$$

The device that is used in the proof is the strong uniform approximation of the empirical process by a Brownian bridge. Hall (1981) employs for density estimation in the one dimensional case the results of Komlós, Major and Tusnády (1975). As in Mack and Silverman (1982), we will make use of an analogous result by Tusnády (1977) for the two dimensional case. Note that although the two dimensional case is considered here, the technique can be extended to higher dimensional design variables $\mathbf{X} = (x^{(1)}, \dots, x^{(d)})$, $d \geq 2$. The assumption, however,

will not be compatible with the case considered here since it is still unknown whether the strong approximation of the multivariate empirical process by a multivariate Brownian bridge has a compatible rate as in the one- or two-dimensional case.

The fundamental connection between $S_n(r)$ and its strong approximation by a Gaussian sequence is established by the following lemma. The proof will be clear from Tusnády (1977) and the fact that $n^{1/2}[F_n(T^{-1}(x, y')) - F(T^{-1}(x', y'))]$, $(x', y') \in [0, 1]^2$ is the empirical process of $\{(X_i, Y_i)\}_{i=1}^n$ (Rosenblatt, 1952).

LEMMA 1. *On a rich enough probability space there is a version of a Brownian bridge $B(x', y')$, $(x', y') \in [0, 1]^2$ such that*

$$P\{\sup_{x,y} |e_n(x, y)| > (C_1 \log n + u) \log n\} < C_2 e^{-C_3 u}$$

where C_1, C_2, C_3 are absolute constants and

$$e_n(x, y) = n[F_n(x, y) - F(x, y)] - n^{1/2}B(T(x, y)).$$

In the following theorem it is now seen that under regularity conditions on the covariances $\sigma(r, s)$ a law of the iterated logarithm (LIL) holds for $m_n(x)$ as defined in (1.1)

THEOREM 2. *Suppose that (2.1) and (2.2a, b) hold and that*

$$(2.3) \quad \lim_{\epsilon \rightarrow 0} \limsup_{n \rightarrow \infty} \sup_{m \in \Gamma_{n,\epsilon}} |\sigma(r(m), r(n)) / \sigma^2(r(n)) - 1| = 0,$$

where $\Gamma_{n,\epsilon} = \{m: |m - n| \leq \epsilon n\}$. Then

$$\limsup_{n \rightarrow \infty} \pm [\phi(n)]^{-1} S_n(r) = 1 \quad \text{a.s.}$$

3. Proofs. To establish Theorem 1 we set

$$T_n = n^{1/2} \int \int_{-\infty}^{\infty} K_r(x)y \, dB(T(x, y)),$$

$B(x', y')$ being the Brownian Bridge of Lemma 1, and show that the difference

$$R_n = n^{-1}(S_n(r) - T_n) = n^{-1} \int \int K_r(x)y \, de_n(x, y)$$

satisfies

$$(3.1) \quad R_n = o(n^{-1/2} \sigma(r) (\log \log n)^{1/2}) \quad \text{a.s.}$$

Note first that T_n has the covariance structure ascribed to it in Theorem 1. This follows from the fact that the Jacobian $J(x, y)$ of $T(x, y)$ is $J(x, y) = f(x, y)$, the joint density of (X, Y) (Rosenblatt, 1952) and the following lemma, stated without proof.

LEMMA 2. Let $G_r(x, y) = K_r(x)y$. Then

$$(Z_1, Z_2) = \left(\int_0^1 \int_0^1 G_{r_1}(T^{-1}(x', y')) dB(x', y'), \right. \\ \left. \int_0^1 \int_0^1 G_{r_2}(T^{-1}(x', y')) dB(x', y') \right)$$

has a bivariate normal distribution with zero means and covariances

$$\text{cov}(Z_1, Z_2) = \int \int K_{r_1}(x)K_{r_2}(x)y^2f(x, y) dx dy \\ - \left[\int \int K_{r_1}(x)yf(x, y) dx dy \right] \left[\int \int K_{r_2}(x)yf(x, y) dx dy \right] \\ = \sigma(r_1, r_2).$$

To demonstrate (3.1), we split up the integration regions and obtain

$$|R_n| \leq \sum_{j=1}^7 R_{j,n}$$

where

$$R_{1,n} = \left| n^{-1} \int_{|x| \leq u_n} \int_{|y| \leq a_n} K_r(x)y de_n(x, y) \right| \\ \leq 2v_n(u_n)a_n n^{-1} \sup_{x,y} |e_n(x, y)|, \\ R_{2,n} = |n^{-1} \sum_{i=1}^n R_{i,n}^{(2)}|, \\ R_{i,n}^{(2)} = [K_r(X_i)I(|X_i| > u_n)Y_iI(|Y_i| \leq a_n)] \\ - E[K_r(X)I(|X| > u_n)YI(|Y| \leq a_n)] \\ R_{3,n} = |n^{-1} \sum_{i=1}^n R_{i,n}^{(3)}|, \\ R_{i,n}^{(3)} = [K_r(X_i)I(|X_i| \leq u_n)Y_iI(|Y_i| > a_n)] \\ - E[K_r(X)I(|X| \leq u_n)YI(|Y| > a_n)] \\ R_{4,n} = |n^{-1} \sum_{i=1}^n R_{i,n}^{(4)}|, \\ R_{i,n}^{(4)} = [K_r(X_i)I(|X_i| > u_n)Y_iI(|Y_i| > a_n)] \\ - E[K_r(X)I(|X| > u_n)YI(|Y| > a_n)], \\ R_{5,n} = n^{-1} \left| \int_{|x| > u_n} \int_{|y| \leq a_n} K_r(x)y dB(T(x, y)) \right|, \\ R_{6,n} = n^{-1} \left| \int_{|x| \leq u_n} \int_{|y| > a_n} K_r(x)y dB(T(x, y)) \right|, \\ R_{7,n} = n^{-1} \left| \int_{|x| > u_n} \int_{|y| > a_n} K_r(x)y dB(T(x, y)) \right|.$$

From Lemma 1 we deduce that $n^{-1} \sup_{x,y} |e_n(x, y)| = O(n^{-1}(\log n)^2)$ a.s., and so by condition (2.1) we conclude that

$$(3.2) \quad R_{1,n} = o(n^{-1/2} \sigma(r)(\log \log n)^{1/2}) \text{ a.s.}$$

Next observe that $\{R_{i,n}^{(2)}\} 1 \leq i \leq n$ are independent and identically distributed random variables. We then have by Markov's inequality that for any $\epsilon > 0$

$$P(n^{-1} |\sum_{i=1}^n R_{i,n}^{(2)}| > \epsilon \cdot \sigma(r)n^{-1/2} \cdot (\log \log n)^{1/2}) \leq \epsilon^{-2} \sigma(r)^{-2} (\log \log n)^{-1} \cdot E(R_{1,n}^{(2)})^2.$$

So with the assumption $EY^2 < \infty$ and condition (2.2 a) it follows with the Borel-Cantelli Lemma that

$$(3.3) \quad R_{2,n} = o(n^{-1/2} \sigma(r)(\log \log n)^{1/2}) \text{ a.s.}$$

The terms $R_{3,n}, R_{4,n}$ may be estimated in the same way using Markov's inequality and condition (2.2b) and we therefore have

$$(3.4) \quad \begin{aligned} R_{3,n} &= o(n^{-1/2} \sigma(r)(\log \log n)^{1/2}) \text{ a.s.} \\ R_{4,n} &= o(n^{-1/2} \sigma(r)(\log \log n)^{1/2}) \text{ a.s.} \end{aligned}$$

The remaining terms, $R_{5,n}, R_{6,n}$ and $R_{7,n}$ are all Gaussian with mean zero and standard deviations

$$\{E(R_{1,n}^{(2)})^2\}^{1/2}, \quad \{E(R_{1,n}^{(3)})^2\}^{1/2}, \quad \{E(R_{1,n}^{(4)})^2\}^{1/2}$$

respectively. Therefore, $R_{5,n}$, for instance, can be computed by

$$\begin{aligned} P(R_{5,n} > \epsilon n^{-1/2} \sigma(r)(\log \log n)^{1/2}) \\ = 2[1 - \Phi\{\epsilon \sigma(r)(\log \log n)^{1/2} / [E(R_{1,n}^{(2)})^2]^{1/2}\}], \end{aligned}$$

where Φ denotes the cdf of the standard normal distribution. A similar equality holds for $R_{6,n}$ and $R_{7,n}$; therefore, we conclude in view of condition (2.2a, b) and the usual approximations to the tails of the normal distribution that

$$(3.5) \quad \begin{aligned} R_{5,n} &= o(n^{-1/2} \sigma(r)(\log \log n)^{1/2}) \text{ a.s.} \\ R_{6,n} &= o(n^{-1/2} \sigma(r)(\log \log n)^{1/2}) \text{ a.s.} \\ R_{7,n} &= o(n^{-1/2} \sigma(r)(\log \log n)^{1/2}) \text{ a.s.} \end{aligned}$$

Theorem 1 follows now by putting together statements (3.2)–(3.5) respectively.

The proof of Theorem 2 follows in the same way as the proof of Theorem 1 in Hall (1981, page 49). We only have to note that Lemma 1 in Hall (1981, page 49) has to be replaced by (2.3).

4. Kernel estimators. Two types of kernel estimates of the regression function $m(x)$ will be considered here. The first is due to Nadaraya (1964) and Watson (1964):

$$m_n^*(x) = (nh)^{-1} \sum_{i=1}^n K((x - X_i)/h) Y_i / [(nh)^{-1} \sum_{i=1}^n K((x - X_i)/h)].$$

We may think of applications where the marginal density f_X of X is known to

the statistician. It is then appropriate to replace the density estimator in the denominator of m_n^* by the known density f_X . This leads to the following estimate:

$$\bar{m}_n(x) = (nh)^{-1} \sum_{i=1}^n K((x - X_i)/h) Y_i / f_X(x)$$

considered by Johnston (1979, 1982).

Let us define $S^2(x) = E(Y^2 | X = x)$, $V^2(x) = S^2(x) - m^2(x)$, and assume that $f_X(x)$, $m(x)$ are twice differentiable and $S^2(x)$ is continuous. We assume further that the kernel $K(\cdot)$ is continuous, has compact support $(-1, 1)$ and that $\int_{-1}^1 uK(u) du = 0$. This implies that $v_n(u_n)$ as defined in (2.1) is constant for large u_n . We will make use of the following assumptions:

(4.1) $nh^5 / \log \log n \rightarrow 0$ as $n \rightarrow \infty$

(4.2) $\sum_{n=3}^{\infty} (h / \log \log n) E[Y^2 I(|Y| > a_n)] < \infty$

where $\{a_n\}$ is as in (2.1), (2.2 a, b) such that

(4.3) $a_n = o((nh^{-1} \log \log n)^{1/2} / (\log n)^2)$
 $\lim_{\epsilon \rightarrow 0} \limsup_{n \rightarrow \infty} \sup_{m \in \Gamma_{n,\epsilon}} |h(m)/h(n) - 1| = 0$.

We then have the following theorem for $\bar{m}_n(x)$.

THEOREM 3. *Under the assumptions above*

$$\begin{aligned} \limsup_{n \rightarrow \infty} \pm [\bar{m}_n(x) - m(x)](nh/2 \log \log n)^{1/2} \\ = [S^2(x) \int K^2(u) du / f_X(x)]^{1/2} \quad \text{a.s.} \end{aligned}$$

The Nadaraya-Watson estimate follows also a LIL.

THEOREM 4. *Under the assumptions above and $\sum_{n=1}^{\infty} n^{-2} h^{-1} < \infty$*

$$\begin{aligned} \limsup_{n \rightarrow \infty} \pm [m_n^*(x) - m(x)](nh/2 \log \log n)^{1/2} \\ = [V^2(x) \int K^2(u) du / f_X(x)]^{1/2} \quad \text{a.s.} \end{aligned}$$

Note that the only difference between Theorem 3 and Theorem 4 is the different scaling factor. As was shown by Johnston (1979), $\bar{m}_n(x)$ has asymptotic variance proportional to $S^2(x)$, whereas $m_n^*(x)$ has asymptotic variance $\sim V^2(x)$. Since in general $S^2(x) \geq V^2(x)$, we expect therefore closer asymptotic confidence intervals for $m_n^*(x)$ than for $\bar{m}_n(x)$.

PROOF OF THEOREM 3. We first show that we could center $\bar{m}_n(x)$ around $E\bar{m}_n(x)$. This follows from

$$E\bar{m}_n(x) = f_X(x)^{-1} h^{-1} \int K((x - u)/h) m(u) f_X(u) du = m(x) + O(h^2)$$

using the smoothness of $m(\cdot)$ and $f_X(\cdot)$ and the assumptions on the kernel $K(\cdot)$ (Parzen, 1962; Rosenblatt, 1971).

From assumption (4.1) it thus follows that the bias term $(E\hat{m}_n(x) - m(x))$ vanishes of higher order. So it remains to show that

$$(4.4) \quad \limsup_{n \rightarrow \infty} \pm [\hat{m}_n(x) - E\hat{m}_n(x)] / (nh^2 \log \log n)^{1/2} = [S^2(x) \cdot f_X(x) \int K^2(u) du]^{1/2} \text{ a.s.}$$

where $\hat{m}_n(x) = \sum_{i=1}^n K((x - X_i)/h) Y_i = \sum_{i=1}^n K_h(X_i) Y_i$.

From the assumptions on the kernel $K(\cdot)$ we conclude that $\delta_n(u) = h^{-1}K(u/h)$ is a delta function sequence (DFS) in the sense of Watson and Leadbetter (1964). We now make use of this general approach in terms of DFS's and obtain the following:

$$h\sigma^2(h) = h \int \delta_n^2(x - u) S^2(u) f_X(u) du - h \left[\int \delta_n(x - u) m(u) f_X(u) du \right]^2 \rightarrow S^2(x) \cdot f_X(x) \int K^2(u) du \text{ as } n \rightarrow \infty.$$

This follows from Watson and Leadbetter (1964) by noting that $S^2(\cdot) f_X(\cdot)$ is continuous and $\{h(\int K^2)^{-1} \delta_n^2(u)\}$ is itself a DFS. We may note that the use of this DFS-technique would also provide a slight simplification of Hall's proof (1981) for Rosenblatt-Parzen kernel density estimates.

To establish (4.4) with the use of Theorem 2, we have to show that (2.3) holds. We thus have to demonstrate that if $h, k \rightarrow 0$ such that $h/k \rightarrow 1$ (in view of assumption (4.3)), then

$$(4.5) \quad h^{-1} \text{cov}\{K((x - X)/h) Y, K((x - Y)/k) Y\} \rightarrow 1.$$

But $EK((x - X)/h) Y = h \int \delta_n(x - u) m(u) \cdot f_X(u) du = o(h^{1/2})$, and so by the computations for $\sigma^2(h)$ above it remains to demonstrate that

$$h^{-1} \int [K((x - u)/h) - K((x - u)/k)]^2 S^2(u) f_X(u) du \rightarrow 0.$$

From the boundedness of $S^2(\cdot)$ and $f_X(\cdot)$ it is clear that the integral above is dominated by

$$M \int [K(u) - K(uh/k)]^2 du.$$

The kernel K is continuous and so $K(uh/k) \rightarrow K(u)$ a.e. and it follows that (4.5) holds.

Assumption (2.1) follows from (4.2) since $K(\cdot)$ has compact support and thus $v_n(u_n) = \text{const.}$ for n large enough. In view of the asymptotic formula for $\sigma^2(h)$ above we have by assumption (4.2)

$$a_n = o((n\sigma^2(h) \log \log n)^{1/2} / (\log n)^2)$$

which is assumption (2.1). Finally, assumptions (2.2a, b) follow immediately from (4.2) since K has compact support and as above $\sigma^2(h) \sim h^{-1}$. Theorem 3 thus follows from Theorem 2.

PROOF OF THEOREM 4. To prove Theorem 4 we decompose

$$m_n^*(x) - m(x) = [(nh)^{-1}\hat{m}_n(x) - m(x)f_n(x)]/f_X(x) + f_X^{-1}(x)[m_n^*(x) - m(x)] \cdot [f_X(x) - f_n(x)]$$

where $f_n(x) = (nh)^{-1} \sum_{i=1}^n K((x - X_i)/h)$ is a density estimate of $f_X(x)$. Now from Hall (1981), Theorem 2 it follows that

$$(4.6) \quad \lim \sup_{n \rightarrow \infty} \pm [f_n(x) - f_X(x)](nh/2 \log \log n)^{1/2} = [f_X(x) \int K^2(u) du]^{1/2} \text{ a.s.}$$

if we use assumption (4.1), ensuring that the bias $(Ef_n(x) - f_X(x)) = O(h^2)$. From Noda (1976) we conclude that with $\sum n^{-2}h^{-1} < \infty$, $m_n^*(x) - m(x) = o(1)$ a.s. This and (4.6) thus yield that the second term on the RHS of the decomposition above is of order $o((nh/2 \log \log n)^{1/2})$ a.s.

The first summand of the decomposition above can be written as

$$\frac{(nh)^{-1}(\hat{m} - E\hat{m})}{f_X} + \frac{(nh)^{-1}E\hat{m} - mf_X}{f_X} - \frac{m(f_n - Ef_n)}{f_X} + \frac{m(f_X - Ef_n)}{f_X}.$$

As in the proof of Theorem 3, it follows by assumption (4.1) that the bias terms $((nh)^{-1}E\hat{m} - mf_X)$ and $(Ef_n - f_X)$ vanish. It remains to show

$$(4.7) \quad (nh)^{-1}(\hat{m} - E\hat{m}) - m(f_n - Ef_n)$$

follows the LIL, i.e.

$$\lim \sup_{n \rightarrow \infty} \pm [(nh)^{-1}(\hat{m} - E\hat{m}) - m(f_n - Ef_n)](nh/2 \log \log n)^{1/2} = [V^2(x) \cdot f_X(x) \cdot \int K^2(u) du]^{1/2} \text{ a.s.}$$

This can be deduced from Theorem 2, if we rewrite (4.7) as

$$(nh)^{-1} \sum_{i=1}^n [K_h(X_i)Y_i - EK_h(X)Y] - m(x)(nh)^{-1} \sum_{i=1}^n [K_h(X_i) - EK_h(X)] = (nh)^{-1} \sum_{i=1}^n \{K_h(X_i)[Y_i - m(x)] - EK_h(X)[Y - m(x)]\}.$$

Next we show that (4.3) holds. The variance for the sequence above is now:

$$h \cdot \sigma^2(h) = h \cdot \int_n^2 \delta_n^2(x - u)[S^2(u) - m^2(x)]f_X(u) du - h \left[\int \delta_n(x - u)[m(u) - m(x)]f_X(u) du \right]^2 \rightarrow V^2(x) \cdot f_X(x) \int K^2(u) du \text{ as } n \rightarrow \infty.$$

As above in the proof of Theorem 3, we conclude that (2.3) holds. Theorem 4 thus follows from Theorem 2.

5. Orthogonal polynomial estimators. Estimators of the regression function $m(x)$ based on orthogonal polynomials fit also in the general framework developed in the first section. We define the estimate based on a system of orthonormal polynomials on $[-1, 1]$ as follows:

$$\hat{m}_n(x) = n^{-1} \sum_{i=1}^n K_m(x; X_i) Y_i / n^{-1} \sum_{i=1}^n K_m(x; X_i)$$

where $m = m(n)$ tends with n to infinity and

$$K_m(x; X_i) = \sum_{j=0}^m e_j(x) e_j(X_i)$$

and $\{e_j(\cdot)\}$ is the orthonormal system of polynomials. As a technical more tractable estimator we consider also:

$$m'_n(x) = n^{-1} \sum_{i=1}^n K_m(x; X_i) Y_i / f_X(x).$$

As in Section 4, let $S^2(x)$ be the second conditional moment of Y and $V^2(x)$ the conditional variance respectively. We further assume that

$$\begin{aligned} f_X(x) \text{ has compact support in } (-1, 1) \\ (1 - x^2)^{-1/4} f_X(x) \text{ is integrable on } (-1, 1). \end{aligned}$$

For reasons of simplicity we only consider the case of $e_j(\cdot) = p_j(\cdot) =$ orthonormal Legendre polynomials here and assume that the following holds:

$$\begin{aligned} (5.1) \quad \lim_{\epsilon \rightarrow 0} \limsup_{n \rightarrow \infty} \sup_{p \in \Gamma_{n,\epsilon}} |m(p)/m(n) - 1| &= 0 \\ (5.2) \quad \sum_{n=3}^{\infty} m^{-1} \cdot (\log \log n)^{-1} E(Y^2 \cdot I(|Y| > a_n)) &< \infty, \end{aligned}$$

when $\{a_n\}$ is as in (2.2), (4.2) a sequence of constants tending to infinity such that

$$\begin{aligned} (5.3) \quad a_n &= o(n^{1/2} m (\log \log n)^{1/2} / (\log n)^2). \\ n / (m^5 \log \log n) &\rightarrow 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

We have then the following Theorem for $m'_n(x)$ and $\hat{m}_n(x)$.

THEOREM 5. *Under the assumptions above*

$$\begin{aligned} \limsup_{n \rightarrow \infty} \pm [m'_n(x) - m(x)] / (n/2m \log \log n)^{1/2} \\ = [S^2(x) / (f_X(x) \cdot \pi)]^{1/2} (1 - x^2)^{-1/4} \quad \text{a.s.} \end{aligned}$$

and

$$\begin{aligned} \limsup_{n \rightarrow \infty} \pm [\hat{m}_n(x) - m(x)] / (n/2m \log \log n)^{1/2} \\ = [V^2(x) / (f_X(x) \cdot \pi)]^{1/2} (1 - x^2)^{-1/4} \quad \text{a.s.} \end{aligned}$$

PROOF. We first show the LIL for $m'_n(x)$. The second assertion will then follow as Theorem 4 from Theorem 3. As in Theorem 3, we show first that the

bias $(Em'_n(x) - m(x))$ is negligible.

$$\begin{aligned} Em'_n(x) &= [f_X(x)]^{-1} \cdot EK_m(x; X)Y \\ &= [f_X(x)]^{-1} \int K_m(x; u)m(u)f_X(u)d \\ &= m(x) + O(m^{-2}) \end{aligned}$$

by a slight modification of the argument proving Theorem 1 in Walter and Blum (1979). By the same arguments as in Hall's (1981) proof of his Theorem 3 (page 60) we conclude that

$$\sigma_m^2 \sim E[K_m^2(x; X)Y^2] \sim m \cdot S^2(x)/([f_X(x)\pi](1 - x^2)^{1/2}).$$

Assumption (2.1) follows now from (5.2) and

$$\int |dK_m(x; u)| = O(m^2).$$

Assumption (2.2) follows also from (5.2) so we finally derive the desired result from Theorem 2, since (2.3) may be proved as in Theorem 3 using (5.1).

REMARK. There is a wide variety of density estimators based on trigonometric series or Fourier transforms. In the same way as orthogonal polynomial regression estimators are deduced from orthogonal polynomial density estimators, one may construct regression estimators based on trigonometric series. It may be possible to show a law of the iterated logarithm for trigonometric series estimators, but as is indicated in Hall (1981) the computations may be more tedious than for the two classes that are considered here.

Acknowledgement. The author wishes to thank R. Carroll, S. Cambanis and an anonymous referee for helpful suggestions and remarks.

REFERENCES

- ČENCOV, N. N. (1962). Evaluation of an unknown distribution density from observations. *Soviet Math.* **3** 1559-1562.
- COLLOMB, G. (1979). Conditions nécessaires et suffisantes de convergence uniform d'un estimateur de la régression, estimation des dérivées de la régression. *C.R. Acad. Sci. Paris* **288** 161-164.
- COLLOMB, G. (1981). Estimation non-paramétrique de la Régression: Revue Bibliographique. *Internat. Statist. Rev.* **49** 75-93.
- CSÖRGŐ, S. and HALL, P. (1982). Upper and lower classes for triangular arrays. *Z. Wahrsch. verw. Gebiete* **61** 207-222.
- DEVROYE, L. P. (1978). The uniform convergence of the Nadaraya-Watson regression function estimate. *Can. J. Statist.* **6** 179-191.
- DEVROYE, L. P. (1981). On the almost everywhere convergence of nonparametric regression function estimates. *Ann. Statist.* **9** 1310-1319.
- DEVROYE, L. P. and WAGNER, T. J. (1980a). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Ann. Statist.* **8** 231-239.
- DEVROYE, L. P. and WAGNER, T. J. (1980b). On the L_1 convergence of kernel estimators of regression functions with applications in discrimination. *Z. Wahrsch. verw. Gebiete* **51** 15-25.

(1984) Härdle, W. A Law of Iterated Logarithm for Nonparametric Regression Function Estimators

- HALL, P. (1981). Laws of the iterated logarithm for nonparametric density estimators. *Z. Wahrsch. verw. Gebiete* **56** 47-61.
- JOHNSTON, G. (1979). Smooth nonparametric regression analysis. Inst. of Stat. Mimeo Series No. 1253, University of North Carolina.
- JOHNSTON, G. (1982). Probabilities of maximal deviation of nonparametric regression function estimation. *J. Multivariate Anal.* **12** 402-414.
- KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1975). An approximation of partial sums of independent rv's and the sample df I. *Z. Wahrsch. verw. Gebiete* **32** 111-131.
- MACK, Y. P. and SILVERMAN, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. verw. Gebiete* **61** 405-415.
- NADARAYA, E. A. (1964). On estimating regression. *Theor. Probab. Appl.* **9** 141-142.
- NODA, K. (1976). Estimation of a regression function by the Parzen kernel type density estimators. *Ann. Inst. Math. Statist.* **28** 221-234.
- PARZEN, E. (1962). On estimation of a probability density function. *Ann. Math. Statist.* **33** 1065-1076.
- ROSENBLATT, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Statist.* **23** 470-472.
- ROSENBLATT, M. (1969). Conditional probability density and regression estimates. In *Multivariate Analysis II*. 23-51. Ed. Krishnaiah.
- RÉVÉSZ, P. (1979). On the nonparametric estimation of the regression function. *Prob. Control. Inform. Theory* **8** 297-302.
- RUTKOWSKI, L. (1982a). On system identification by nonparametric function fitting. *IEEE Trans. Int. Control* **27** 225-227.
- RUTKOWSKI, L. (1982b). On-line identification of time-varying systems by nonparametric techniques. *IEEE Trans. Int. Control* **27** 228-230.
- SCHUSTER, E. F. (1972). Joint asymptotic distribution of the estimated regression function at a finite number of distinct points. *Ann. Math. Statist.* **43** 84-88.
- SCHUSTER, E. F. and YAKOWITZ, S. (1979). Contributions to the theory of nonparametric regression, with application to system identification. *Ann. Statist.* **7** 139-149.
- STONE, C. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595-645.
- STUTE, W. (1982). A law of logarithm for kernel density estimators. *Ann. Probab.* **10** 414-422.
- TUSNÁDY, G. (1977). A remark on the approximation of the sample df in the multidimensional case. *Period. Math. Hung.* **8** 53-55.
- WALTER, G. and BLUM, J. (1979). Probability density estimation using delta sequences. *Ann. Statist.* **7** 328-340.
- WANDL, H. (1980). On kernel estimation of regression functions. *Wiss. Sitz. zur Stochastik. WSS-03* 1-25.
- WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā A* **26** 359-372.
- WATSON, G. S. and LEADBETTER, M. R. (1964). Hazard analysis II. *Sankhyā A* **26** 101-116.
- WEGMAN, E. J. and DAVIES, H. I. (1979). Remarks on some recursive estimators of a probability density. *Ann. Statist.* **7** 316-327.

UNIVERSITÄT HEIDELBERG
 SONDERFORSCHUNGSBEREICH 123
 IM NEUENHEIMER FELD 293
 D-6900 HEIDELBERG 1
 WEST GERMANY

Robust Smoothing Applied to White Noise and Single Outlier Contaminated Raman Spectra

BERND-M. BUSSIAN* and WOLFGANG HÄRDLE

Anorganisch-Chemisches Institut, Universität Heidelberg, Im Neuenheimer Feld 270, 6900 Heidelberg 1, Federal Republic of Germany (B.-M.B.) and Institut für angewandte Mathematik, Sonderforschungsbereich 123, Universität Heidelberg, Im Neuenheimer Feld 293, 6900 Heidelberg 1, Federal Republic of Germany (W.H.)

There are several smoothing procedures for spectral data which are affected by occasionally occurring outliers. Most of the known methods are based on local averages (or fits) of the spectral data. We introduce here an outlier-insensitive, robust smoothing method which rejects the influence of huge noise spikes. The proposed smoothing algorithm can be tuned by two parameters. The first corresponds to the signal-to-noise ratio, the second to the halfwidths of the spectral bands. We apply this new technique to several spectra and prove the advantages of our method of identifying peaks and baselines in Raman spectroscopy.

Index Headings: Raman spectroscopy; Noise reduction; Robust smoothing; Non-linear filtering.

INTRODUCTION

Noise always accompanies the recording and evaluation of spectra and thus introduces a lot of difficulty into

the identifying of certain elements of the spectra. In many cases, smoothing of the experimental data is necessary because of the unfavorable signal-to-noise ratio caused by, for instance, low concentrations of the sample in solution. Smoothing is also useful whenever parameters of bands with low intensity have to be determined exactly.

If we are interested in shape and/or location of a spectral band we have to suppress the "noise part" of our data. Mathematically speaking, this is the same as "estimating a curve" or "smoothing contaminated data." Smoothing of the raw data is recommended, especially in the three following cases:

1. The interpretation of spectra of highly diluted samples in solutions leads to the problem of identifying bands buried in the noise or superimposed by solvent bands. In this case we emphasize the need of Raman difference spectroscopy, described by Laane and Kiefer.¹ This subtraction is valid only when the experimental parameters are set equal for both spectra. Since the noise adds to

Received 24 January 1983; revision received 20 July 1983.

* Present address: Department of Chemistry, University of Oregon, Eugene, OR 97403-1210.

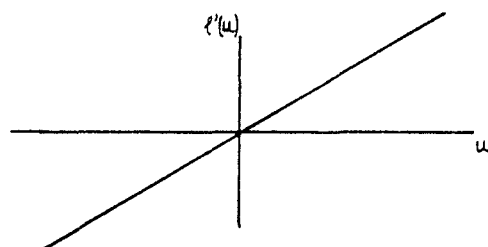


Fig. 1

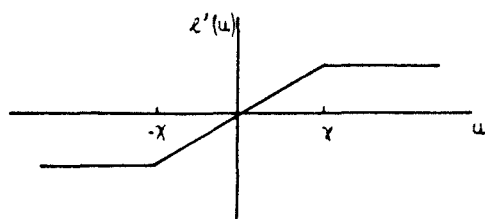


Fig. 2

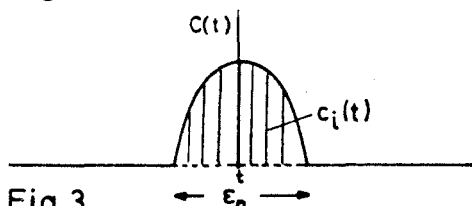


Fig. 3

FIG. 1. Influence curve for LS-methods.

FIG. 2. Bounded influence curve.

FIG. 3. Weighting function $c_i(t)$.

the signal, one can surely not expect that the noise cancels out by subtraction of spectra.

2. The determination of intensities, the exact calculation of the depolarization ratio, and the determination of the differential profile of ρ over the band need spectra with high signal-to-noise ratios, because even small uncertainties in the intensity of the perpendicular spectra lead to large errors in ρ .

3. Smoothed spectra are necessary when band shape analyses such as the determination of Gaussian and Lorentzian contributions to the profile or separation of overlapping bands are studied. Gans *et al.*² propose a manual guess of the parameters and carry out the separation of the bands on a graphical display. This method may be advantageous because the experimenter gets a feeling for the spectra. Yet strongly contaminated spectra may lead to a wrong choice of parameters since they are subjectively estimated.

In the last few years several papers on smoothing have been published which are all based on the method of "least-squares" fitting.²⁻⁸ We will explain why the "least-square techniques" are necessarily sensitive to single outliers and may therefore lead to wrong conclusions if, for example, we are concerned with the determination of peak height and bandwidth.

In the second part of this paper, we describe the origin and properties of the noise, which demands a specific mathematical model, developed in the third section. In the last section we present our results and apply the robust smoothing algorithm to very strongly contaminated raw data, as published recently by Hillig and Morris.⁹

SOURCE AND PROPERTIES OF INSTRUMENTATION NOISE

The amplitude of noise and its statistical behavior depend on the source of the noise. Most of the recording techniques in use cause noise generation. According to the source of the instrumentation noise, we classify it into two groups.

One group is the so-called white noise which is statistically distributed around the true signal. The amplitude of the white noise can be influenced by the time constant of the amplifier circuitry. The white noise arises partly from the electronic equipment and partly from the dark noise of the photomultiplier. The latter is the dominating noise source.

Besides this kind of noise, there exists a type of noise which is caused by random external events such as high frequency signals, bubbles in the sample by which scattering is possible, or shock-waves which occur within the optical path. Also, errors in data handling, such as misprints or punching errors of perforated tapes, may introduce huge, absurd spikes. We call this type of noise pattern a "single spike outlier."

Obviously the presence of such a single spike outlier causes difficulties in the smoothing of spectral data. Outlying spikes near a spectral band should not be included in a smoothing procedure. In the following section, a mathematical procedure is developed which is adapted to the twofold noise pattern described above.

MATHEMATICAL CONSIDERATION

In this section we explain how large single spike outliers in spectra may affect the value of the estimates. We then define the robust smoothing procedure and show how the influence of single spike outliers may be bounded.

The sampling of contaminated spectral data is formulated in the following model:

$$Y_i = f(t_i) + z_i \quad i = 1, \dots, n \quad (2.1)$$

where Y_1, \dots, Y_n are the observations at the points t_i , and z_i represents the noise. In Raman spectroscopy, the spacing $\Delta = t_i - t_{i-1}$ between two successive points on the wavenumber scale is usually constant. The function $f(t)$, denoting the true intensities, is to be estimated.

A procedure often applied to estimate $f(t)$, the true spectrum, on the basis of the observations Y_1, \dots, Y_n is the moving average (or linear filter)

$$f_n^*(t) = \sum_{i=1}^n c_i(t) Y_i \quad (2.2)$$

where $\sum_{i=1}^n c_i(t) = 1$ for all t , and $c_i(t)$ are weighting constants corresponding to a window of certain extent (see Refs. 2-8). Tuning the bandwidth of the window in accordance with the signal-to-noise ratio gives a smooth estimate of the intensities. Since this estimate is based on an average of the observations Y_i near t , only one huge single spike outlier may distort the linear filter (Eq. 2.2). Thus the estimate 2.2 depends greatly on the amount of outliers in the noise z_i . Whenever the noise

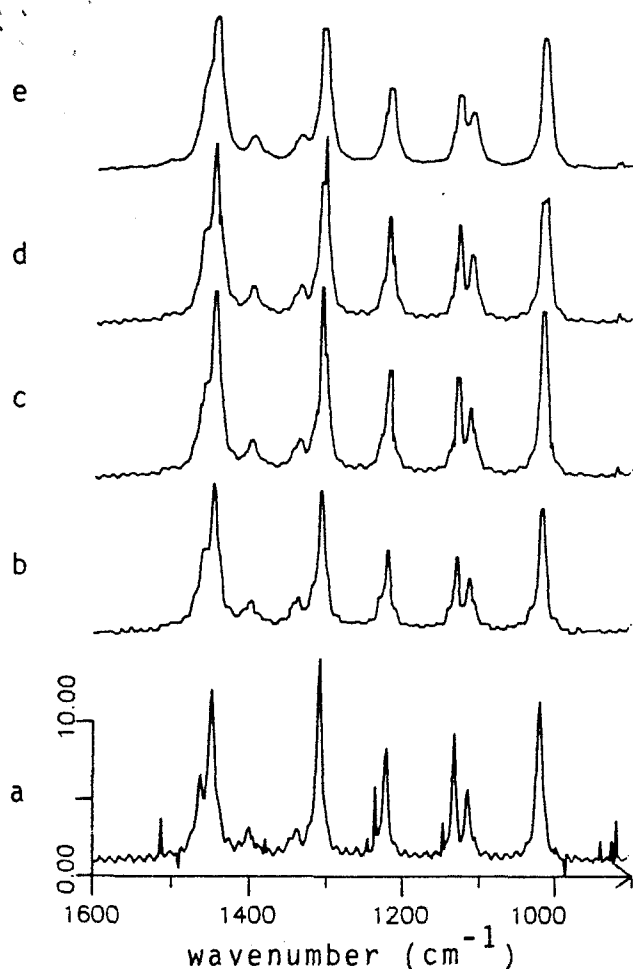


FIG. 4. Inverse Raman spectra of liquid *p*-dioxane: a, original data after Hillig and Morris;⁹ b, Savitzky-Golay³ fit, 5 points; c, robust smoothing, $\epsilon_n = 5$, $\chi = 1.0$; d, robust smoothing, $\epsilon_n = 5$, $\chi = 2.0$; e, robust smoothing, $\epsilon_n = 7$, $\chi = 1.0$.

contains single spike outliers, the moving average (2.2) will be misleading, in other words "not robust against outlying single spikes." We can see the influence of single spike outliers if we rewrite 2.2. The estimate $f_n^*(t)$ can be considered as the solution of

$$\sum_{i=1}^n c_i(t) \ell(Y_i - f_n^*(t)) = \min!$$

where $\ell(u) = u^2/2$. The distance of $f_n^*(t)$ to the observations Y_i is measured quadratically: Huge single spike outliers tow $f_n^*(t)$ away from the true spectral value $f(t)$. Construction of ψ , the derivative of ℓ , results in the so-called influence curve (IC) which is shown for the estimate 2.2 in Fig. 1.

To obtain a robust estimate, we bend down the tails of the IC, bounding the influence of single spike outliers. The robust smoothing method is thus defined through a ψ -function which is bounded (and also antisymmetric, as in Fig. 2). The robust estimate is $f_n(t)$, a solution of

$$\sum_{i=1}^n c_i(t) \psi(Y_i - f_n(t)) = 0 \quad (2.3)$$

with the same window $c_i(t)$ as in Eq. 2.2.

In Fig. 2 we give an example of a bounded IC (a ψ -function, which goes back to Huber¹⁰). This IC is also

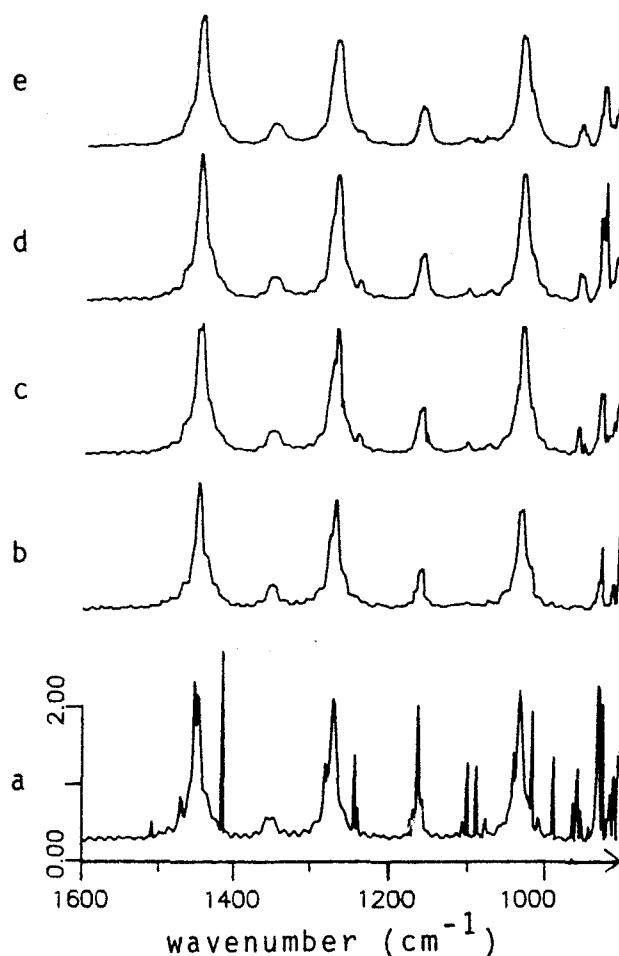


FIG. 5. Inverse Raman spectra of liquid cyclohexane: a, original data after Hillig and Morris;⁹ b, Savitzky-Golay³ fit, 5 points; c, robust smoothing, $\epsilon_n = 5$, $\chi = 0.3$; d, robust smoothing, $\epsilon_n = 5$, $\chi = 1.0$; e, robust smoothing, $\epsilon_n = 7$, $\chi = 0.3$.

implemented in the algorithm stated below. Note that any other bounded antisymmetric function, such as an arctan-curve, may be used. The IC of Fig. 2 is

$$\psi(u) = \ell'(u) = \min(\chi, \max(u, -\chi)) \quad (2.4)$$

where χ is the robustness parameter which corresponds to the amount of single spike outliers in the noise. The ℓ -function corresponding to ψ in 2.4 is a function which is, in the middle, like a parabola, and, in the tails, a straight line. Thus, outlying single spikes have less influence on the estimate $f_n(t)$.

By tuning χ in the ψ -function in 2.4 one can vary the degree of robustness of the estimate $f_n(t)$ against single spike outliers. By increasing χ , the solution of 2.3 approaches the "least-square" estimate $f_n^*(t)$, i.e. the tails of the IC are lifted. If χ approaches zero we obtain the moving median. A bulk of huge single spike outliers may be removed with a small value of χ , whereas spectra contaminated only by white noise of small variance should be smoothed with a larger value of χ .

As a digression, we may note that the same robustness considerations described above hold for parametric models, where a parameter is estimated by the least-square method. Such a parametric model is, for instance, a Gauss-Lorentzian curve mixture which is frequently

TABLE I. Fortran Code.

```

SUBROUTINE RAMSMO ( DATA, WINDOW, EXT, XP, ES-
  TIM, EPS, KAPPA, Z)
DIMENSION WINDOW(1), DATA(1), Z(EXT)
REAL ESTIM, EPS
INTEGER XP, EXT
C ----- Smoothing of Raman spectra -----
C DATA(1) spectral data input
C WINDOW(1) window, generated by GENWIN input
C EXT extension of window input
C XP point where to smooth input
C ESTIM estimate of intensity at XP output
C EPS precision of zero in Eq. 2.3 input
C KAPPA cutoff point of psi-fct. input
C Z(EXT) buffer input
C -----
ILOW=XP-EXT/2
DO 1 I=1,EXT
1 Z(I)=DATA(ILOW+I)
C --- start with initial estimate
T=MEDIAN(Z)
C --- Newton-Raphson loop with Huber's psi-fct., see Eq. 2.4.
100 SUMMA=0.
SUMMA2=0.
DO 2 I=1,EXT
YPSI=Z(I)-T
W=WINDOW(I)
YP=KAPPA
YPS=0.
IF( YPSI .GT. KAPPA) GOTO 3
YP= - KAPPA
IF( YPSI .LT. -KAPPA) GOTO 3
YP=YPSI
YPS=1.
3 SUMMA=SUMMA + W*YP
SUMMA2=SUMMA2 + W*YPS
2 CONTINUE
H=SUMMA / SUMMA2
IF( ABS( H ) .LE. EPS ) RETURN
T=T + H
GOTO 100
END
C
C
C SUBROUTINE GENWIN( WINDOW, EXT, SPACE )
DIMENSION WINDOW(EXT)
REAL SPACE
INTEGER EXT
C ----- Generation of Smoothing Window -----
C WINDOW(1) window, generated here output
C EXT extension of window (odd) input
C SPACE spacing distance on wave- input
C number axis
C -----
IMID=EXT/2 + 1
DO 1 I=1, (IMID-1)
X=SPACE * I / IMID
C --- use quadratic kernel for instance
W= .75 * (1. - X * X)
WINDOW(IMID+I)= W
1 WINDOW(IMID-I)= W
WINDOW(IMID)=.75
RETURN
END
    
```

in use. For this model, a robust estimate of the parameter may be introduced in the same way as above.

The numerical algorithm for the robust smoothing procedure is given by the FORTRAN code which appears as Table I. The procedure uses the Newton-Raphson algorithm to solve the implicit equation 2.3 (loop 100 in the code below).

First, the window $c_i(t)$ has to be generated; this is accomplished in SUBROUTINE GENWIN. We have chosen a window of parabolic shape (see Fig. 3), but any other window may be used in GENWIN.

RESULTS

We now apply the robust smoothing procedure to some experimental data containing different noise amplitudes. In particular, we consider the inverse Raman spectra published recently by Hillig and Morris.⁹

Hillig and Morris emphasize that one has need of a spike-detecting routine in the case of sensitive absorption measurements. In Figs. 4 and 5 we reproduce the inverse Raman spectra of *p*-dioxane and cyclohexane/carbon black, respectively. By comparison of the different spectra given in Figs. 4 and 5, we can demonstrate the limitations of the Savitzky-Golay filter³ and the method proposed by Hillig and Morris,⁹ and some advantages of the robust smoothing technique.

The first example shows *p*-dioxane (Fig. 4). The original data are affected by several single spike outliers (at $\sim 1150 \text{ cm}^{-1}$ for instance), and contain white noise with a scale of about 5% of the intensity of the strongest band. By applying the Savitzky-Golay (five points) fit, one sees that the single outlying spikes are reduced but retained as shoulders. A second more serious effect, generated by the local parabolic fit, is that the intensity of single outliers adds up to the peaks in the neighborhood. This fact can be drawn from Fig. 4b where the intensity ratio between the highest bands changes relative to the original data. Figure 4c-e shows the same spectra fitted by the robust smoothing algorithm. While in Fig. 4c and d, the intensities are well reproduced, in Fig. 4e the intensity ratio changes. This is due to a larger value of ϵ_n (see Fig. 3) and to the fact that the spectral band near 1450 cm^{-1} has a higher halfwidth than the spectral band at 1320 cm^{-1} . The effect of increasing χ can be seen by comparison of Fig. 4c and 4d. Changing χ from 1.0 to 2.0 allows for more influence of single spike outliers. For instance, the shoulder at the spectral band 1210 cm^{-1} , introduced by the outlier at 1230 cm^{-1} , is more visible in Fig. 4d. With their method, Hillig and Morris⁹ observe a large broadening of the bands while all the spikes are removed. This is not so with our method, as can be seen in Fig. 4c-e.

For the second example (cyclohexane/carbon black, see Fig. 5) Hillig and Morris⁹ emphasize that the Savitzky-Golay smoother does a poor job of removing the spikes. But even the Hillig and Morris⁹ procedure does not remove all outlying spikes, as can be drawn from their fig. 2c. The application of our algorithm to their spectral data shows the advantage of the robust smoothing technique. In all cases shown in Fig. 5c-e, the single outliers are removed and the ratio of the peak intensities is preserved. The cutoff-parameter χ , as defined above, was set to 0.3 and 1.0, respectively, yielding a highly robust estimate of the spectral intensities. In that example many single outliers occurred; therefore, the robustness parameter χ should be low, as discussed above. Due to the small amount of data points for one band, all smoothed spectra show diminished peak intensities. In the case of more observations this effect will not be so pronounced.

The fact that not all spikes are removed by the Hillig and Morris⁹ procedure shows that their procedure fails in the case in which a single outlier is close to a spectral band. In contrast, Fig. 5e suggests that our robust smoothing algorithm removed all single outliers in the neighborhood of the dominant spectral peaks. Last, but not least, we may note that the robust smoothing algorithm usually stopped after one Newton-Raphson iteration (see loop 100 in the FORTRAN code). The computation time is thus very low; for instance, the smoothing presented in Fig. 5e consumed 0.8 s, which is about ten times faster than the Hillig and Morris⁹ procedures.

SUMMARY

A new robust smoothing method for spectral data is proposed. A comparison with two other methods—the Savitzky-Golay and the Hillig-Morris procedures—shows the advantage of the new algorithm. While in spectral data containing single outliers near a spectral band the two other smoothers fail (by introduction of shoulders),

the robust smoothing algorithm as proposed here does a good job when the relevant parameters are tuned correctly.

ACKNOWLEDGMENTS

Special thanks go to Prof. Dr. H. H. Eysel for helpful discussion. We also thank M. Morris for the permission to use his data.

This work was in part financially supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 123.

Computations were performed on IBM 370-168 at the Universitätsrechenzentrum, Heidelberg.

1. J. Laane and W. Kiefer, *Appl. Spectrosc.* **35**, 267 (1981).
2. P. Gans and J. B. Gill, *Appl. Spectrosc.* **31**, 451 (1977).
3. A. Savitzky and M. J. E. Golay, *Anal. Chemistry* **36**, 628 (1964).
4. W. F. Edgell, E. Schmidlin, and M. W. Balk, *Appl. Spectrosc.* **34**, 420 (1980).
5. W. F. Edgell, E. Schmidlin, T. J. Kuriakose, and P. Lurix, *Appl. Spectrosc.* **30**, 168 (1976).
6. W. F. Maddams, *Appl. Spectrosc.* **34**, 245 (1980).
7. D. A. Stephenson and R. J. Blint, *Appl. Spectrosc.* **33**, 41 (1979).
8. R. W. Chrisman, J. C. English, and R. S. Tobias, *Appl. Spectrosc.* **30**, 168 (1976).
9. K. Hillig and M. Morris, *Appl. Spectrosc.*, **36**, 700 (1982).
10. P. Huber, *Robust Statistics* (Wiley and Sons, New York, 1981).

Biometrika (1985), 72, 2, pp. 481-4
 Printed in Great Britain

Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression

BY W. HÄRDLE

Fachbereich Mathematik, Johann Wolfgang Goethe Universität, D-6000 Frankfurt am Main, Federal Republic of Germany

AND J. S. MARRON

Department of Statistics, University of North Carolina, Chapel Hill, N.C. 27514, U.S.A.

SUMMARY

The bandwidth selection problem in nonparametric kernel regression is considered. Bandwidth selectors based on cross-validation and on Akaike's information criterion, AIC, and his finite prediction error, FPE, are among those compared. It is seen that they are not necessarily asymptotically equivalent. Conditions are given under which the equivalence holds and modifications are suggested which make the selectors equivalent.

Some key words: Bandwidth selection; Kernel estimator; Model selection; Nonparametric regression.

1. INTRODUCTION

Let $(X, Y), (X_1, Y_1), \dots, (X_n, Y_n)$ be independent, identically distributed random vectors and let

$$m(x) = E(Y | X = x)$$

denote the regression curve of Y on X . Consider the estimator

$$\hat{m}_h(x) = n^{-1} h^{-1} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) Y_i / f(x),$$

(Johnston, 1982), where K is a 'kernel' or 'window' function, $h = h_n$ is a bandwidth, and the marginal density $f(x)$, of X , is assumed to be known. In the present paper, several bandwidth selectors, most of which are derived from model selection procedures, are compared. It will be shown that, contrary to what may be expected in view of the results of Stone (1977), Shibata (1981) and Rice (1984), in quite simple cases, these selection procedures are not all asymptotically equivalent to each other.

It may appear as a drawback that the estimator is defined in terms of a known marginal density, $f(x)$. This is only done for clarity of presentation. In the slightly more complicated case of the Nadaraya-Watson estimator, $m_h^* = \hat{m}_h f / \hat{f}_h$, with \hat{f}_h a kernel density estimator, the approximations used in an unpublished paper by the present authors can be employed to see that the ideas of this paper also apply to m_h^* . These results also hold in the case of a multivariate design vector, X .

The basic idea of most bandwidth selection rules is to choose the bandwidth h to make $\hat{m}_h(X_i)$ an effective predictor of Y_i . A crude attempt at this would be to minimize the resubstitution estimate of the prediction error,

$$p(\hat{m}_h) = n^{-1} \sum_{i=1}^n \{Y_i - \hat{m}_h(X_i)\}^2 w(X_i),$$

where w is a nonnegative weight function. This estimate has an optimistic bias because Y_i is used in the prediction of Y_i . Thus the bandwidth selector which minimizes $p(\hat{m}_h)$ has a tendency to undersmooth or, in other words, take h too small.

The above optimistic bias can be avoided by the method of cross-validation. For bandwidth selection, this leads (Wahba & Wold, 1975; Clark, 1975) to minimizing

$$p'(\hat{m}_h) = n^{-1} \sum_{i=1}^n \{Y_i - \hat{m}_{h,i-}(X_i)\}^2 w(X_i),$$

where

$$\hat{m}_{h,i-}(x) = (n-1)^{-1} h^{-1} \sum_{j \neq i} K\left(\frac{x-X_j}{h}\right) Y_j / f(x).$$

Consider the mean integrated squared error distance given by

$$d_M(\hat{m}_h, m) = E \int \{\hat{m}_h(x) - m(x)\}^2 w(x) f(x) dx.$$

To see that the task of minimizing $p'(\hat{m}_h)$ is asymptotically equivalent to the task of minimizing $d_M(\hat{m}_h, m)$, write

$$p'(\hat{m}_h) = p(m) + d'_A(\hat{m}_h, m) + 2c(h), \tag{1}$$

where

$$p(m) = n^{-1} \sum_{i=1}^n \{Y_i - m(X_i)\}^2 w(X_i),$$

$$d'_A(\hat{m}_h, m) = n^{-1} \sum_{i=1}^n \{\hat{m}_{h,i-}(X_i) - m(X_i)\}^2 w(X_i),$$

$$c(h) = n^{-1} \sum_{i=1}^n \{Y_i - m(X_i)\} \{m(X_i) - \hat{m}_{h,i-}(X_i)\} w(X_i).$$

The unpublished paper by the present authors shows that, under suitable assumptions,

$$\sup_h |d'_A(\hat{m}_h, m) - E\{d'_A(\hat{m}_h, m)\}| / d_M(\hat{m}_h, m) \rightarrow 0,$$

almost surely, and

$$\sup_h |c(h) - E\{c(h)\}| / d_M(\hat{m}_h, m) \rightarrow 0$$

almost surely, where \sup_h denotes supremum over h in an interval $(h_*, h^*) \subseteq \mathbb{R}^+$. Thus, since

$$\sup_h |E\{d'_A(\hat{m}_h, m)\} - d_M(\hat{m}_h, m)| / d_M(\hat{m}_h, m) \rightarrow 0, \quad E\{c(h)\} = 0,$$

(1) may be written as

$$p'(\hat{m}_h) = p(m) + d_M(\hat{m}_h, m) + o\{d_M(\hat{m}_h, m)\}, \tag{2}$$

where o is uniform over h in the above sense.

Thus, the task of minimizing $p'(\hat{m}_h)$ is asymptotically equivalent to the task of minimizing $d_M(\hat{m}_h, m)$. For this reason, a bandwidth selector will be said to be 'asymptotically equivalent to $d_M(\hat{m}_h, m)$ ' whenever it has an asymptotic representation as a sum of $d_M(\hat{m}_h, m)$, a term independent of h , and negligible terms as in (2).

2. NONEQUIVALENCE OF SOME BANDWIDTH SELECTORS

To verify the poor performance of the bandwidth selector which minimizes $p(\hat{m}_h)$, as in (2) write

$$p(\hat{m}_h) = p(m) + d_M(\hat{m}_h, m) - 2n^{-1} h^{-1} K(0) \int V(x) w(x) dx + o\{d_M(\hat{m}_h, m)\}, \tag{3}$$

where $V(x)$ denotes the conditional variance $V(x) = E(Y^2 | X = x) - \{m(x)\}^2$. The third term on the right-hand side of (3) is of the same order as the variance of \hat{m}_h . Thus $p(\hat{m}_h)$ is not asymptotically equivalent to $d_M(\hat{m}_h, m)$.

Several other bandwidth selection rules have been proposed. Most of these selectors were originally developed in the context of model selection (Rice, 1984). As above, these involve minimization of a function of h . Each of these may be thought of as multiplying $p(\hat{m}_h)$ by a selection penalty, $\Xi(t)$, which is a function of

$$t(h) = n^{-1} h^{-1} K(0) n^{-1} \sum_{i=1}^n \{f(X_i)\}^{-1}.$$

Examples are as follows:

- (i) generalized cross-validation (Craven & Wahba, 1979), $GCV(h) = p(\hat{m}_h) \{1 - t(h)\}^{-2}$;
- (ii) Akaike's information criterion, AIC (Akaike, 1974), $\exp AIC(h) = p(\hat{m}_h) \exp\{2t(h)\}$;
- (iii) finite prediction error, FPE (Akaike, 1970), $FPE(h) = p(\hat{m}_h) \{1 + t(h)\} / \{1 - t(h)\}$;
- (iv) a selector of Shibata (1981), $S(h) = p(\hat{m}_h) \{1 + 2t(h)\}$;
- (v) a selector of Rice (1984), $T(h) = p(\hat{m}_h) / \{1 - 2t(h)\}$.

Observe that, by Taylor's theorem, each of the above selectors may be written in the form $p(\hat{m}_h) [1 + 2t(h) + o\{t(h)\}]$, which motivates the definition of a general selection penalty function $\Xi(t)$ which includes all of the above.

For Ξ , with $\Xi(0) = 1$, $\Xi'(0) = 2$, Ξ'' bounded on a neighbourhood of the origin,

$$G(h) = p(\hat{m}_h) \Xi\{t(h)\}.$$

The bandwidth selector G , and hence (i)–(v) above as well, may now be analysed by noting that, as in (2) and (3),

$$G(h) = p(m) + d_M(\hat{m}_h, m) - 2n^{-1} h^{-1} K(0) \int V(x) w(x) dx + 2E\{t(h)\} E\{p(m)\} + o\{d_M(\hat{m}_h, m)\}.$$

Straightforward computations yield $E\{p(m)\} = \int V(x) f(x) w(x) dx$, and, if the additional assumption is made that f is supported and bounded above 0 on the support of w , say the interval (a, b) , then $E\{t(h)\} = n^{-1} h^{-1} K(0) (b-a)$. Hence, G , and (i)–(v), are not asymptotically equivalent to $d_M(\hat{m}_h, m)$, unless

$$\int V(x) w(x) dx = (b-a) \int V(x) f(x) w(x) dx.$$

In general, this seems quite unlikely, but note that it does happen in two commonly considered special cases:

- (a) $f(x)$ is constant on (a, b) ,
- (b) $V(x)w(x)$ is constant on (a, b) .

In the setting of fixed design regression, Rice (1984) has established the asymptotic equivalence of (i)–(v) to $d_M(\hat{m}_h, m)$ under assumption (a). In the setting of spline regression, Craven & Wahba (1979) have demonstrated a weaker, expected value version of the asymptotic equivalence of (i) to $d_M(\hat{m}_h, m)$ under the assumption that both V and w are constant.

From another point of view, (b) can be interpreted as saying that if one wants to use the selector $G(h)$, or any of the others (i)–(v), one should choose $w(x) = V(x)^{-1}$. This has already been suggested by Silverman (1985, §5). An obvious drawback to this is that typically, in practice, the function V is unknown. It is an advantage of $p'(\hat{m}_h)$ to be asymptotically equivalent to $d_M(\hat{m}_h, m)$ independently of the choice of w .

There are three readily apparent ways to overcome the above difficulties. First, a reasonable estimate of $\int V(x)w(x)dx$ is provided by

$$p^*(\hat{m}_h) = n^{-1} \sum_{i=1}^n \{Y_i - \hat{m}_{h,i-}(X_i)\}^2 w(X_i) \{f(X_i)\}^{-1}.$$

Thus the selector $S(h)$, for example, could be modified to $S(h) = p(\hat{m}_h) + 2t(h)p^*(\hat{m}_h)$, and the other selectors can be similarly modified. The second way is to find an estimate of $V(x)$, possibly a smoothing of the squared residuals and substitute its inverse for $w(x)$ in $p(\hat{m}_h)$. Thirdly, using the idea of prewhitening as suggested by D. Brillinger, transform the data so that X_1, \dots, X_n can be thought of as uniform variables, by plugging them into the inverse of the cumulative distribution function.

ACKNOWLEDGEMENTS

This research was conducted during the visit of the second author to the Universität Heidelberg. The support of the Sonderforschungsbereich 123 of the Deutsche Forschungsgemeinschaft is gratefully acknowledged.

REFERENCES

- AKAIKE, H. (1970). Statistical predictor information. *Ann. Inst. Statist. Math.* **22**, 203-17.
 AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Auto Control* **AC 19**, 716-23.
 CLARK, R. M. (1975). A calibration curve for radio carbon dates. *Antiquity* **49**, 251-66.
 CRAVEN, P. & WAHBA, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31**, 377-403.
 JOHNSTON, G. J. (1982). Probabilities of maximal deviations for nonparametric regression function estimates. *J. Mult. Anal.* **12**, 402-14.
 RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215-30.
 SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45-54.
 SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. R. Statist. Soc. B* **47**. To appear.
 STONE, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Statist. Soc. B* **39**, 44-7.
 WAHBA, G. & WOLD, S. (1975). A completely automatic french curve: fitting spline functions by cross-validation. *Comm. Statist.* **4**, 1-17.

[Received June 1984. Revised January 1985]

J. R. Statist. Soc. B, (1984),
 46, No. 1, pp. 42-51

Robust Non-parametric Function Fitting

By W. HÄRDLE and T. GASSER

University of Heidelberg, Germany

[Received December 1982]

SUMMARY

A robust non-parametric function fitting method is introduced. The estimate is motivated from the theory of M -estimation and of kernel estimation of regression functions. Consistency and asymptotic normality are shown. Bias and variance rates are the same as those previously obtained by Gasser and Müller (1979) for linear smoothers. The estimate satisfies a minimax property, i.e. it minimizes the maximal asymptotic variance as the error distributions vary over a suitable contamination neighbourhood.

Keywords: NONPARAMETRIC REGRESSION; ROBUST SMOOTHING; FUNCTION FITTING

1. INTRODUCTION

Let $F = \{f(y, x) : x \in [0, 1]\}$ be a family of probability density functions (p.d.f.) indexed by "x" and let $m(x) = \int yf(y, x) dy$ be the regression function of y on x . Suppose that we have made n observations

$$Y_i = m(x_i) + \epsilon_i, \quad 0 \leq x_1 \leq x_2 \leq \dots \leq x_n = 1,$$

where ϵ_i has p.d.f. $f(y - m(x_i); x_i)$. The intention is to estimate $m(x)$ on the basis of the observations $\{(x_1, Y_1), \dots, (x_n, Y_n)\}$.

Several estimators of the regression function $m(x)$ have been introduced. Priestley and Chao (1972) proposed the following estimator

$$m_n^*(x) = \sum_{i=1}^n \alpha_i(x) Y_i, \tag{1}$$

where $\alpha_i(x) = \alpha_i^{PC}(x) = h_n^{-1} K((x - x_i)/h_n) (x_i - x_{i-1})$, $x_0 = 0$, $K(\cdot)$ is a kernel function and $\{h_n\}$ is a sequence of positive bandwidths tending to zero as the sample size n tends to infinity. This estimator is essentially a generalization to regression of kernel methods introduced for density estimation by Rosenblatt (1956) and Parzen (1962). Benedetti (1977) showed the asymptotic normality of the estimator (1) with $\alpha_i = \alpha_i^{PC}$. Gasser and Müller (1979) used the weights

$$\alpha_i(x) = h_n^{-1} \int_{s_{i-1}}^{s_i} K((x - u)/h_n) du \tag{2}$$

with a sequence of interpolating points $\{s_i\}_{i=0}^n$, such that $x_{i-1} \leq s_i \leq x_i$, $x_0 = 0$, $i = 1, \dots, n$ and $s_0 = 0$, $s_n = 1$. Cheng and Lin (1981) showed uniform consistency of $m_n^*(x)$ under mild conditions, setting $s_i = x_i$ in formula (2).

Present address: Sonderforschungsbereich 123, Universität Heidelberg, Im Neuenheimer Feld 293, D-6900 Heidelberg, West Germany.

Whenever the residuals ϵ_i are normally distributed a linear smoother, defined in (1), is appropriate to obtain a non-parametric estimate for the function $m(\cdot)$ but for longer-tailed residual p.d.f.'s gross misinterpretations are to be suspected.

Not only will the variance be inflated, moreover the outlying residuals might initiate (smooth) peaks and troughs due to the averaging property of (1).

If one is interested in structural elements such as extrema of the regression curve, it seems to be more advantageous to apply an outlier insensitive method; i.e. an estimator which is robust with respect to spurious huge spikes in the residuals.

We therefore propose to estimate the function $m(x)$ by $m_n(x)$ a zero of $H_n(x, \cdot)$, where

$$H_n(x, \cdot) = \sum_{i=1}^n \alpha_i(x) \psi(Y_i - \cdot) \tag{3}$$

and $\psi(u)$ is a bounded odd function. Observe that $\sum_{i=1}^n \alpha_i(x) = 1$ for the weights defined in (2). Thus taking $\psi(u) = u$ in equation (3) gives the linear local average $m_n^*(x)$. The definition of $m_n(x)$ is a straightforward application of M -estimation of location to regression function estimation. The only difference to (weighted) M -estimation (Huber, 1981) is that the observations $\{(x_i, Y_i)\}_{i=1}^n$ are not identically distributed and thus introduce bias problems if $m(x)$ is not a constant.

However, the weights $\{\alpha_i(x)\}$ concentrate asymptotically like a delta spike making the observations "locally i.i.d." and we thus obtain asymptotic unbiasedness.

Cleveland (1979) introduced another robust estimator of $m(x)$ involving robust local polynomial fits in several steps. A variety of papers are concerned with linear estimators of type (1). Rosenblatt (1969) computed (asymptotic) bias and variance rates for the case that X is also allowed to be a random variable.

Gasser and Müller (1979) considered a wide class of kernel-functions. Collomb (1981) gives a bibliographic review on non-parametric regression function estimation.

We show consistency and asymptotic normality of $m_n(x)$ under mild conditions on the ψ -function and the sequence of band-widths $\{h_n\}$. We further compute rates of bias and variance and show an asymptotic minimax property of $m_n(x)$. More precisely, the asymptotic variance of $m_n(x)$ has a saddlepoint when $f(y; x)$, the underlying p.d.f., varies over a contamination neighbourhood of some fixed $g(y; x)$. A small Monte Carlo study was carried out to demonstrate the different behaviour of $m_n(x)$ and $m_n^*(x)$ in the diagnosis of extrema. Finally, the two estimators are applied to real data: the fitting of Raman spectra, which is a standard smoothing procedure in physical chemistry to determine location and height of spectral bands.

It will be convenient to introduce the following assumptions.

- (A1) $\psi(u)$ is a monotone, odd and bounded function having a bounded derivative $\psi'(u)$ (except at a finite number of points), $\psi'(0)$ exists and is positive.
- (A2) $f(y; x)$ is symmetric for all $x \in [0, 1]$ and $\partial^2 f / \partial x^2 (y; x)$ exists for all $x \in [0, 1]$.
- (A3) The bandwidths $\{h_n\}$ satisfy

$$(I) \quad h_n \rightarrow 0. \quad (II) \quad nh_n \rightarrow \infty.$$

- (A4) The interpolating sequence $\{s_i\}_{i=0}^n$ satisfies

$$\sup_i |s_i - s_{i-1} - n^{-1}| = O(n^{-\delta}), \quad \delta > 1.$$

- (A5) The kernel function $K(u)$ is Lipschitz-continuous and has compact support $[-A, A]$. It further satisfies

$$(I) \quad \int K(u) du = 1. \quad (II) \quad \int uK(u) du = 0.$$

Assumption (A1) guarantees the robustness of $m_n(x)$. We excluded the "local median" which

has the ψ -function $\kappa \cdot \text{sign}(u)$, $\kappa > 0$, since the existence of the derivative $\psi'(0)$ immediately implies

$$c_0 = \inf_x \int \psi'(y - m(x)) f(y; x) dy > 0$$

which is needed in the proofs. However, it will be clear that with a proper interpretation of c_0 in the case the median is used the proofs are also valid for the median.

Assumption (A4) may also be found in Gasser and Müller (1979), and guarantees an asymptotic uniform spacing of the sequence $\{s_i\}$.

A variety of ψ -functions may be chosen in defining $m_n(x)$ through equation (3). For instance, Huber's ψ -function (1964):

$$\psi(u) = \max\{-\kappa, \min\{u, \kappa\}\}, \kappa > 0 \tag{4}$$

or a suitably scaled arctan-curve or normal distribution function. Assumption (A1) excludes for the moment those ψ -functions which redescend to zero as $|u| \rightarrow \infty$. We therefore skip the discussion of this point here and refer to the next section.

In all statements that follow, x will denote an interior point $x \in (0, 1)$ of the interval $[0, 1]$. Integration with respect to the p.d.f. $f(y; x)$ will be indicated with an index "x"; $E_x y$ is thus $\int y f(y; x) dy = m(x)$.

2. CONSISTENCY AND ASYMPTOTIC NORMALITY

We first characterize the solutions of (3). It is evident that if ψ is not strictly monotone several solutions of (3) may exist.

Lemma 2.1. For each x , for each n , the set $L_n(x) = \{m: m \text{ solves (3)}\}$ is non-empty, compact provided at least one observation (x_i, Y_i) is contained in the support of the kernel function. If, in addition the kernel is positive then $L_n(x)$ is moreover convex.

The proof is the same as in Huber's fundamental paper (1964, Lemma 1); we only have to cope with the kernel weights. We therefore define $m_n(x)$, the robust estimate of $m(x)$, as any representative of $L_n(x)$.

The weak consistency follows easily from the (local) monotony of ψ (assumption (A1)).

Proposition 2.1. Suppose that (A1) to (A4) hold, then $m_n(x)$ is weakly consistent, i.e. $m_n(x) \xrightarrow{P} m(x)$.

The proof follows immediately from the weak law of large numbers. Applying it to $H_n(x, m) = \sum_{i=1}^n \alpha_i(x) Z_i$, with the bounded random variables (r.v.) $Z_i = \psi(y_i - m)$, it follows that $H_n(x, m) - EH_n(x, m) \xrightarrow{P} 0$ for all x, m .

Since $EH_n(x, m(x)) \rightarrow 0$ as $n \rightarrow \infty$ by the following lemma, the assertion follows from (A1) the antisymmetry of ψ and (A2) the symmetry of $f(y; x)$.

Lemma 2.2. Let $h(x): [0, 1] \rightarrow \mathbb{R}$ be a Lipschitz continuous function. Suppose that (A3), (A4) and (A5) hold, then we obtain

$$\lim_{n \rightarrow \infty} \sup_{a \leq x \leq b} \left| \sum_1^n \alpha_i(x) h(x_i) - h(x) \right| = 0, \quad 0 < a < b < 1,$$

where the sequence of weights $\{\alpha_i(x)\}$ is defined by (2).

The proof follows from the mean value theorem and the following inequality

$$\left| h_n^{-1} \sum_{i=1}^n \int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{h_n}\right) h(x_i) du - h_n^{-1} \int_0^1 K\left(\frac{x-u}{h_n}\right) h(x) du \right|$$

$$\leq h_n^{-1} \sup_u |K(u)| \sum_{J_n} |s_j - s_{j-1}| \cdot |h(x_j) - h(\xi_j)|,$$

where ξ_j are suitable mean values and J_n is the set of indices $J_n = \{j: |x - x_j| < h_n A\}$ which satisfies $|T_n| = O(nh_n)$ by assumption (A4). From the Lipschitz-continuity of $h(\cdot)$ the assertion follows. For ψ -functions $\psi(u)$ bending down to zero as $|u| \rightarrow \infty$ (which obviously have strong robustness properties) neither the proof of Lemma 2.1 nor that of Proposition 2.1 will work since the technique there heavily depends on the monotony of $\psi(u)$. However, we may obtain consistency for rebending $\psi(u)$ as well if we couple the solutions of (3) in a suitable way to a consistent estimate of $m(x)$. That is define $\tilde{m}_n(x)$, the robust estimator for not necessarily monotone ψ , as that solution of (3) which is nearest to some other consistent estimate $c_n(x)$ of the regression function $m(x)$.

$$|\tilde{m}_n(x) - c_n(x)| = \inf \{ |m - c_n(x)| : m \text{ solves (3)} \}.$$

By standard arguments it can be shown that $\tilde{m}_n(x)$ is also consistent (Andrews *et al.*, 1972). By taking different $c_n(x)$ as coupled estimates one obtains a different behaviour of $\tilde{m}_n(x)$. By Proposition 2.1 we may put $c_n(x) = m_n(x)$ or we may use $m_n^*(x)$ which is also consistent (Gasser and Müller, 1979).

To formulate the theorem on the asymptotic normality we need some more notation. Let

$$B_n(x) = EH_n(x), \quad \gamma(x) = E_x \psi'(y - m(x)),$$

$$\sigma^2(x) = E_x \psi^2(y - m(x)) = \text{var}_x \psi(y - m(x))$$

$$\beta_n = \left[(nh_n)^{-1} \sigma^2(x) \int K_2 \int K^2(u) du (u) du \right]^{-\frac{1}{2}}, \text{ where } H_n(x) = H_n(x, m(x)).$$

Theorem 2.1. Suppose that Assumptions (A1)-(A5) hold, then

$$Z_n(x) = \beta_n \gamma(x) [(m_n(x) - m(x) - B_n(x))/\gamma(x)]$$

is asymptotically standard normally distributed.

Proof. By Taylor expansion it follows that

$$m_n(x) - m(x) = H_n(x)/D_n(x),$$

where

$$D_n(x) = \sum_{i=1}^n \alpha_i(x) \psi'(Y_i - m(x) + W_i)$$

and $|W_i| \leq |m_n(x) - m(x)|$. From the weak law of large numbers, Proposition 2.1 and the boundedness of ψ' it follows that $D_n(x) \xrightarrow{P} \gamma(x)$. It remains to show that $\beta_n(H_n(x) - B_n(x))$ is asymptotically normally distributed.

Recalling the definition of $H_n(x)$ and $B_n(x)$, it follows that

$$H_n(x) - B_n(x) = \sum_{i=1}^n \alpha_i(x) Z_i, \quad Z_i = \psi(Y_i - m(x)) - E_{x_i} \psi(Y - m(x)),$$

with bounded and independent r.v. $\{Z_i\}_{i=1}^n$.

The asymptotic normality now follows from Assumption (A4) ($nh_n \rightarrow \infty$) and the central limit theorem applying Ljapunov's condition.

The term $B_n(x)$ may be interpreted as a bias term and is due to the centering around $EH_n(x)$. However it is evident that $B_n(x) \rightarrow 0$ as $n \rightarrow \infty$. So if we are interested in (asymptotic) confidence bands for $m(x)$ without the bias term $B_n(x)$, we have to require that $B_n(x)$ is asymptotically negligible of order $o((nh_n)^{-\frac{1}{2}})$.

From Proposition 3.1 we have that the bias $B_n(x) = O(h_n^2)$ for kernel functions $K(u)$ satisfying (A5). Hence it suffices to require $nh_n^5 \rightarrow 0$ as $n \rightarrow \infty$ and the following corollary holds.

Corollary 2.1. Suppose that in addition to the assumptions of the theorem $nh_n^5 \rightarrow 0$ as $n \rightarrow \infty$. Then

$$W_n(x) = \beta_n[\gamma(x) (m_n(x) - m(x))]$$

is asymptotically standard normally distributed.

The condition $nh_n^5 \rightarrow 0$ may also be found in some work on the Nadaraya-Watson estimator, compare Schuster (1972) or Johnston (1979).

The above corollary enables us to compute confidence bands provided consistent estimates of $\sigma^2(x)$ and $\gamma(x)$ exist. These are, for instance,

$$\begin{aligned} \sigma_n^2(x) &= \sum_{i=1}^n \alpha_i(x) \psi^2(Y_i - m_n(x)), \\ \gamma_n(x) &= \sum_{i=1}^n \alpha_i(x) \psi'(Y_i - m_n(x)) \end{aligned} \tag{5}$$

which converge by the weak law of large numbers to $\sigma^2(x)$ and $\gamma(x)$ respectively. So an asymptotic pointwise confidence region for $m(x)$ may be obtained by the following:

Corollary 2.2. Suppose that the conditions of the theorem hold and that $nh_n^5 \rightarrow 0$. Let $\sigma_n^2(x)$ and $\gamma_n(x)$ defined as in (6), then

$$(nh_n)^{\frac{1}{2}} \left[\sigma_n^2(x) \int K^2(u) du \right]^{-\frac{1}{2}} \gamma_n(x) (m_n(x) - m(x))$$

is asymptotically normally distributed.

Note that the theorem on the asymptotic normality does not depend on the monotony assumption of (A1). It is also true for rebending ψ -functions (provided some assumptions concerning consistency are fulfilled). We only used the continuity of ψ and the boundedness of ψ in the theorem.

3. BIAS, VARIANCE

In this section we compute the rates of the bias term and of the variance of $m_n(x)$. Both together give us the rate for the mean squared error (M.S.E.) of $m_n(x)$. We compare this with the respective quantities of $m_n^*(x)$. We begin with the approximation of the bias term $B_n(x)$.

Lemma 3.1. Suppose that conditions (A1) to (A5) hold, then the bias can be approximated by

$$B_n(x) = \int K(u) E_{x-uh_n} \psi(Y - m(x)) du + O(n^{-1}).$$

Proof. Since $B_n(x) = EH_n(x) = \sum_{i=1}^n \alpha_i(x) E_{x_i} \psi(Y - m(x))$ we estimate

$$\left| \sum_{i=1}^n \alpha_i(x) E_{x_i} \psi(Y - m(x)) - h_n^{-1} \int_0^1 K\left(\frac{x-u}{h_n}\right) E_u \psi(y - m(x)) du \right|$$

$$\leq h_n^{-1} \sum_{i=1}^n \left| \int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{h_n}\right) [E_{x_i} \psi(y - m(x)) - E_u \psi(y - m(x))] du \right|$$

$$\leq M(\psi, f) \max_u |K(u)| h_n^{-1} \sum_{j=1}^p |s_j - s_{j-1}| |x_i - \xi_j| \quad \text{by (A1) and (A2)}$$

with suitable mean values $\{\xi_j\}$ and where $M(\psi, f)$ denotes the upper bound of $|\psi| \cdot |\partial f / \partial x|$.

By assumption (A4) the right-hand term in $O(n^{-1})$ and the Lemma follows. A hierarchy of kernels may now be defined.

Definition 3.1. A kernel K is of order p i.f.f. the following conditions hold:

- (I) $\int K(u) u^j du = 0, \quad j = 1, \dots, p - 1,$
- (II) $\int u^p K(u) du \neq 0.$

Note that Assumption (A5) define a kernel function K , that is of order 2. Kernel functions of higher order allow higher rates for the bias at the price that the assumptions on the regression function have to be more stringent.

Proposition 3.1. Let $K(\cdot)$ be a kernel function of order p and the regression function $m(x)$ be p -times differentiable. Then

$$B_n(x) / \gamma(x) = (-1)^p / p! h_n^p \int K(u) u^p du \cdot \frac{d^p m}{dx^p} + O(n^{-1}) + o(h_n^p).$$

Proof. Consider the function $T(\epsilon) = E_{x-\epsilon} \psi(y - m(x))$. Then from the Taylor expansion applied to $m(x)$, we have

$$T(\epsilon) = \int \Delta \cdot f(y; x - \epsilon) dy \cdot E_x \psi'(y - m(x)) + O(\Delta),$$

where $\Delta = m(x - \epsilon) - m(x) = m'(x) \epsilon + \epsilon^2 m''(x) + \dots$. If we plug in $T(x - uh_n)$ into the right-hand side of Lemma 3.1 the assertion follows.

The bias rate and the constants are thus the same as for $m_n^*(x)$ (Gasser and Müller, 1979). From the proof of Theorem 2.1 it is clear that

$$(nh_n) \text{ var } H_n(x) \rightarrow \int K^2(u) du \sigma^2(x) \text{ as } n \rightarrow \infty.$$

So an approximation to the asymptotic variance is given by $\text{var } H_n(x)$. Define

$$v(x) = E_s [\psi(y - m(x)) - E_s \psi(y - m(x))]^2$$

where the dependence of $v(\cdot)$ on x is omitted for simplicity. The next proposition yields an approximation to $\text{var } H_n(x)$.

Proposition 3.2. Suppose that (A1)-(A5) hold, then

$$\text{var } H_n(x) = (nh_n)^{-1} \int K^2(u) v(x - uh_n) du + O(n^{-\delta} h_n^{-1}) + O(n^{-2} h_n^{-2})$$

The proof is similar to that of Lemma 3.1 and is therefore omitted. From the last two propositions the MSE rate and the MSE at the optimal rate may be computed.

Proposition 3.3. Suppose that assumptions (A1)-(A5) hold and that K is a kernel of order p . Then

$$\begin{aligned} \text{MSE}(x) &= (nh_n)^{-1} \int_{-A}^A K^2(u) v(x - uh_n) du / \gamma^2(x) \\ &+ h_n^{2p} / (p!)^2 \left(\int_{-A}^A K(u) u^p du \right)^2 \frac{d^p m}{dx^p}(x) \end{aligned}$$

and the MSE optimal bandwidth is

$$h_n^* = C(x) \cdot \left(\frac{(p!)^2}{2p} \cdot \frac{\left(\int K^2(u) du \right)}{\left(\int K(u) u^p du \right)^2} \cdot \frac{1}{n} \right)^{1/(2p+1)},$$

where

$$C(x) = \left[\sigma^2(x) / \left((\gamma(x))^2 \frac{d^p m}{dx^p}(x) \right) \right]^{1/(2p+1)}.$$

For comparison we state this constant for $m_n^*(x)$ in the case that $f(y; x) = f_0(y - m(x))$ with a fixed p.d.f. $f_0(y)$. Here, the constant occurring above, turns out to be (Gasser and Müller, 1979)

$$C^*(x) = \left[\sigma^{*2}(x) / \frac{d^p m}{dx^p}(x) \right]^{1/(2p+1)},$$

where $\sigma^{*2}(x) = E_x(y - m(x))^2$ is constant. $C(x)/C^*(x)$ is thus $\sigma^2(x)/[\sigma^{*2}(x)(\gamma(x))^2]$ in general not equal to unity.

Optimal kernels can now be constructed in the same way it was done by Gasser and Müller (1979).

We may now proceed to two stages. First, we may optimize the “smoothing part” $\int K^2(u) du$ (also occurring in density estimation; Rosenblatt, 1971), secondly, we may optimize the “robustness part” $\sigma^2(x)/[\gamma(x)]^2$ of the asymptotic variance. The first optimization was studied by Gasser *et al.* (1982). To look at the robustness part, let us assume that

$$f(y; x) = (1 - \epsilon) \varphi(y - m(x)) + \epsilon h(y - m(x)),$$

with φ the p.d.f. of the standard normal distribution and $h(x)$ the heavy tailed double exponential p.d.f. Then from Table 1 in Huber (1964) we obtain the asymptotic variance $\sigma^2(x)/(\gamma(x))^2$. Suppose that $\epsilon = 10$ per cent and we use $\psi(u)$ as defined in (4) with varying κ , we then obtain the following quantities for $\sigma^{*2}(x)/[\sigma^2(x)/(\gamma(x))^2]$.

Table 1 shows also the convexity of $\sigma^2(x)/(\gamma(x))^2$ as a parameter of κ , the folding point of Huber’s ψ -function (4). The optimal κ corresponding to the contamination rate $\epsilon = 10$ per cent is about 1.2.

Example 3.1

In physical chemistry the analysis of Raman-spectra is used to identify the location and size of peaks and troughs of spectral bands. Huge spikes added to relatively small instrumental noise are introduced by vibrations of the experimental installation or simply by switch-on operations generating a small peak in the electrical circuit. In Fig. 1 a typical spectrum is presented. Estimating with $m_n^*(x)$ leads to the rather spurious two neighbouring peaks in Fig. 2; using $m_n(x)$ the robust candidate with ψ -function as in (4) ($\kappa = 0.9, h_n = 9$) gives us the rather insensitive estimate presented in Fig. 3).

TABLE 1
 Relative efficiencies of $m_n^*(x)$ vs $m_n(x)$ with varying cut off point x

κ	$\sigma^2(x)/(\gamma(x))^2$	$\sigma^{*2}(x)$	$\sigma^{*2}(x) (\gamma(x))^2/\sigma^2(x)$
0.2	1.778	2.309	1.29
0.4	1.659		1.39
0.6	1.576		1.46
0.8	1.523		1.51
1.0	1.495		1.544
1.2	1.491		1.548
1.4	1.507		1.532
1.6	1.542		1.497
1.8	1.595		1.44
2.0	1.665		1.38
2.2	1.75		1.31
2.4	1.85		1.24
2.6	1.963		1.17
2.8	2.09	1.10	
3.0	2.229	1.035	

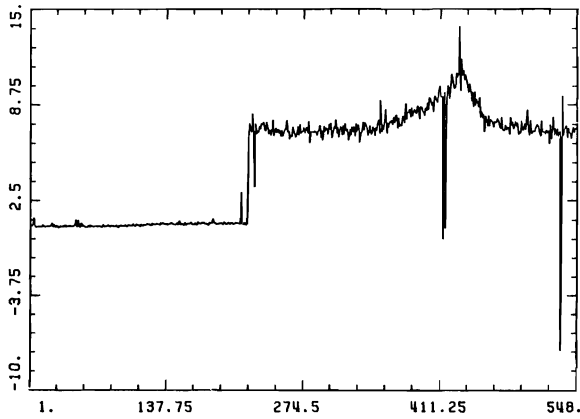


Fig. 1.

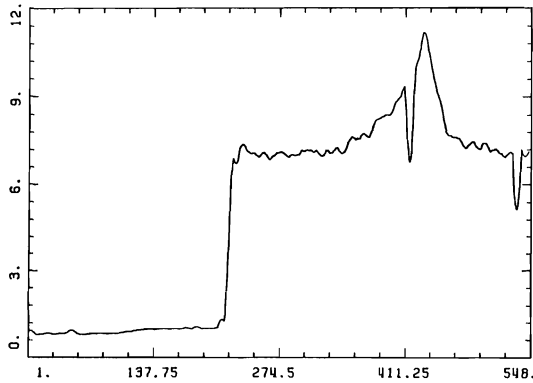


Fig. 2.

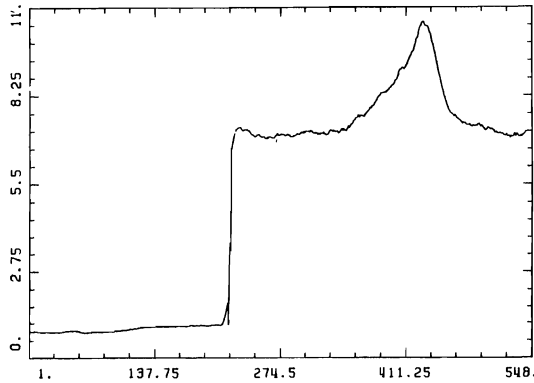


Fig. 3.

In a Monte Carlo investigation we studied the difference between $m_n(x)$ and $m_n^*(x)$ in predicting peaks and troughs. We choose $m(x) = 0.5x + 1$ on $x_i = i, i = 1, \dots, 100$, and residuals p.d.f.'s

$$f(y; x) = \frac{2}{10} \varphi(y - m(x)) + \frac{1}{30} \varphi(y - m(x))/3,$$

φ denoting the p.d.f. of the standard normal distribution. After smoothing with a certain bandwidth h_n , $m_n^*(x)$ ($m_n(x)$) was said to produce a peak i.f.f.

$$m_n^*(x) (m_n(x)) > m(x) + t_0,$$

where t_0 denotes some tolerance level. The same for troughs respectively. The results for different h_n and t_0 are shown in Table 2.

TABLE 2

$h_n \backslash t_0$	2.5		3.0		3.5		method
	$m_n(x)$	$m_n^*(x)$	$m_n(x)$	$m_n^*(x)$	$m_n(x)$	$m_n^*(x)$	
2.5	16	105	12	47	4	12	Peaks
	24	133	1	44	4	12	Troughs
3.5	7	85	0	36	0	5	Peaks
	8	80	0	29	0	1	Troughs
4.5	4	4	1	1	0	1	Peaks
	1	4	0	2	0	0	Troughs
5.5	3	10	0	0	0	0	Peaks
	5	9	1	3	0	0	Troughs

The number of peaks and troughs in 1000 Monte Carlo replications is shown. It is easy to see that the robust method predicts much less extrema than the linear method based on $m_n^*(x)$.

In particular as to be expected for low tolerance level t_0 ($t_0 = 2.5$) the difference is drastic. When both h_n and t_0 are chosen big enough the performance of $m_n^*(x)$ and $m_n(x)$ is quite similar, but for small bandwidth h_n , $m_n(x)$ still detects less extrema than $m_n^*(x)$.

4. CONCLUDING REMARKS

Based on the asymptotic variance of $m_n(x)$, computed in theorem 2.1, we may now carry out the same optimality considerations as in Huber (1981, Chapter 4). Note that the asymptotic variance splits up into two factors

$$S = \int K^2(u) du \text{ ("smoothing part")}$$

and

$$R = \int \psi^2(y - m(x)) f(y; x) dy / [\int \psi'(y - m(x)) f(y; x) dy]^2 \text{ ("robustness part").}$$

Since S does not depend on the family of residual distributions $\{f(y; x): x \in (0, 1)\}$ it suffices to consider optimality of R only. From Huber (1981, Chapter 4). We obtain the following corollary.

Corollary 4.1. Let $f(y; x) = (1 - \epsilon)g(y - m(x)) + \epsilon h(y - m(x))$, where $\epsilon = \epsilon(x)$, g fixed p.d.f., h arbitrary symmetric and $-\log g(\cdot - m(x))$ convex. Then

$$R = R(\psi, f) = E_x \psi^2(y - m(x)) / [E_x \psi'(y - m(x))]^2$$

has a saddlepoint, i.e. there exists a p.d.f. f_0 and a function ψ_0 such that

$$\sup_f R(\psi_0, f) = R(\psi_0, f_0) = \inf_{\psi} R(\psi, f_0).$$

This corollary shows that $m_n(x)$ is robust in a strict sense, i.e. $m_n(x)$ is minimax as f varies over the contamination neighbourhood $\{f = (1 - \epsilon)g + \epsilon h\}$.

ACKNOWLEDGEMENT

This research was financed through the Deutsche Forschungsgemeinschaft.

REFERENCES

- Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. and Tukey, J. W. (1972) *Robust Estimation of Location*. Princeton: University Press.
- Benedetti, J. K. (1977) On the non-parametric estimation of regression functions. *J. R. Statist. Soc. B*, **39**, 248–253.
- Cheng, K. F. and Lin, P. E. (1981) Non-parametric estimation of a regression function. *Z. f. Wahrsch. u. verw. Gebiete*, **57**, 223–233.
- Cleveland, W. S. (1977) Robust locally weighted regressions and smoothing scatterplots. *J. Amer. Statist. Ass.*, **74**, 829–836.
- Collomb, G. (1981) Estimation non-paramétrique de la régression. *Revue Bibliographique. Int. Statist. Rev.*, **6**, 75–93.
- Gasser, T. and Müller, H. G. (1979) Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*. Springer Lecture Note 757. (ed. T. Gasser, M. Rosenblatt).
- Gasser, T., Müller, H. G. and Mammitsch, V. (1982) Kernels for non-parametric curve estimation. Preprint University of Heidelberg.
- Hampel, F. R. (1973) Robust estimation. A condensed partial survey. *Z. f. Wahrsch. u. verw. Gebiete*, **27**, 87–104.
- Huber, P. J. (1964) Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73–101.
- (1981) *Robust Statistics*. New York: Wiley.
- Johnston, G. (1982) Probabilities of maximal deviations for non-parametric regression function estimates. *J. Multiv. Anal.* **12**, 402–414.
- Parzen, E. (1962) On estimation of a probability density function. *Ann. Math. Statist.*, **33**, 1065–1076.
- Priestley, M. B. and Chao, M. T. (1972) Non-parametric function fitting. *J. R. Statist. Soc. B*, **34**, 385–392.
- Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.*, **27**, 832–837.
- (1969) Conditional probability density and regression estimators. In *Multivariate Analysis II* (P. R. Krishnaiah, ed.), pp. 25–31. New York: Academic Press.
- (1971) Curve estimates. *Ann. Math. Statist.*, **42**, 1815–1842.
- Schuster, E. F. (1972) Joint asymptotic distributions of the estimated regression function at a finite number of points. *Ann. Math. Statist.*, **43**, 84–88.

UNIFORM CONSISTENCY OF A CLASS OF REGRESSION FUNCTION ESTIMATORS¹

BY W. HÄRDLE AND S. LUCKHAUS

University of Frankfurt and University of Heidelberg

We study a wide class of nonparametric regression function estimators including kernel estimators and robust smoothers. Under different assumptions on the kernel and the sequence of bandwidths, we obtain weak uniform consistency rates on a bounded interval. The uniform consistency is shown in a "stochastic design model" and in a "fixed design model".

1. Introduction. Let $(X_1, Y_1), (X_2, Y_2), \dots$ be independent bivariate random data sampled either with stochastic design rv's X_1, X_2, \dots or with fixed design points x_1, x_2, \dots . In the stochastic design model $(X_1, Y_1), (X_2, Y_2), \dots$ are independent bivariate random variables identically distributed as a bivariate random variable (X, Y) whose joint cumulative distribution function is F and whose joint probability density is $f(x, y)$. In the fixed design model (noisy sampled data) we have an underlying family of probability density functions $\{f(\cdot; x): x \in [0, 1]\}$ and $\mathcal{P}_n = \{x_1, x_2, \dots, x_n\}$ where $0 \leq x_1 \leq x_2 \leq \dots \leq x_n = 1$ is a partition of $[0, 1]$ determined by the experimenter.

The nonparametric regression problem is the problem of estimating the regression curve of Y on X . Equivalently, the nonparametric regression problem requires finding $m(x) = m_{\psi, F}(x)$, given observations

$$\{X_i, m(X_i) + N_i\}_{i=1}^n.$$

The function ψ is used here as an indexing parameter, since, as is shown in examples below, the shape of ψ determines the regression curve $m(x)$. Different choices of ψ yield the conditional mean or the conditional median for instance. The N_i being an independent noise variable which may depend on X_i and $m_{\psi, F}$ being the trend satisfying

$$(1.1) \quad E_x \psi(Y - m(x)) = 0$$

where $E_x(\cdot) = E(\cdot | X = x)$ in the stochastic design case or $E_x(\cdot) = \int \cdot f(y; x) dy$ in the fixed case and $\psi(\cdot)$ is a monotone continuous function.

We propose to estimate $m(x)$ by $m_n(x)$ a solution (with respect to θ) of

$$(1.2) \quad \sum_{i=1}^n \alpha_i(x) \psi(Y_i - \theta) = 0$$

where $\alpha_i(x) = \alpha_i^{(n)}(x)$ are (localizing) weights depending on X . In the present

Received June 1982; revised May 1983.

¹This research was made possible by the Deutsche Forschungsgemeinschaft. Sonderforschungsbereich 123 "Stochastische Mathematische Modelle" and Air Force Scientific Research Contract AFOSR-F49620-82-C-0009.

AMS 1980 subject classification. Primary 62G15; secondary 60E15, 62F25.

Key words and phrases. Regression, robust smoothing, kernel estimators, uniform convergence rates.

paper we derive—under mild conditions on the weight sequence $\alpha_i^{(n)}(\cdot)$ —the uniform consistency of $m_n(\cdot)$ on the interval $I = [0, 1]$. We show that

$$(UC) \quad r_n^{-1} \sup_{0 \leq t \leq 1} |m_n(t) - m(t)| = O_p(1)$$

with rate $r = r_n$. In the derivation of this result we shall need bounds on moments of sums of independent rv's, as given by Whittle (1960), Theorem 2.

The quite general setup of $m_n(x)$ as the solution of (1.2) and $m(x)$ as the solution of (1.1) allows us by tuning $\alpha_i(\cdot)$ and $\psi(\cdot)$ to obtain a wide class of estimators and regression functions as will be shown in the following examples.

One of the following examples (Example 5) will give a partial answer to a question raised by C. J. Stone in his special invited paper on optimal rates of convergence (Stone, 1982, page 1044, Question 4).

EXAMPLE 1. Take $\psi(u) = u$ in both (1.1) and (1.2) and define

$$\alpha_i(x) = n^{-1}h^{-1}K((x - X_i)/h)$$

for kernel $K(\cdot)$ and a sequence of bandwidth $h = h(n)$ tending to zero. The resulting regression curve in the stochastic design case is

$$m_{\psi,F}(x) = m(x) = E(Y | X = x)$$

and the estimator is

$$m_n^*(x) = (nh)^{-1} \sum_{i=1}^n K((x - X_i)/h) Y_i / [(nh)^{-1} \sum_{j=1}^n K((x - X_j)/h)].$$

The estimator was proposed independently by Nadaraya (1964) and Watson (1964). Rosenblatt (1969) and Collomb (1977, 1979) computed bias and variance rates. Schuster (1972) demonstrated the multivariate normality at a finite number of distinct points; Schuster and Yakowitz (1979) derived uniform consistency of $m_n^*(x)$ on a finite interval. Recently, Johnston (1979) in his thesis proved a uniform consistency result (with rates) for the related estimator

$$m_n^*(x) \cdot [(nh)^{-1} \sum_{i=1}^n K((x - X_i)/h) / f_X(x)],$$

$f_X(\cdot)$ denoting the marginal density of X . Further (uniform) consistency results for $m_n^*(x)$ were obtained by Major (1973), Konakov (1977), Nadaraya (1973, 1974), Stone (1977) among others. A bibliographic review on the estimation of $m(x) = E(Y | X = x)$ may be found in Collomb (1981).

EXAMPLE 2. Take $\psi(u) = u$ and define in the fixed design case

$$\alpha_i(x) = h^{-1} \int_{s_{i-1}}^{s_i} K\left(\frac{x - u}{h}\right) du$$

where $s_0 = 0, s_{j-1} \leq x_j \leq s_j, s_n = 1, \int K(u) du = 1$ and $h = h(n)$ is as above a sequence of bandwidths tending to zero as $n \rightarrow \infty$. Since $\sum_{i=1}^n \alpha_i(x) = 1$, the resulting estimator is

$$\bar{m}_n(x) = \sum_{i=1}^n \alpha_i(x) Y_i,$$

first discussed by Gasser and Müller (1979) and recently considered by Cheng

and Lin (1981) (with $s_j \equiv x_j$ in $\alpha_j(x)$). The Priestley and Chao (1972) estimator does not fall in the class of estimators here, but is, as shown by Cheng and Lin (1981), also uniform consistent obtaining the same rate as $\bar{m}_n(x)$.

In the following example we will assume symmetry of $f(y|x)$. Note that for the results of this paper neither symmetry of $f(y|x)$ nor antisymmetry of ψ are required. This assumption is only made to obtain in a convenient way the conditional mean from equation (1.1).

EXAMPLE 3. Take $f(y|x)$ respectively $f(y;x)$ be symmetric and ψ a bounded, antisymmetric function. Then again (1.1) gives for the stochastic design case

$$m_{\psi,F}(x) = m(x) = E(Y|X = x)$$

respectively

$$m(x) = \int yf(y; x) dy$$

in the fixed design case. The regression curve is thus a quantity $m_{\psi,F}(x)$ which minimizes (w.r.t. θ)

$$\int \rho(Y - \theta)f(y|x) dy$$

where we assume ρ to be positive, even, convex and differentiable with derivative $\rho' = \psi$. This is exactly the notation of a M -functional (Bickel and Lehmann, 1975, page 1053) and shows that $m_n(x)$ from (1.2) (with weights $\{\alpha_i(x)\}$ as in example 1 or example 2) is a robust estimator of $m(x)$.

In the stochastic design case $m_n(x)$ is a solution (with respect to θ) of

$$n^{-1}h^{-1} \sum_{i=1}^n K((x - X_i)/h)\psi(Y_i - \theta) = 0.$$

In the fixed design case the estimator is a solution (with respect to θ) of

$$h^{-1} \sum_{i=1}^n \left[\int_{i-1}^{s_i} K\left(\frac{x-u}{h}\right) du \right] \psi(Y_i - \theta) = 0.$$

Pointwise consistency and asymptotic normality along with some numerical results are shown in Härdle (1983) and Härdle and Gasser (1982). In the last paper it is also shown that the robust estimator $m_n(x)$ proves to be useful in the evaluation of Laser spectra (Raman spectra). If we take for instance

$$\psi(u) = \max\{-\kappa, \min\{u, \kappa\}\}, \quad \kappa \geq 0$$

we obtain a Huber-type (Huber, 1964) robust nonparametric regression function estimator. Bias and variance rates for this Huber-type estimator with a uniform window, i.e. $K(u) = I_{[-.5,.5]}(u)$ were computed by Stuetzle and Mittal (1979).

EXAMPLE 4. Taking $\psi(u) = \alpha u^{\alpha-1}$, $u \geq 0$ and $\psi(u) = -\alpha(-u)^{\alpha-1}$, $u < 0$, $1 < \alpha < 2$ allows us by tuning α to steer from the (local) least square estimator, which is $m_n^*(x)$ as $\alpha = 2$, to the (local) median (as $\alpha \rightarrow 1$) and vice versa. The whole class of these estimators will also be covered by our theorems.

EXAMPLE 5. Take $\psi(u) = \frac{1}{2} - I(u \leq 0)$, a ψ -function leading to the conditional median $m(x) = \text{med}(Y | X = x)$ as the regression curve. Stone (1982) raised the question if $\{n^{-r}\}$ ($r = (p - m)/(2p + d)$ in his notation) is still an achievable rate of convergence. The results of this paper give a partial answer to that question. We show that for $p = 1, d = 1, m = 0$ a subclass of his $\{T(\theta)\}$ indeed, $\{(n^{-1} \log n)^r\}$, the optimal rate of his Theorem 1 (for the type of distance considered here) is achieved. To see this in the "stochastic design model", note that assumption (A4) of Section 2 is trivially fulfilled and assumption (A3) is satisfied if there exists a constant c_0 such that $f(m(x) | x) > 2c_0, x \in I$. Assumptions (A1) and (A5) are only technical and (A2) is the definition of $m(x) = \text{med}(Y | X = x)$. Assume now that m is continuously differentiable so that the modulus of continuity $\omega_m(\delta)$ is linear in δ . Then, Theorem 1 below says that with $r_n \sim h_n$ and $r_n \sim (\log n)^{1/2} (nh_n)^{-1/2}$ uniform consistency of m_n can be achieved with rate $r_n = n^{-1/3} (\log n)^{1/3}$ which is the optimal rate given in Stone (1982). Quite analogous conclusions can be drawn in the fixed design model.

We present the result (UC) for $\alpha_i(x)$ as in example 1 and example 2 for the stochastic design case in Theorem 1 and a following remark. Theorem 2 shows (UC) in the fixed design case with $\alpha_i(x)$ as in example 2. All theorems require a certain amount of smoothness of $m(\cdot)$, expressed through the behaviour of the modulus of continuity of m which we denote by ω_m . These results are improvements over some previous work. Our assumptions are weaker than those of Major (1973) in that Y is not required to be bounded a.s. and our results are stronger than those of Schuster and Yakowitz (1978) because we were able to compute uniform convergence rates for $m_n^*(x)$ as in Mack and Silverman (1982).

2. Results. We will make the following assumptions on the kernel function and on moments of $[\psi(Y - m(x) + s)]$.

(A1) The kernel K is positive, continuously differentiable with compact support

$$[-A, A] \quad \text{and} \quad \int_{-A}^0 K(u) du = \int_0^A K(u) du = \frac{1}{2}.$$

(A2) ψ is a monotone, locally bounded function with $E_x \psi(Y - m(x)) = 0$.

(A3) There are constants $c_0, c_1 > 0$ such that for every $x \in I = [0, 1]$

$$|E_x \psi(Y - m(x) + s)| > c_0 |s|, \quad |s| \leq c_1.$$

(A4,k) For some $k \geq 2$ let $\sup_{x \in I} E_x |\psi(Y - m(x) \pm c_1)|^k < \infty$.

(A4, ∞) ψ is bounded, $\sup_{u \in \mathbb{R}} |\psi(u)| \leq B_\psi < \infty$.

(A5) The marginal density of X is bounded from above and below

$$0 < a \leq f_X(u) \leq b < \infty \quad \text{for all } u \in I.$$

(A6) There exists a constant C_0 such that for every $x \in I$

$$|E_x \psi(Y - m(x) + s)| < C_0 |s|.$$

Some remarks about the assumptions should be made. The first assumption is

very common in nonparametric regression and needs no further explanation (Collomb, 1981). The second assumption is just the proper (implicit) definition of the regression function. Assumption (A3) needs some more motivation. Assume for simplicity that we have a homoscedastic error structure that is $f(y|x) = f(y - m(x))$ and $f(y|x)$ is symmetric. If we have that $\psi(u) = u$ then (A3) is trivially fulfilled. For the nonlinear ψ functions, (A3) is satisfied if $|\int \psi(y + s)/sf(y) dy| > c_0$ for small s . So (A3) can be interpreted as a criterion for $E_x \psi'(y + s)$, (s small) staying away from zero, provided it exists at all. Assumption (A6) is trivially fulfilled for $\psi(u) = u$. For nonlinear ψ functions (A6) is obviously fulfilled if $|\int \psi(y + s)/sf(y) dy| < C_0$, which can be interpreted as an upper bound for $E_x \psi'(y - m(x) + s)$, s small. We have chosen this quite technical way of formulation to include the conditional median corresponding to $\psi(u) = \frac{1}{2} - I(u \leq 0)$ which is nondifferentiable at $u = 0$. The assumption (A4,k) will be used for unbounded ψ functions only, (A4, ∞) is just the definition of a bounded ψ function making m_n a robust estimator of m .

As already mentioned, the modulus of continuity of m will be denoted by

$$\omega_m(\delta) = \sup_{x \in I} \sup_{|x-x'| < \delta} |m(x) - m(x')|.$$

As long as there is no confusion, the index “ n ” will be dropped in the sequel.

The following theorems will split up into a statement on unbounded ψ functions (i.e. containing as a special case the Nadaraya-Watson estimator) and a statement on bounded ψ functions. The theorems tell us how we have to choose the sequence $h = h(n)$ in dependence of the sample size n and the rate $r = r_n$ in order to obtain (UC).

We begin with the uniform consistency in the stochastic design case.

THEOREM 1. *Let the data be generated with stochastic design $\{X_i\}_{i=1}^n$, and let $\alpha_i(t) = (nh)^{-1}K((t - X_i)/h)$. Assume that (A1)–(A5) hold and let*

$$\omega_m(2Ah) < r, \quad nh^2/\log n \geq d > 0.$$

If (A4,k) holds let

$$nh^{1+2/(k-1)}r^{2+2/(k-1)} \rightarrow \infty$$

and if (A4, ∞) holds

$$nhr^2/\log n \geq \xi_1$$

ξ_1 depending on c_0, c_1, B_ψ, a, b . Then $m_n(x)$ satisfies (UC). If in addition (A6) holds, only

$$nr^2h^{1+2/(k-1)} \rightarrow \infty$$

and (A4,k) suffice to establish (UC).

REMARK. It can be shown that (UC) also holds for the situation described in example 2 for the stochastic design case. Very similar arguments that are used

to prove Theorem 2 yield that if

$$\begin{aligned} nh^{1+2/(k-1)}r^{2+2/(k-1)}/\log n &\rightarrow \infty && \text{in case of (A4,k),} \\ nhr^2/(\log n)^2 &\geq \xi_2 && \text{in case of (A4,\infty)} \end{aligned}$$

the uniform consistency (UC) with rate $r = r_n$ follows.

THEOREM 2. *Let the data be generated with fixed design points $\{x_i\}_{i=1}^n$, satisfying $\sup_i |x_i - x_{i-1}| = O(n^{-1})$ and set $\alpha_i(t)$ as in example 2. Assume that (A1)-(A5) hold and $\omega_m(2Ah) < r$. If (A4,k) holds, let*

$$nh^{1+2/(k-1)}r^{2+2/(k-1)} \rightarrow \infty$$

and if (A4,\infty) holds

$$nhr^2/\log n \geq \xi_2$$

ξ_2 depending on $c_1, c_0, B_\psi, \int K^2$. Then $m_n(x)$ satisfies (UC). If in addition (A6) holds, the condition

$$nh^{1+2/(k-1)}r^2 \rightarrow \infty$$

together with (A4,k) suffice to establish (UC).

3. Proofs. To show that the class of estimators defined through (1.2) satisfies (UC) for the various choices of ψ -functions and weights $\{\alpha_i(x)\}_{i=1}^n$, we have to show that

$$P\{\sup_{x \in I} |m_n(x) - m(x)| > r_n\}$$

is arbitrarily small. Now by monotonicity of ψ , this can be estimated by

$$P(\Omega_n) + P(\Omega'_n)$$

where

$$\Omega_n = \{\sup_{x \in I} g_n(x, -r) \geq 0\}, \quad \Omega'_n = \{\inf_{x \in I} g_n(x, r) \leq 0\}$$

and

$$g_n(x, s) = \sum_{i=1}^n \alpha_i(x) \psi(Y_i - m(x) + s).$$

By the symmetry of the problem it will suffice to consider $P(\Omega_n)$.

The principal idea of the proof is to lay an equidistant mesh $0 = t_0 < t_1 < \dots < t_{\ell_n} = 1$, where $\ell_n \ll n$, to sum the probabilities at the meshpoints and to use the mean value theorem applied to $\alpha_i(t)$ between them. More precisely we have

$$\begin{aligned} P(\Omega_n) &\leq \ell_n \sup_{t=t_j} P[\{g_n(t, -r_n/2) \geq -\eta_n\} \cap M_n] \\ &\quad + \ell_n \sup_{t=t_j} P[\{\sup_{|u-t| < \ell_n^{-1}} g_n(u, -r_n/2) > \eta_n + g_n(t, -r_n/2)\} \cap M_n] \\ &\quad + P(M_n^c) = \ell_n [\sup_{t=t_j} U_{1n}(t) + \sup_{t=t_j} U_{2n}(t)] + U_{3n}, \end{aligned}$$

where ℓ_n, η_n are arbitrary sequences to be specified later and M_n is an arbitrary set to be chosen for the different cases (stochastic/fixed design and the particular

$\{\alpha_i(x)\}_{i=1}^n$. We will also make use of the following fact that in the fixed design

$$\sum_{i=1}^n \alpha_i^2(t) = O(n^{-1}h^{-1}) \quad \text{uniformly in } t$$

(Gasser and Müller, 1979) and in the stochastic design

$$\sum_{i=1}^n E\alpha_i^2(t) = O(n^{-1}h^{-1}) \quad \text{uniformly in } t$$

(Johnston, 1979).

PROOF OF THEOREM 1. Suppose that (A4,k) holds, then with $\eta_n = \beta r_n$, β small enough to satisfy the assumptions of Lemma 1, we obtain from (A.1)

$$\sup_{t=t_j} U_{1n}(t) \leq \nu_1 r^{-k} (nh)^{-k/2},$$

and if $\ell_n^{-1} < Ah$ we have from (A.3)

$$\sup_{t=t_j} U_{2n}(t) \leq \nu_2 (r\ell_n)^{-k} [h^{-k} + (nh^3)^{-k/2}],$$

and if (A6) holds

$$\sup_{t=t_j} U_{2n}(t) \leq \nu'_2 (r\ell_n)^{-k} [h^{-k} r^{-k} + (nh^3)^{-k/2}]$$

where ν denote large constants and M_n is chosen as in Lemma 3. Then with $\ell_n^2 = nh^{-1}$ (such that $h^{-1} \ll \ell_n \ll n$) we have from Lemma 3

$$\begin{aligned} P(\Omega_n) &\leq \mu_1 \ell_n \{(nhr^2)^{-k/2} + (\ell_n^2 r^2 h^2)^{-k/2} + [nh(\ell_n^2 r^2 h^2)]^{-k/2}\} \\ &\quad + \mu_2 \ell_n \exp(-\mu_3 nh^2) \\ &\leq \mu_4 (nh^{-1})^{1/2} (nhr^2)^{-k/2} + \mu_2 \exp(\log n - \mu_3 nh^2) \end{aligned}$$

which is small by the assumptions of the theorem. A similar inequality shows that if (A6) is fulfilled, $P(\Omega_n)$ can be made arbitrarily small. Suppose now that (A4,∞) holds and choose ℓ_n such that

$$\sum_{i=1}^n |\alpha'_i(t)| \leq \sup |K'| |n^{-1}h^{-2}4b\epsilon^{-1}n(Ah + \ell_n^{-1}) \leq \eta_n \ell_n / (2B_\psi)$$

to fulfill the assumptions of Lemma 2. This can be made for $\eta_n = \beta r_n$ and $r\ell_n h \geq \mu_5$, μ_5 a large constant. So we get by Lemma 1 and Lemma 2

$$P(\Omega_n) \leq \mu_6 \exp(-\mu_7 r^2 nh - \log r - \log h) + \mu_8 \exp(-\epsilon^{-1} nh^2 - \log r - \log h)$$

which is small by the assumptions of the theorem.

PROOF OF THEOREM 2. Define $M_n = \{(x_1, x_2, \dots, x_n) : \sup_{2 \leq i \leq n-1} |s_{i+1} - s_{i-1}| < \gamma/n\}$. If γ is chosen large enough we have that $M_n^c = \phi$ by assumption on the fixed design. Since

$$\alpha_i(t) = h^{-1} \int_{s_{i-1}}^{s_i} K\left(\frac{t-u}{h}\right) du$$

we have

$$\begin{aligned} \sum_{i=1}^n \alpha_i(t) &= \int_{(t-Ah, t+Ah) \cap I} K(u) du \geq \frac{1}{2} \\ \sum_{i=1}^n |\alpha_i(t)|^2 &\leq h^{-1} \sup |s_{i+1} - s_{i-1}| \int |K^2| \\ \sum_{i=1}^n |\alpha'_i(t)| &\leq h^{-1} \int |K'|. \end{aligned}$$

Choosing $\eta_n = \beta r_n$, β small enough, we get from (A1) and (A4,k)

$$\begin{aligned} U_{1n}(t) &\leq \mu_9 r^{-k} [1/(nh)]^{k/2} \\ U_{2n}(t) &\leq \mu_{10} [(\ell_n r h)^{-k} + (1/(nh^3))^{k/2} (r \ell_n)^{-k}]. \end{aligned}$$

Respectively if (A6) holds

$$U_{2n}(t) \leq \mu_{11} [(\ell_n h)^{-k} + (1/(nh^3))^{k/2} (r \ell_n)^{-k}]$$

taking $\ell_n^2 = nh^{-1}$ respectively $\ell_n^2 = nh^{-1} r^2$ (for the (A6) case) shows that $P(\Omega_n)$ is small. Now in the case that ψ is bounded we see that

$$\ell_n = \varepsilon^{-1} B_\psi \int |K'| / (c_0 h), \quad \varepsilon \text{ small}$$

ensures

$$U_{2n}(t) = 0$$

and

$$U_{1n}(t) \leq \mu_{12} \exp(-\mu_{13} r^2 n h - \log h).$$

APPENDIX

It is shown here how the terms $U_{1n}(t)$, $U_{2n}(t)$, U_{3n} may be estimated in the different cases (stochastic design, fixed design). Lemma 1 and Lemma 2 are shown for the fixed design case only. The proofs for the stochastic design case are essentially the same by conditioning on $\{X_1, \dots, X_n\}$.

LEMMA 1. Suppose that the modulus of continuity of $m(\cdot)$ satisfies $\omega_m(Ah_n) \leq r_n/4$ and let $\eta_n \leq c_0 \delta r_n/8$ where δ is a small constant, c_0 is the constant of (A3) and $M_n \subset \{\sum_{i=1}^n \alpha_i(t) > \delta\}$. Then if (A4,k) holds

$$\begin{aligned} (A.1) \quad U_{1n}(t) &\leq \eta_n^{-k} \Lambda_k^{(1)} [\sup_{\mathbf{x} \in M_n} (\sum_{i=1}^n \alpha_i^2(t))^{k/2}] \\ &\quad \times \sup_{0 \leq x \leq 1} E_x(|\psi(Y - m(x) \pm c_1)^k|). \end{aligned}$$

Otherwise if (A4,∞) holds

$$(A.2) \quad U_{1n}(t) \leq \exp[-\Lambda_\infty^{(1)} \eta_n^2 (B_\psi^2 \sum_{i=1}^n \alpha_i^2(t))^{-1}]$$

where $\Lambda^{(1)}$ denote constants.

PROOF. Using the assumption on ω_m and the monotonicity of ψ near the origin we have

$$E_{x_i}\psi(Y_i - m(t) - r_n/2) < E_{x_i}\psi(Y_i - m(x_i) - r_n/4)$$

for all $i \in \{j: |t - x_j| < Ah\}$ and therefore

$$\begin{aligned} \sum_{i=1}^n \alpha_i(t)[\psi(Y_i - m(t) - r_n/2) - E_{x_i}\psi(Y_i - m(t) - r_n/2)] \\ > c_0 r_n/4 \sum_{i=1}^n \alpha_i(t) + g_n(t, -r_n/2) \end{aligned}$$

by assumption (A3). So by Chebychev's inequality and Theorem 2 of Whittle (1960) we have that

$$\begin{aligned} U_{1n}(t) &\leq P\{\sum_{i=1}^n \alpha_i(t)[\psi(Y_i - m(t) - r_n/2) - E_{x_i}\psi(Y_i - m(t) - r_n/2)] > \eta_n\} \\ &\leq \eta_n^{-k} \lambda_k [\sum_{i=1}^n \alpha_i^2(t)]^{k/2} \\ &\quad \times \sup_{|t-x_i| < Ah} E_{x_i}[|\psi(Y - m(t) - r_n/2) - E_{x_i}\psi(Y - m(t) - r_n/2)|^k] \end{aligned}$$

in the case that (A4,k) is used. Otherwise, if (A4,∞) holds, by an easy extension of Whittle's Theorem 2

$$U_{1n}(t) \leq \exp[-\eta_n^2 (4eB_\psi^2 \sum_{i=1}^n \alpha_i^2(t))^{-1}]$$

for bounded ψ functions which shows that (A.1), (A.2) hold. \square

The next lemma estimates $U_{2n}(t)$.

LEMMA 2. Suppose that the modulus of continuity of $m(\cdot)$ satisfies $\omega_m(\ell_n^{-1} + Ah) < r_n/2$. Then, if (A4,k) holds

$$\begin{aligned} U_{2n}(t) &\leq \eta_n^{-k} \ell_n^{-k} \Lambda_k^{(2)} \{ \sup_{\mathbf{x} \in M_n, u \in I} |\sum_{i=1}^n \alpha'_i(u)|^k \sup_{u \in I} E_x |\psi(Y - m(x) - r_n)|^k \\ &\quad + \sup_{\mathbf{x} \in M_n, u \in I} [\sum_{i=1}^n [\alpha'_i(u)]^2]^{k/2} \sup_{u \in I} E_x |\psi(Y - m(x) \pm c_1)|^k \}, \end{aligned}$$

$\Lambda_k^{(2)}$ a constant.

On the other hand if (A4,∞) is true, then $U_{2n}(t) = 0$ provided that

$$M_n \subset \mathcal{M}_n = \{ \sum_{i=1}^n |\alpha'_i(t)| < \eta_n \ell_n / [2B_\psi] \}.$$

PROOF. By the assumption on the modulus of continuity of $m(\cdot)$ and the mean value theorem we conclude that

$$U_{2n}(t) \leq P \left\{ \int_{\Gamma_n(t)} \left| \sum_{i=1}^n \alpha'_i(u) \psi(Y_i - m(t) - r_n/2) \right| du > \eta_n \right\}$$

where $\Gamma_n(t) = \{u: |u - t| \leq \ell_n^{-1}\}$. This already shows that $U_{2n}(t) = 0$ if $M_n \subset \mathcal{M}_n$ and (A4,∞) holds.

We now further estimate the RHS of the inequality above using Chebyshev's

inequality. We then have

$$\begin{aligned}
 U_{2n}(t) &\leq \Lambda_k^{(2)} \eta_n^{-k} \left\{ E \left| \int_{\Gamma_n(t)} \sum_{i=1}^n \alpha'_i(u) E_{x_i} \left| \psi(Y_i - m(t) - r_n/2) \right| du \right|^k \right. \\
 &\quad \left. + E \left| \int_{\Gamma_n(t)} \left| \sum_{i=1}^n \alpha'_i(u) [\psi(Y_i - m(t) - r_n/2) \right. \right. \right. \\
 &\quad \quad \left. \left. \left. - E_{x_i} \left| \psi(Y_i - m(t) - r_n/2) \right| \right] du \right|^k \right\} \\
 &= V_{1n} + V_{2n}, \text{ say.}
 \end{aligned}$$

Now by Hölder's inequality (with $p = k$) and Theorem 2 of Whittle (1960), we have

$$\begin{aligned}
 V_{2n} &\leq \left[\int_{\Gamma_n(t)} du \right]^{k-1} E \int_{\Gamma_n(t)} \left| \sum_{i=1}^n \alpha'_i(u) [\psi(Y_i - m(t) - r_n/2) \right. \\
 &\quad \left. - E_{x_i} \left| \psi(Y_i - m(t) - r_n/2) \right| \right|^k du \\
 &\leq [2\ell_n^{-1}]^k \{ \sup_{u \in I, x \in M_n} [\sum_{i=1}^n [\alpha'_i(u)]^2]^{k/2} 2^k \\
 &\quad \times \sup_{u \in I, |u-x| < Ah + \ell_n^{-1}} E_{x_i} \left| \psi(Y_i - m(u) - r_n/2) \right|^k \}.
 \end{aligned}$$

Applying now the assumption on the modulus of continuity, we have the desired upper bound for both V_{1n} and V_{2n} (after an application of Hölder's inequality to V_{1n} , too). □

In the following lemma we estimate the term U_{3n} for different sets M_n .

LEMMA 3. *Let*

$$M_n = \{(X_1, \dots, X_n): \sum_{i=1}^n \alpha_i(t_j) > a/4 \text{ and}$$

$$\#\{|X_i - t_j| < Ah + \ell_n^{-1}\} < 4bn(Ah + \ell_n^{-1})\varepsilon^{-1} \text{ for } j = 0, \dots, \ell_n, 0 < \varepsilon \leq 1\}$$

in the stochastic design case. Then

$$U_{3n} \leq \Lambda^{(3)} \ell_n \exp[-\lambda_3 \varepsilon^{-1} n(h_n^2 + \ell_n^{-2})],$$

where $\Lambda^{(3)}, \lambda_3$ are constants.

PROOF. Since $E\alpha_i(t) = n^{-1} \int_{[-Ah+t, Ah+t] \cap I} K(u) f_x(t + uh) du \geq a(2n)^{-1}$, $\sum_{i=1}^n \alpha_i(t) \leq a/4$ implies $|\sum_{i=1}^n [\alpha_i(t) - E\alpha_i(t)]| > a/4$. Now by Whittle's theorem we have

$$P(|\sum_{i=1}^n [\alpha_i(t) - E\alpha_i(t)]| > a/4) \leq \exp\left(-\lambda_1 \frac{\sup K^2}{\inf f_X^2} n\right), \quad \lambda_1 = \text{const.}$$

On the other hand

$$\#\{|X_i - t| < Ah + \ell_n^{-1}\} = \sum_{i=1}^n I_{\Delta_n(t)}(X_i) = n^{-1} \sum_{i=1}^n \bar{Z}_i$$

where

$$\Delta_n(t) = \{u: |u - t| < Ah + \ell_n^{-1}\}.$$

Since $n^{-1}E\bar{Z}_i(t) = \int_{\Delta_n(t)} f_X(u) du \leq 2b(Ah + \ell_n^{-1})$ we have by Whittle's theorem that

$$P(n^{-1} \sum_{i=1}^n \bar{Z}_i \geq 4bn(Ah + \ell_n^{-1})\varepsilon^{-1}) \leq \exp[-\lambda_2 n(Ah + \ell_n^{-1})], \quad \lambda_2 = \text{const.} \quad \square$$

Acknowledgment. The authors would like to thank C. Jennen, R. Lerche and B. Silverman for many fruitful comments and suggestions. We also want to thank the referees for substantial improvements of the paper.

REFERENCES

- BICKEL, P. J. and LEHMANN, E. L. (1975). Descriptive statistics for nonparametric models II. *Ann. Statist.* **3** 1045-1069.
- CHENG, KUANG FU and LIN, PI-ERH. (1981). Nonparametric estimation of a regression function. *Z. Wahsch. verw. Gebiete* **57** 223-233.
- COLLOMB, G. (1979). Quelques propriétés de la methode du noyau pour l'estimation non-paramétrique de la régression en un point fixé. *C. R. Acad. Sci. Paris* **285A** 289-292.
- COLLOMB, G. (1979). Conditions nécessaires et suffisantes de convergence uniforme d'un estimateur de la régression, estimation des dérivées de la régression. *C. R. Acad. Sci. Paris* **288A** 161-164.
- COLLOMB, G. (1981). Estimation non-paramétrique de la régression: revue bibliographique. *ISR* 75-93.
- GASSER, T. and MÜLLER, H. G. (1979). Kernel estimation of regression functions. *Springer Lecture Notes in Math.* **757**.
- HÄRDLE, W. and GASSER, T. (1983). Robust nonparametric function fitting. *J. Roy. Statist. Soc. (B)*, to appear.
- HÄRDLE, W. (1984). Robust regression function estimation. *J. Multivariate Anal.*, to appear.
- HUBER, P. J. (1964). Robust estimation of location. *Ann. Math. Statist.* **35** 73-101.
- JOHNSTON, G. J. (1979). Smooth nonparametric regression analysis. Ph.D. dissertation, University of North Carolina, Chapel Hill.
- KONAKOV, V. D. (1977). On a global measure of deviation for an estimate of the regression line. *Theor. Probab. Appl.* **22** 858-868.
- MACK, Y. P. and SILVERMAN, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. verw. Gebiete* **61** 405-415.
- MAJOR, P. (1973). On a non-parametric estimation of the regression function. *Stud. Sci. Math. Hung.* **8** 347-361.
- NADARAYA, E. A. (1964). On estimating regression. *Theor. Probab. Appl.* **9** 141-142.
- NADARAYA, E. A. (1973). Some limit theorems related to nonparametric estimates of regression curve (in Russian). *Bull. Acad. Sci. Georgian S.S.R.* 71n n°, 57-60.
- NADARAYA, E. A. (1974). The limit distribution of the quadratic deviation of nonparametric estimates of the regression function. *Soobshch. Akad. Nauk. Gruz. SSR* 74, 33-36 (in Russian).
- PRIESTLEY, M. B. and CHAO, M. T. (1972). Nonparametric function fitting. *J. Roy. Statist. Soc. B* **34** 385-392.
- ROSENBLATT, M. (1969). Conditional probability density and regression estimates. In *Multivariate Analysis II*, ed. Krishnaiah.
- SCHUSTER, E. F. (1972). Joint asymptotic distribution of the estimated regression function at a finite number of distinct points. *Ann. Math. Statist.* **43** 84-88.

- SCHUSTER, E. F. and YAKOWITZ, S. (1979). Contributions to the theory of nonparametric regression, with application to system identification. *Ann. Statist.* **7** 139-145.
- STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595-620.
- STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040-1053.
- STUETZLE, W. and MITTAL, Y. (1979). Some comments on the asymptotic behavior of robust smoothers in *Smoothing Techniques for Curve Estimation* (ed. T. Gasser and M. Rosenblatt). *Springer Lecture Notes* **757**, Heidelberg.
- WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theor. Probab. Appl.* **5** 302-305.

DEPARTMENT OF APPLIED MATHEMATICS
UNIVERSITY OF FRANKFURT
D-6000 FRANKFURT/MAIN
WEST GERMANY

SONDERFORSCHUNGSBEREICH 123
IM NEUENHEIMER FELD 293
UNIVERSITY OF HEIDELBERG
D-6900 HEIDELBERG
WEST GERMANY

How to determine the bandwidth of
some nonlinear smoothers in practice *)

Wolfgang Härdle
Fachbereich Mathematik
Johann-Wolfgang Goethe Universität
D - 6000 Frankfurt/M.

Abstract. A nonlinear smoothing procedure which estimates a regression curve is proposed. A kernel operates on data which are first transformed in the way which is familiar in the theory of M-estimators. The bandwidth of the kernel is chosen by a "crossvalidatory" device and asymptotic optimality properties are proven. The proposed method is compared with AIC and FPE and shown to be asymptotically equivalent. An application to Raman-Spectra and a Monte Carlo study show how well our method works in practice.

1. Introduction

Let us assume that we observed a triangular sequence of datapoints

$$(1.1) \quad Y_t^{(T)} = \mu_t^{(T)} + Z_t^{(T)}, \quad t=1,2,\dots,T$$

with expectation

$$E Y_t^{(T)} = \mu_t^{(T)} = m(t/T), \quad t=1,2,\dots,T$$

*) Research partially financed by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 123, "Stochastische Mathematische Modelle".

and independent identically distributed errors $\{z_t^{(T)}\}_{t=1}^T$ with variance σ^2 . The unknown function $m \in C^2[0,1]$, the regression curve, is to be estimated from the observations $\{Y_t^{(T)}\}_{t=1}^T$. In this paper we propose a nonlinear smoothing procedure. We choose a zero of the function

$$\theta \mapsto \sum_{s=1}^T \alpha_s^{(T)}(x) \psi(Y_s^{(T)} - \theta)$$

and call it the M-smoother $S^{(T)}(x)$ derived from ψ and the weights $\alpha_s^{(T)}$. We assume the weights to be given by a kernel function K as follows

$$(1.2) \quad \alpha_s^{(T)}(x) = T^{-1} h^{-1}(T) K((x - s/T)/h(T)).$$

ψ is a given monotone and bounded function, $\psi \in C^2$, with $\psi(0)=0$, $E\psi(Z_t^{(T)}) = 0$. The parameter $h = h(T)$ in the weights (1.2) is called bandwidth. Interpreting $\{\alpha_s^{(T)}(x)\}_{s=1}^T$ as a window (Brillinger, 1975, chapter 3.3) the bandwidth h regulates the size (or span) of the window. In practice, one must select a particular size of the window. It seems desirable to use a bandwidth which makes the averaged square error (ASE) small

$$(1.3) \quad ASE(h;T) = T^{-1} \sum_{t=1}^T [S^{(T)}(t/T) - \mu_t^{(T)}]^2.$$

Denote by $h_A = h_A(T)$ the bandwidth which minimizes ASE. The ASE is a discrete approximation to the mean integrated square error (MISE)

$$MISE(h;T) = E \int_0^1 [S^{(T)}(s) - m(s)]^2 ds$$

of the estimated regression curve $S^{(T)}(\cdot)$. Since the regression curve $m(\cdot)$ is unknown we cannot determine h_A from the data.

We discuss a data-driven procedure for approximating h_A which is based on cross-validation in the sense of Stone (1974). In our case the cross-validatory choice of $h(T)$ is that value $h_C = h_C(T)$ which minimizes

$$(1.4) \quad CRVD(h;T) = T^{-1} \sum_{s=1}^T [S_{-s}^{(T)}(t/T) - Y_t^{(T)}]^2$$

where $S_{-s}^{(T)}(t/T)$ denotes the M-smoother computed from the subsample $\{Y_s^{(T)}\}_{s \neq t}$, i.e. $S_{-s}^{(T)}(t/T)$ is a suitable zero of

(1984) Härdle, W. How to determine the bandwidth of nonlinear smoothers in practice?

$$\theta \mapsto \sum_{s \neq t} \alpha_s^{(T)} \{t/T\} \psi(Y_s^{(T)} - \theta).$$

This is not exactly Stone's (1974) "leave-one-out" statistic, which would be obtained if we would use the weights $(T/T-1)\alpha_s^{(T)}$ rather than $\alpha_s^{(T)}$. For technical convenience we prefer our modified definition of $S_{-}^{(T)}$.

We show that

$$(1.5) \quad \text{CRVD}(h;T) = T^{-1} \sum_{t=1}^T [Z_t^{(T)}]^2 + \text{ASE}(h;T) + R(h;T)$$

where $R(h;T)$ is a remainder term which tends to zero "uniformly in h " (in a sense to be specified later). Remark that the first term on the RHS of (1.5) is independent of h and therefore the task to minimize $\text{CRVD}(h;T)$ is similar to the task to minimize $\text{ASE}(h;T)$ over h . We shall in fact prove

$$h_A/h_C \xrightarrow{p} 1 \quad \text{as } T \rightarrow \infty.$$

From this asymptotic behavior we could expect that $h_C(T)$, the cross-validatory choice of $h(T)$, is a reasonable selection for the bandwidth in practical situations.

A small Monte Carlo study and an application of M -smoothers to Raman-Spectra shows how the method works in practice. We also consider the relationship of CRVD to other devices such as Akaike's (1970, 1974) AIC or FPE and show that they are equivalent to CRVD.

Cross-validation as a method for choosing the degree of smoothing has been proposed by several authors in slightly different situations. Wahba and Wold (1975) discuss spline nonparametric regression; Chow, Geman and Wu (1983) studied kernel density estimators with a bandwidth selected by cross-validation; Wong (1983) showed consistency of the Nadaraya-Watson estimator for regression curves (fixed, equispaced design) with a cross-validatory choice of the bandwidth. Our device is competing with running medians, considered by Tukey (1977), Velleman and Hoaglin (1981). This estimation device admits a similar presentation.

One has to choose

$$\psi(u) = I(u > 0) - 1/2, K(u) = I(|u| \leq 1/2).$$

Our theory does not apply, since these functions do not satisfy our regularity conditions.

The M-smoothers which we investigate here were proposed in a time series setting by Velleman (1980). There are also relations to the work of Mallows (1980) who considered some nonlinear smoothers in the frequency domain but left open the question how the span of such nonlinear smoothers ought to be chosen in practice.

The feature which distinguishes our model from those treated by the authors mentioned above is the possibility of sampling a curve finer and finer. We find this feature in Raman spectroscopy, a field of anorganic chemistry; here indeed the spacings between two successive wavenumbers may be decreased (Bussian and Härdle, 1984). It seems also that our methods can be used in geophysics, in order to identify so-called "nugget-effects" (Cressie, 1983).

For notational convenience we will suppress the index T where it seems unnecessary for the understanding. In particular we shall write $\alpha_s(t)$ instead of $\alpha_s^{(T)}(t/T)$. Similarly S_{t-} instead of $S_{-}^{(T)}(t/T)$, S_t instead of $S^{(T)}(t/T)$ and Y_t instead of $Y_t^{(T)}$.

2. The bandwidth selection problem

In this section we will show that approximation (1.5) holds under the following assumptions on K , m and $h = h(T)$:

- (2.1) The kernel K is twice differentiable, symmetric, integrates to one and vanishes outside $[-1, 1]$;
- (2.2) the sequence of bandwidth $h = h(T)$ tends to zero such that $Th(T) \rightarrow \infty$, as $T \rightarrow \infty$;
- (2.3) the regression curve $m: [0, 1] \rightarrow \mathbb{R}$ is twice continuously differentiable with

$$\int_0^1 [m''(x)]^2 dx > 0$$

and

$$m^{(p)}(0) = m^{(p)}(1), \quad p=0,1,2.$$

Assumption (2.1) is fulfilled by many kernels K which have been proposed in the literature. A good example is the Bartlett-Priestley window (Priestley, 1981, page 569; Epanechnikov, 1969). The assumption (2.3) is introduced for technical convenience. It allows us to treat the problem without modification at the boundary points. In a practical situation where we suspect that (2.3) is not fulfilled we would use a weighted version of $ASE(h;T)$. This is suggested by work of Gasser and Müller (1979) who showed that the rate of convergence of ASE is different at the boundary points.

A linear approximation $\{\tilde{S}_t\}$ to $\{S_t\}$ will be defined in order to simplify technical details. For $\{\tilde{S}_t\}$ an asymptotic relation similar to (1.5) holds. It is then seen that the problem of approximating $h_A(T)$ can be solved via a cross-validation device based on the linear approximation $\{\tilde{S}_t\}$.

S_t is a zero of the equation $\sum_s \alpha_s(t) \psi(Y_s - \theta) = 0$.

Define

$$(2.4) \quad \tilde{S}_t = \sum_{s=1}^T \alpha_s(t) \tilde{Y}_s,$$

where $\tilde{Y}_s = \mu_s + \psi(Z_s)/q$, $q = E \psi'(Z)$. Note that \tilde{S}_t can be interpreted as a classical kernel regression estimate which linearly operates on the non-observable pseudo-data $\{\tilde{Y}_s\}_{s=1}^T$, while the M -smoother $\{S_t\}$ operates nonlinearly on the original data $\{Y_s\}_{s=1}^T$.

In analogy to (1.3) and (1.4) define now the following quantities for $\{\tilde{S}_t\}$.

$$(2.5) \quad \begin{aligned} \widetilde{ASE}(h;T) &= T^{-1} \sum_{s=1}^T [\tilde{S}_t - \mu_t]^2 \\ \widetilde{CRVD}(h;T) &= T^{-1} \sum_{s=1}^T [\tilde{S}_t - \tilde{Y}_t]^2 \end{aligned}$$

where $\tilde{S}_{t-} = \sum_{s \neq t} \alpha_s(t) \tilde{Y}_s$. Define also

$$\widetilde{MASE}(h;T) = E \{ \widetilde{ASE}(h;T) \}.$$

By computations very similar to Parzen (1962) it is seen that

$$(2.6) \quad \widetilde{\text{MASE}}(h;T) = (Th)^{-1} \int K^2(u) du \text{E}\psi^2(Z)/q^2 \\ + 1/4 h^4 \int_0^1 [m''(x)]^2 dx \int u^2 K(u) du + o(T^{-1}h^{-1}+h^4).$$

Neglecting the third summand on the RHS in the equation above, we see that $h(T) = \eta T^{-1/5}$, $\eta > 0$ balances the trade-off between the variance - and the (bias)²- part. The value η minimizing

$$(2.7) \quad \widetilde{M}(\eta) = \lim_{T \rightarrow \infty} T^{4/5} \widetilde{\text{MASE}}(\eta T^{-1/5}; T)$$

is obviously

$$c_1 = \left\{ \int_{-1}^1 K^2(u) du \text{E}\psi^2(Z)/q^2 \int_0^1 [m''(x)]^2 dx \int_{-1}^1 u^2 K(u) du \right\}^{1/5}.$$

Fix now two constants $a < c_1 < b$ and define $\underline{h} = aT^{-1/5}$, $\bar{h} = bT^{-1/5}$. It will be seen in Theorem 2.1, that the remainder terms R_i , $i=1,2$ in the following equation vanish uniformly over $h \in [\underline{h}, \bar{h}]$,

$$(2.8) \quad \text{ASE}(h;T) = \widetilde{\text{ASE}}(h;T) + R_1(h;T) \\ = \widetilde{\text{MASE}}(h;T) + R_2(h;T).$$

To be precise, we show that for all $\varepsilon > 0$

$$(2.9) \quad P\left\{ \sup_{\underline{h} \leq h \leq \bar{h}} T^{4/5} |R_i(h;T)| > \varepsilon \right\} \rightarrow 0, \quad i=1,2, \text{ as } T \rightarrow \infty.$$

Therefore the problem of finding $h_A \in \arg \min_{h \in [\underline{h}, \bar{h}]} \text{ASE}(h;T)$

reduces to selecting a bandwidth between $aT^{-1/5}$ and $bT^{-1/5}$ which minimizes

$$\widetilde{\text{MASE}}(h;T) + R_2(h;T).$$

The first approximation in (2.8) is a consequence of the following lemma.

Lemma 2.1

Consider a sequence $Y_t^{(T)}$ with the properties specified in (1.1) and assume that (2.1) - (2.3) holds, then for all $\varepsilon > 0$, as $T \rightarrow \infty$

$$P \left\{ \sup_{\underline{h} \leq h \leq \bar{h}} T^{4/5} \left| T^{-1} \sum_{t=1}^T (S_t - \tilde{S}_t)^2 \right| > \varepsilon \right\} \rightarrow 0.$$

Proof

Consider the following two functions $\phi : \mathbb{R}^T \rightarrow \mathbb{R}^T$,

$$\psi : \mathbb{R}^T \rightarrow \mathbb{R}^T$$

$$\phi = (\phi_1, \dots, \phi_T) ; \quad \psi = (\psi_1, \dots, \psi_T)$$

where

$$\phi_t(\underline{\xi}) = -q^{-1} \sum_{s=1}^T \alpha_s(t) \psi(Y_s - \xi_t),$$

$$\psi_t(\underline{\xi}) = \xi_t - \sum_{s=1}^T \alpha_s(t) \psi(Z_s) / q - \sum_{s=1}^T \alpha_s(t) \mu_s,$$

$$\underline{\xi} = (\xi_1, \dots, \xi_T).$$

By definition of $\underline{S} = (S_1, \dots, S_T)$, $\tilde{\underline{S}} = (\tilde{S}_1, \dots, \tilde{S}_T)$ we have

$$\phi_t(\underline{S}) = 0, \quad t=1, 2, \dots, T$$

$$\psi_t(\tilde{\underline{S}}) = 0.$$

Applying Taylor's theorem to ϕ_t yields

$$\begin{aligned} \phi_t(\underline{\xi}) &= -q^{-1} \sum_{s=1}^T \alpha_s(t) \psi(Z_s) - q^{-1} \sum_{s=1}^T \alpha_s(t) \psi'(Z_s) [\mu_s - \xi_t] \\ &\quad - 1/2 q^{-1} \sum_{s=1}^T \alpha_s(t) \psi''(Z_s + a_{s,t}) [\mu_s - \xi_t]^2, \end{aligned}$$

where $a_{s,t}$ is between 0 and $\mu_s - \xi_t$.

The difference between ϕ_t and ψ_t is then

$$\begin{aligned} \phi_t(\underline{\xi}) - \psi_t(\underline{\xi}) &= -q^{-1} \sum_{s=1}^T \alpha_s(t) \mu_s [\psi'(Z_s) - q] \\ &\quad + q^{-1} \sum_{s=1}^T \alpha_s(t) \xi_t [\psi'(Z_s) - q] \\ (2.10) \quad &\quad - 1/2 q^{-1} \sum_{s=1}^T \alpha_s(t) \psi''(Z_s + a_{s,t}) [\mu_s - \xi_t]^2 \\ &= R_{1,t} + R_{2,t} + R_{3,t}. \end{aligned}$$

We investigate now the rates at which these $R_{i,t}$, $i=1, 2, 3$ tend uniformly (in the sense of (2.9)) to zero as $T \rightarrow \infty$.

1) Define $V_s = (\psi'(Z_s) - q)/q$ and note that $\{V_s\}_{s=1}^T$ are i.i.d. rv's with mean zero. Summation by parts yields

$$R_{1,t} = \sum_{s=1}^T \alpha_s(t) \mu_s V_s = \sum_{s=1}^T \Delta w_{s,t} \left\{ T^{-1} \sum_{r=1}^s \mu_r V_r \right\}$$

with $\Delta w_{s,t} = h^{-1} \{K((t-1)/(hT)) - K((t-s-1)/(hT))\}$. Assumption (2.1) implies

$$\sum_{s=1}^T |\Delta w_{s,t}| \leq C_1 T^{-1} h^{-2}$$

with a constant C_1 depending on K , and therefore

$$|R_{1,t}| \leq C_1 T^{-1} h^{-2} \left| T^{-1} \sup_{1 \leq s \leq T} \sum_{r=1}^s \mu_r V_r \right|.$$

By Kolmogorov's inequality we have with a constant C_2 , bounding the variances of $\mu_s V_s$,

$$\begin{aligned} P \left\{ \sup_{1 \leq s \leq T} \left| T^{-1} \sum_{r=1}^s \mu_r V_r \right| \geq \varepsilon \right\} &\leq \varepsilon^{-2} \text{var} \left\{ T^{-1} \sum_{s=1}^T \mu_s V_s \right\} \\ &\leq C_2 \varepsilon^{-2} / T, \end{aligned}$$

which shows that

$$\sup_{1 \leq t \leq T} \sup_{\underline{h} \leq h \leq \bar{h}} |R_{1,t}| = o_p(T^{-2/5}).$$

2) The term $R_{2,t}$ is estimated similarly.

3) The third term $R_{3,t}$ splits up into the following three summands.

$$\begin{aligned} R_{3,t} &= \sum_{s=1}^T \alpha_s(t) \psi''(Z_s + a_{s,t}) [\mu_s - \mu_t]^2 \\ &\quad + 2 \sum_{s=1}^T \alpha_s(t) \psi''(Z_s + a_{s,t}) [\mu_s - \mu_t] [\mu_t - \xi_t] \\ &\quad + \sum_{s=1}^T \alpha_s(t) \psi''(Z_s + a_{s,t}) [\mu_t - \xi_t]^2 \\ &= U_{1,t} + U_{2,t} + U_{3,t} \end{aligned}$$

If $C_3 \geq 2b$, then as $T \rightarrow \infty$

$$P \left\{ \sup_{\underline{h} \leq h \leq \bar{h}} T^{-1} \sum_{t=1}^T (\tilde{S}_t - \mu_t)^2 < C_3 T^{-4/5} \right\} \rightarrow 1$$

(Marron and Härdle, 1983).

(1984) Härdle, W. How to determine the bandwidth of nonlinear smoothers in practice?

Define the set

$$\mathcal{F}_T = \{ \underline{\xi} \in \mathbb{R}^T : T^{-1} \sum_{t=1}^T (\xi_t - \mu_t)^2 \leq C_3 T^{-4/5} \}.$$

Then, if $\underline{\xi} \in \mathcal{F}_T$ it is easily seen that with a constant C_4 bounding K and ψ''

$$U_{3,t} \leq C_4 T^{-3/5}.$$

The Cauchy-schwarz inequality shows that there exist constants C_5, C_6 with:

$$U_{1,t} \leq C_5 T^{-3/5}$$

and

$$U_{2,t} \leq C_6 T^{-3/5}.$$

Putting these statements together we finally have that for $\underline{\xi} \in \mathcal{F}_T$

$$(2.11) \quad \sup_{\underline{h} \leq \underline{h} \leq \bar{h}} T^{-1} \sum_{t=1}^T (\phi_t(\underline{\xi}) - \psi_t(\underline{\xi}))^2 = o_p(T^{-4/5}).$$

Now the triangle inequality yields

$$\sup_{\underline{h} \leq \underline{h} \leq \bar{h}} T^{-1} \sum_{t=1}^T (\phi_t(\underline{\xi}) - (\xi_t - \mu_t))^2 = o_p(T^{-4/5}).$$

Therefore the function $\underline{\eta} \mapsto \underline{\eta} - \phi(\underline{\eta} + \underline{\mu})$ maps the compact, convex set $\mathcal{F}_T - \underline{\mu}$, $\underline{\mu} = (\mu_1, \dots, \mu_T)$ into itself and by a suitable fixed-point theorem there exists a fixed point $\hat{\underline{\eta}}$ in $\mathcal{F}_T - \underline{\mu}$.

Setting $\underline{S} = \hat{\underline{\eta}} + \underline{\mu}$ we see that $\phi(\underline{S}) = 0$.

We furthermore have by (2.11)

$$\begin{aligned} & T^{-1} \sum_{t=1}^T [\phi_t(S_t) - \psi_t(S_t)]^2 \\ &= T^{-1} \sum_{t=1}^T (S_t - \tilde{S}_t)^2 = o_p(T^{-4/5}) \end{aligned}$$

with the " o_p " denoting a rv which uniformly in $h \in [\underline{h}, \bar{h}]$ tends to zero. This proves the lemma.

With this lemma we obtain that the difference between ASE and $\widetilde{\text{ASE}}$ is of smaller order than $T^{-4/5}$ uniformly over $h \in [\underline{h}, \bar{h}]$.

This result, together with the second equality in (2.8), yields that $h_A T^{1/5} \xrightarrow{P} c_1$ as $T \rightarrow \infty$.

Theorem 2.1

Consider the sequence $Y_t^{(T)}$, as defined in (1.1), and assume that (2.1) - (2.3) hold. Then for all $\epsilon > 0$, as $T \rightarrow \infty$,

$$(2.12) \quad P \left\{ \sup_{h \in [\underline{h}, \bar{h}]} T^{4/5} |ASE(h;T) - \widetilde{ASE}(h;T)| \geq \epsilon \right\} \rightarrow 0$$

and $h_A \in \arg \min_{h \in [\underline{h}, \bar{h}]} ASE(h;T)$ satisfies $T^{1/5} h_A(T) \xrightarrow{P} c_1$, with

$$(2.13) \quad c_1 = \left\{ \int_{-1}^1 K^2(u) du E \psi^2(Z) / (q^2 \int_0^1 [m''(x)]^2 dx \int_{-1}^1 u^2 K(u) du) \right\}^{1/5} ..$$

Proof

Statement (2.12) follows by lemma 2.1 and an application of the Cauchy-Schwarz-inequality. Note that the remainder term in (2.6) is tending to zero uniformly in $h \in [\underline{h}, \bar{h}]$. Therefore $\widetilde{M}(\cdot)$, as defined in (2.7) reads as

$$\widetilde{M}(\eta) = \eta^{-1} \int K^2(u) du E_F \psi^2(Z) / q^2 + 1/4 \eta^4 \int_0^1 [m''(x)]^2 dx$$

which is a continuous and convex function for $\eta \in [a, b]$ and has its unique minimum at $c_1 = \arg \min_{h \in [\underline{h}, \bar{h}]} \widetilde{M}(hT^{1/5})$. Now

(2.12) and Theorem 1 of Marron and Härdle (1983) yield that, as $T \rightarrow \infty$

$$\begin{aligned} & \sup_{\eta \in [a, b]} |T^{4/5} ASE(\eta T^{-1/5}; T) - \widetilde{M}(\eta)| \\ & \leq \sup_{\eta \in [a, b]} |T^{4/5} \{ASE(\eta T^{-1/5}; T) - \widetilde{ASE}(\eta T^{-1/5}; T)\}| \\ & + \sup_{\eta \in [a, b]} |T^{4/5} \{\widetilde{ASE}(\eta T^{-1/5}; T) - \widetilde{MASE}(\eta T^{-1/5}; T)\}| \\ & + \sup_{\eta \in [a, b]} |T^{4/5} \widetilde{MASE}(\eta T^{-1/5}; T) - \widetilde{M}(\eta)| \xrightarrow{P} 0. \end{aligned}$$

The following arguments are as in Rice (1983). For any $\delta > 0$ define

$$D(\delta) = \inf_{|\eta - c_1| > \delta} (\widetilde{M}(\eta) - \widetilde{M}(c_1)).$$

(1984) Härdle, W. How to determine the bandwidth of nonlinear smoothers in practice?

Then

$$\begin{aligned}
 & P\{|h_A T^{1/5} - c_1| > \delta\} \\
 & \leq P\{\tilde{M}(h_A T^{1/5}) - \tilde{M}(c_1) > D(\delta)\} \\
 & \leq P\{\tilde{M}(h_A T^{1/5}) - T^{4/5} \text{ASE}(h_A; T) + T^{4/5} \text{ASE}(c_1 T^{-1/5}; T) \\
 & \quad - \tilde{M}(c_1) > D(\delta)\} \\
 & \leq P\{\tilde{M}(h_A T^{1/5}) - T^{4/5} \text{ASE}(h_A; T) \geq D(\delta)/2\} \\
 & \quad + P\{T^{4/5} \text{ASE}(c_1 T^{-1/5}; T) - \tilde{M}(c_1) \geq D(\delta)/2\} \\
 & \rightarrow 0, \text{ which proves (2.13).}
 \end{aligned}$$

Recall now the definition of S_{t-} and of $\text{CRVD}(h; T)$. The next theorem shows that (1.5) holds. Therefore, for large T , instead of minimizing the (unknown) function $\text{ASE}(\cdot; T)$, we may minimize $\text{CRVD}(\cdot; T)$.

Theorem 2.2

Consider $\{Y_t^{(T)}\}$, as defined in (1.1) and assume that (2.1)-(2.3) holds.

Then, for all $\varepsilon > 0$,

$$\begin{aligned}
 (2.14) \quad & P\{\sup_{h \in [\underline{h}, \bar{h}]} T^{4/5} |\text{CRVD}(h; T) - T^{-1} \sum_{t=1}^T Z_t^{(T)2} - \text{ASE}(h; T)| \geq \varepsilon\} \\
 & \rightarrow 0, \text{ as } T \rightarrow \infty.
 \end{aligned}$$

and $h \in \arg \min_{h \in [\underline{h}, \bar{h}]} \text{CRVD}(h; T)$ satisfies

$$(2.15) \quad h_C T^{1/5} \xrightarrow{P} c_1$$

where c_1 is the same constant as in Theorem 2.1.

Proof

Consider the following decomposition

$$\begin{aligned}
 (2.16) \quad & (\tilde{Y}_t - \tilde{S}_{t-})^2 = (Y_t - S_{t-})^2 + (S_{t-} - \tilde{S}_{t-})^2 + (\psi(Z_t)/q - Z_t)^2 \\
 & \quad + 2(Y_t - S_{t-})(\psi(Z_t)/q - Z_t + S_{t-} - \tilde{S}_{t-}) \\
 & \quad + 2(S_{t-} - \tilde{S}_{t-})(\psi(Z_t)/q - Z_t)
 \end{aligned}$$

where \tilde{S}_{t-} is the "leave-one-out" statistic based on the pseudo-data $\{Y_s\}_{s \neq t}$. From Härdle and Marron (1983), Theorem 1 we have that, uniformly over $h \in [\underline{h}, \bar{h}]$

$$\begin{aligned}
 \widetilde{CRVD}(h;T) &= T^{-1} \sum_{t=1}^T (\tilde{Y}_t - \tilde{S}_{t-})^2 \\
 (2.17) \qquad &= T^{-1} \sum_{t=1}^T (\psi(Z_t)/q)^2 + \widetilde{ASE}(h;T) + o_p(T^{-4/5}).
 \end{aligned}$$

Now by Theorem 2.1 and the Cauchy-Schwarz inequality we have

$$\sup_{h \in [\underline{h}, \bar{h}]} T^{-1} \sum_{t=1}^T (S_t - \tilde{S}_{t-})^2 = o_p(T^{-4/5}).$$

In view of (2.16) it remains therefore to show that the sum

$T^{-1} \sum_{t=1}^T$ {of the following terms}

$$\begin{aligned}
 &(\psi(Z_t)/q - Z_t)^2 + 2(\mu_t - S_{t-})(\psi(Z_t)/q - Z_t + S_{t-} - \tilde{S}_{t-}) \\
 &+ 2 Z_t(\psi(Z_t)/q - Z_t + S_{t-} - \tilde{S}_{t-}) \\
 &+ 2(S_{t-} - \tilde{S}_{t-})(\psi(Z_t)/q - Z_t) - \psi(Z_t)/q
 \end{aligned}$$

equals

$$T^{-1} \sum_{t=1}^T Z_t^2 + o_p(T^{-4/5})$$

uniformly over $h \in [\underline{h}, \bar{h}]$. Observing that some terms cancel each other, we have to show that the " $T^{-1} \sum_{t=1}^T$ " sum of

$$\begin{aligned}
 &(\mu_t - S_{t-})(\psi(Z_t)/q - Z_t) + (\mu_t - S_{t-})(S_{t-} - \tilde{S}_{t-}) \\
 &+ (\psi(Z_t)/q)(S_{t-} - \tilde{S}_{t-}) \\
 &= W_{1,t} + W_{2,t} + W_{3,t} = o_p(T^{-4/5}).
 \end{aligned}$$

By Theorem 2.1 and (2.6) we have that

$$\begin{aligned}
 T^{-1} \sum_{t=1}^T W_{2,t} &\leq (ASE(h;T))^{1/2} (T^{-1} \sum_{t=1}^T (S_{t-} - \tilde{S}_{t-})^2)^{1/2} \\
 &= o_p(T^{-4/5}), \text{ uniformly over } h \in [\underline{h}, \bar{h}].
 \end{aligned}$$

The third term is estimated as in the proof of Lemma 2.1 by setting $\xi_t = S_{t-}$ in (2.10) and observing that S_{t-} and \tilde{S}_{t-} are independent of $\psi(Z_t)/q$. It remains to show that

$$\sup_{h \in [\underline{h}, \bar{h}]} T^{-1/5} \sum_{t=1}^T (\mu_t - s_{t-}) (\psi(Z_t)/q) = o_p(1), \text{ since the analysis}$$

of the term where $\psi(Z_t)/q$ is replaced by Z_t is the same. Adding and subtracting \tilde{S}_{t-} and repeating the argument for $W_{3,t}$ it remains to show that

$$\begin{aligned} \sup_{h \in [\underline{h}, \bar{h}]} T^{-1/5} \sum_{t=1}^T (\mu_t - \sum_{s \neq t} \alpha_s(t) \mu_s - \sum_{s \neq t} \alpha_s(t) \psi(Z_s)/q) \\ \cdot (\psi(Z_t)/q) \\ = o_p(1) \end{aligned}$$

Consider the bias term

$$\sup_{h \in [\underline{h}, \bar{h}]} T^{-1/5} \sum_{t=1}^T (b_T(t) \psi(Z_t)/q)$$

where $b_T(t) = \mu_t - \sum_{s \neq t} \alpha_s(t) \mu_s = O(T^{-4/5})$ in the range $h \in [\underline{h}, \bar{h}]$.

This shows that the bias term is $o_p(1)$. Using now the independence of $\psi(Z_t)/q$ from $\sum_{s \neq t} \alpha_s(t) \psi(Z_s)/q$ it follows by

similar calculations as in the proof of Lemma 2.1 that

$$T^{-1} \sum_{t=1}^T W_{1,t} = o_p(T^{-4/5}), \text{ uniformly over } h \in [\underline{h}, \bar{h}]. \text{ This proves (2.14).}$$

We show now (2.15). Recall the definition of $\tilde{M}(\eta)$ and $D(\delta)$

then with (2.14) and $\hat{\sigma}_T^2 = T^{-1} \sum_{t=1}^T Z_t^{(T)2}$ we have,

$$\begin{aligned} P\{|T^{1/5} h_c - c_1| > \delta\} &\leq P\{\tilde{M}(T^{1/5} h_c) - \tilde{M}(c_1) > D(\delta)\} \\ &\leq P\{\tilde{M}(T^{1/5} h_c) - CRVD(h_c; T) - \hat{\sigma}_T^2 + CRVD(T^{-1/5} h_c; T) + \hat{\sigma}_T^2 \\ &\quad - \tilde{M}(c_1) > D(\delta)\} \\ &\leq P\{\tilde{M}(T^{1/5} h_c) - ASE(h_c; T) \geq D(\delta)/4\} \\ &\quad + P\{ASE(T^{-1/5} c_1; T) - \tilde{M}(c_1) \geq D(\delta)/4\} \\ &\rightarrow 0 \quad \text{by Theorem 2.1.} \end{aligned}$$

3. Relations to other devices for selecting a bandwidth

In section 2 we studied the selection of the bandwidth $h_c \in \arg \min_{\underline{h} \leq h \leq \bar{h}} \text{CRVD}(h; T)$ on the basis of a modified form of Stone's (1974) crossvalidation function. This was mainly done for historical reasons, since Wahba and Wold (1975) introduced the crossvalidation method as a device to pick up "asymptotically correct" sequences of bandwidth in the setting of regression function estimation. Stone (1977) showed an asymptotic equivalence of the crossvalidation method and Akaike's information criterion (AIC) in the context of model selection. It is therefore of interest to study the equivalence of other devices, such as AIC, FPE, to cross-validation in our context.

Note that in the proof of Theorem 2.2 we have essentially shown that

$$\text{CRVD} = \widetilde{\text{CRVD}} - T^{-1} \sum_{t=1}^T \psi^2(z_t)/q^2 + T^{-1} \sum_{t=1}^T z_t^2 + o_p(T^{-4/5}).$$

Since the two middle terms on the RHS do not depend on h and the last term vanishes uniformly in $h \in [\underline{h}, \bar{h}]$, we conclude with the techniques developed in section 2, that the minima of CRVD approximate asymptotically the minima of $\widetilde{\text{CRVD}}$. We therefore consider only $\widetilde{\text{CRVD}}$ in the following.

Let us rewrite $\widetilde{\text{CRVD}}$:

$$\begin{aligned} \widetilde{\text{CRVD}}(h; T) &= T^{-1} \sum_{t=1}^T (\tilde{Y}_t (1 + T^{-1} h^{-1} K(o)) - \tilde{S}_t)^2 \\ (3.1) \quad &= T^{-1} \sum_{t=1}^T (\tilde{Y}_t - \tilde{S}_t)^2 + T^{-1} \sum_{t=1}^T (T^{-1} h^{-1} K(o) \tilde{Y}_t)^2 \\ &\quad + 2T^{-1} \sum_{t=1}^T (\tilde{Y}_t - \tilde{S}_t) (T^{-1} h^{-1} K(o) \tilde{Y}_t). \end{aligned}$$

It is easy to see that $\sup_{h \in [\underline{h}, \bar{h}]} |T^{-1} \sum_{t=1}^T (T^{-1} h^{-1} K(o) \tilde{Y}_t)^2| = o_p(T^{-4/5})$.

The third term is equal to

$$2T^{-1} h^{-1} K(o) E_F \psi^2(z)/q^2 + o_p(T^{-4/5})$$

uniformly over $h \in [\underline{h}, \bar{h}]$.

Define the residual sum of squares

$$\widetilde{\text{RSS}}(h;T) = T^{-1} \sum_{t=1}^T (\tilde{Y}_t - \tilde{S}_t)^2.$$

Then as we have shown above

$$\widetilde{\text{CRVD}} = \widetilde{\text{RSS}} + 2T^{-1}h^{-1}K(o)V(\psi,F) + o_p(T^{-4/5})$$

where $V(\psi,F) = E_P \psi^2(Z)/q$.

Define the leading term

$$(3.2) \quad C^*(h;T) = \widetilde{\text{RSS}}(h;T) + 2T^{-1}h^{-1}K(o)V(\psi,F) - V(\psi,F).$$

We will see in Theorem 3.1 that the minima of $C^*(\cdot;T)$ approximate asymptotically the minima of the following functions.

$$(3.3) \quad \exp(\text{AIC}(h;T)) = \widetilde{\text{RSS}}(h;T) \exp(2T^{-1}h^{-1}K(o))$$

$$\text{AIC}(h;T) = \log(\widetilde{\text{RSS}}(h;T)) + 2T^{-1}h^{-1}K(o)$$

(Akaike, 1974),

$$(3.4) \quad \text{FPE}(h;T) = (1 - T^{-1}h^{-1}) / (1 - T^{-1}h^{-1}) \widetilde{\text{RSS}}(h;T)$$

(Akaike, 1970),

$$(3.5) \quad \text{SHI}(h;T) = \widetilde{\text{RSS}}(h;T) (1 + 2T^{-1}h^{-1}K(o))$$

(Shibata, 1981).

This list may be extended to GXV (generalized cross-validation, Craven and Wahba, 1979) or $\text{FPE}(\alpha)$, a modified FPE criterion from Bhansali and Downham (1977).

Note that all the devices listed from (3.2) to (3.5) carry the same structure. They contain a term involving $\widetilde{\text{RSS}}$ which is decreasing as $h \downarrow 0$ and a penalty term getting bigger if h is too small. The next theorem states that a small random or nonrandom disturbance $\delta(h;T)$ of $C^*(h;T)$ does not affect the asymptotic optimality of h .

Theorem 3.1

Suppose that for all $\varepsilon > 0$

$$P\left\{ \sup_{h \in [\underline{h}, \bar{h}]} T^{4/5} |\delta(h; T)| > \varepsilon \right\} \rightarrow 0, \text{ as } T \rightarrow \infty.$$

Then a sequence of bandwidth $h_{C, \delta}(T)$ chosen so as to minimize

$$C_{\delta}^*(.; T) = C^*(.; T) + \delta(.; T)$$

approximates asymptotically $h_C \in \arg \min_{h \in [\underline{h}, \bar{h}]} \text{CRVD}(h; T)$, i.e.

$$h_{C, \delta} - h_C \xrightarrow{P} 0, \text{ as } T \rightarrow \infty.$$

The proof of this theorem follows closely the arguments that were used in the proof of Theorem 2.2.

Shibatas criterion function (3.5) may be written as

$$\begin{aligned} \text{SHI}(h; T) &= (C^*(h; T) + V(\psi, F) - 2T^{-1}h^{-1}K(o)V(\psi, F)) \\ &\quad (1 + 2T^{-1}h^{-1}K(o)) \\ &= C^*(h; T) + \delta(h; T) + V(\psi, F) \end{aligned}$$

where $\delta(h; T) = o_p(T^{-4/5})$ uniformly over $h \in [\underline{h}, \bar{h}]$.

The other functions may be expanded in Taylor-series to see that they are asymptotically equivalent to $\text{SHI}(h; T)$.

4. An example and a Monte Carlo study

We report here the results of an application and of a small Monte Carlo simulation. M -smoothers of the function $m(s) = \sin(2\pi s)$ were computed from a sample of $T=100$ equispaced data points t/T , $1 \leq t \leq T$. The residuals $\{z_t\}$ were generated according to the pdf

$$(4.1) \quad g(z) = 9\phi(10z) + 1/9\phi(z/9)$$

where ϕ denotes the pdf of a standard normal distribution. By direct computation one sees $\sigma^2 = 8.19$. The kernel we implemented was the so-called Bartlett-Priestley window (Epanechnikov, 1969)

$$(4.1) \quad \begin{aligned} K(u) &= .75(1-u^2) & |u| \leq 1 \\ &= 0 & |u| > 1 \end{aligned}$$

(1984) Härdle, W. How to determine the bandwidth of nonlinear smoothers in practice?

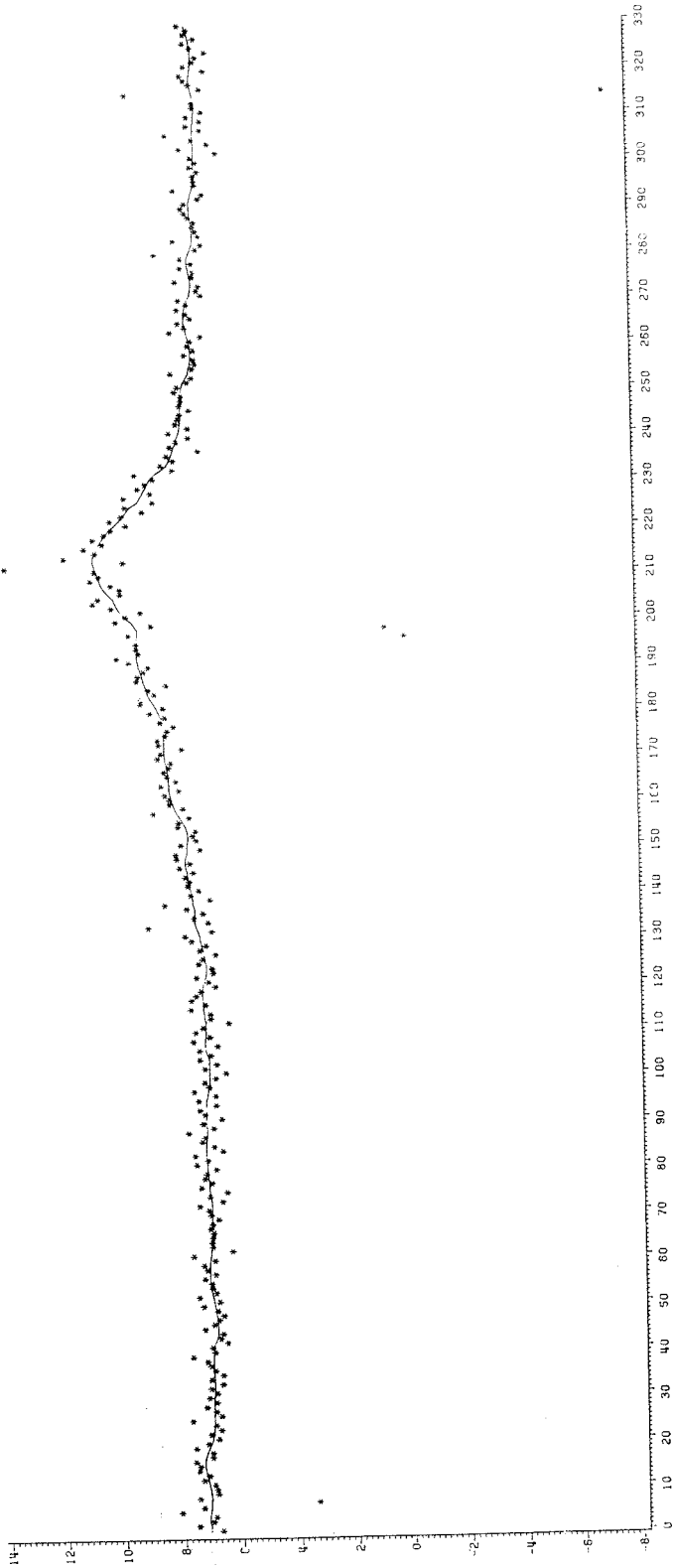
The IMSL routine GGNPM was used to generate the Gaussian pseudo random numbers. For each of the 160 Monte Carlo runs, the functions $CRVD(h)$ and $ASE(h)$ for $h = i/200$, $i = 3, 5, \dots, 15$ were computed. We used Huber's ψ -function

$$(4.3) \quad \psi(u) = \max(-\mathcal{K}, \min(u, \mathcal{K})) , \mathcal{K} > 0$$

for $\mathcal{K} = 1.2, 1.5, 3$. The mean and the standard deviation of $CRVD$ and ASE for different bandwidth h and tuning parameter \mathcal{K} , together with the correlation between ASE and $CRVD$, are shown in Table 1. The numbers shown there are consistent with the theory: the averaged $CRVD$ and ASE curves have both their minimum at .065 for $\mathcal{K} = 1.2, 1.5, 3$.

An application of M-smoothing to Raman spectroscopic data was also carried out. For various reasons spiky outliers may corrupt the recorded Raman spectrum. Intermittent high frequency signals, bubbles in the sample, furthermore shock waves within the optical instrumentation may introduce absurd spikes (Bussian and Härdle, 1984). In Figure 1 a typical data sequence, $T = 330$, together with the smoothed series $\{S_t\}_{t=1}^T$ is shown. Huber's ψ -curve (4.3) was used and S_t was computed by the Newton-Raphson algorithm. In Figure 2 a batch of $CRVD$ curves for different levels of \mathcal{K} is shown. To simplify the interpretation, on the horizontal axis the scale $2hT$ is used rather than h itself. The solid line in Figure 2 corresponds to $\mathcal{K} = .2$ and the finest dotted line belongs to $\mathcal{K} = .4$; the three other graphs were computed for $\mathcal{K} = .25, .3, .35$ respectively. The five curves have their minimum all in the range between 6 and 8. Selecting $2hT = 7$ and $\mathcal{K} = .25$ gives the smooth curve of Figure 1. There the M-smoother $\{S_t\}$, overlaid with the original data $\{Y_t\}$, is shown. Obviously $\{S_t\}$ is not affected by the spurious observations at $t \approx 200$ and $t \approx 310$. We tested our assumption on the noise sequence $\{Z_t\}$ by means of Bartlett's test (Priestley, 1981). The test did not reject the white noise hypothesis at a 5% significance level. The programs, written in FORTRAN, can be obtained from the author.

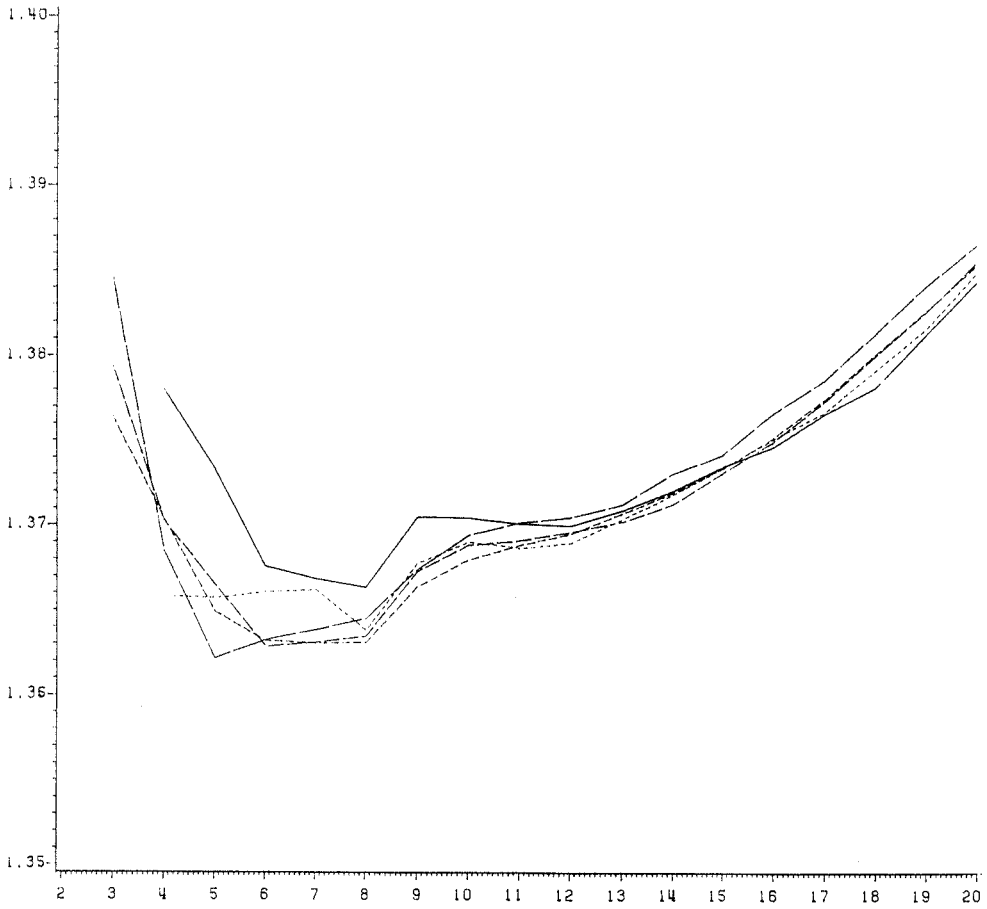
ROBUST SMOOTHED RAMAN DATA



$T = 330, K = .25, \text{span} = 13$

Figure 1

CRVD WITH RAMAN DATA $T = 330$
 RAMAN DATA M0181055



$(\text{span}+1)/2$

$\kappa = .2 \text{ to } .4 \text{ by } .05$

Figure 2

$(200h+1)/2$		2			3			4			5			6			7			8		
		1.2	1.5	3.	1.2	1.5	3.	1.2	1.5	3.	1.2	1.5	3.	1.2	1.5	3.	1.2	1.5	3.	1.2	1.5	3.
\mathcal{X}	CRVD	205	207	186	166	122	23	60	37	11	40	27	9.7	47	21	8.6	18	7.9	32	20	8	
	ASE	198	200	179	159	115	16	53	29	3.7	33	20	2.1	40	13	1.0	20	10	.3	24	12	.4
std	CRVD	172	197	219	169	141	70	117	85	28	78	64	23	88	57	10	68	50	4	77	50	4.5
	ASE	170	196	218	168	141	70	118	85	27	77	62	21	87	57	9	67	49	.1	75	48	.2
corr ASE,CRVD		.997	.997	.996	.994	.991	.735	.852	.8	.637	.771	.698	.585	.74	.661	.534	.603	.645	.492	.604	.673	.449

Table 1

The curves CRVD and ASE averaged over 160 Monte Carlo experiments.

$$\mu_t = \sin(2\pi t/T), \quad 1 \leq t \leq T=100.$$

$$z_t = 9\phi(10x) + 1/90 \phi(x/9).$$

Acknowledgement

I would like to thank my colleagues H. Dinges, J.B. Ferebee, J. Franke, G. Kersting and all other patient souls for many helpful discussions.

References

- Akaike H (1970). Statistical predictor identification. Ann. Inst. Stat. Math. 22, 203 - 217
- Akaike H (1974). A new look at statistical model identification. IEEE Trans. Auto. Cont. 19, 716 - 723
- Bhansali R J and Downham D Y (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion, Biometrika, 64, 547-551
- Brillinger D R (1975). Time Series, Data Analysis and Theory; Holt, Rinehard and Winston, New York
- Bussian B and Härdle W (1984). Robust smoothing applied to white noise and single outlier contaminated Raman spectra. Applied Spectroscopy, 38, 309 - 313
- Chow Y S , Geman S and Wu L D (1983). Consistent cross-validated density estimation. Ann Stat. 11, 25 - 38
- Craven P and Wahba G (1979). Smoothing noisy data with spline functions. Numerische Mathematik, 31, 377 - 403
- Cressie N (1983). Personal communication.
- Epanechnikov V A (1969). Nonparametric estimation of a multivariate probability density. Theory Probab. Appl. 14, 153 - 158
- Gasser T and Müller H G (1979). Kernel estimation of regression functions. in "Smoothing Techniques for Curve Estimation". Springer Lecture Notes 757
- Härdle W and Marron S (1983). Optimal bandwidth selection in nonparametric regression function estimation. Inst. of Stat. Mimeo Series #1530, Chapel Hill, N.C.
- Härdle W and Gasser T (1984). Robust nonparametric function fitting. J. Royal Stat. Soc. B 46, 42 - 51

(1984) Härdle, W. How to determine the bandwidth of nonlinear smoothers in practice?

- Mallows C (1980). Some theory of nonlinear smoothers. *Ann. Stat.* 8, 695 - 715
- Marron S and Härdle W (1983). Random approximations to an error criterion of nonparametric statistics. *Inst. of Stat. Mimeo Series #1538*, Chapel Hill, N.C.
- Parzen E (1962). On the estimation of a probability density and mode. *Ann. Math. Stat.* 33, 1065 - 1076
- Priestley M B (1981). *Spectral analysis and time series*. Academic Press, London
- Rice J (1983). Bandwidth choice for nonparametric kernel regression. Unpublished manuscript
- Shibata R (1981). An optimal selection of regression variables. *Biometrika*, 68, 45 - 54
- Stone M (1974). Crossvalidatory choice and assessment of statistical predictions (with discussion). *J. Royal Stat. Soc. B*, 36, 111 - 147
- Stone M (1977). An asymptotic equivalence of choice of model by crossvalidation and Akaike's criterion. *J. Royal Stat. Soc. B*, 39, 44 - 47
- Tukey (1977). *Exploratory data analysis*. Addison-Wesley, Reading Massachusetts
- Velleman P F (1980). Definition and Comparison of Robust Nonlinear Data Smoothing Algorithms. *J. Amer. Stat. Ass.*, 75, 609 - 615
- Velleman P F, Hoaglin D C (1981). *Applications, Basics, and computing of Exploratory Data Analysis*. Duxbury Press, Boston Massachusetts
- Wahba G and Wold S (1975). A completely automatic French curve: fitting splines functions by cross-validation. *Communications in Statistics* 4, 1 - 17
- Wong W H (1983). On the consistency of cross-validation in kernel nonparametric regression. *Ann. Stat.* 11, 1136 - 1141

Cahiers du C.E.R.O., Volume 28, n^o 1-2, 1986
Coll. Approches non paramétriques et analyse chronologique, Bruxelles, 1985

QUELQUES ASPECTS DE LA PREDICTION NON PARAMETRIQUE :
TRAVAUX DE GERARD COLLOMB (1951-1985) EN
ANALYSE NON PARAMETRIQUE DES SERIES TEMPORELLES (*)

Wolfgang HÄRDLE

1. POURQUOI LA PREDICTION NON PARAMETRIQUE DE SERIES CHRONOLOGIQUES ?

Soit $\{Z_i, i = 0, \pm 1, \dots\}$ un processus stationnaire à valeurs réelles. On suppose qu'une réalisation de longueur N du processus $\{Z_i\}_{i=1}^N$ ont été observées et l'on veut prédire Z_{N+1} . Lorsque des hypothèses sur la structure du processus permettent de le caractériser par un nombre fini de paramètres, le problème de la prédiction de Z_{N+1} revient à estimer convenablement ces paramètres. Une attention a été portée au cas de processus Gaussiens puisque ceux-ci sont entièrement caractérisés par leur moyenne et leurs autocorrélations. Même lorsqu'on pense que le processus observé n'est pas Gaussien, les méthodes basées sur l'estimation des autocorrélations sont utilisées, l'interprétation des paramètres estimés étant considérée comme plus importante que l'efficacité de la méthode utilisée. Ces dernières années, de nombreuses approches ont été développées pour la modélisation de processus non Gaussiens. De nombreux auteurs se sont intéressés à des modèles autorégressif non Gaussiens plus ou moins modifiés (par exemple Lawrence et Lewis (1980), Martin et Yohai (1985), Tong et Lin (1980) ...). Une liste de références, riche mais certainement incomplète, se trouve dans les actes de la conférence "Robust and Nonlinear Time Series Analysis" (1984). En particulier, l'approche robuste de l'analyse des séries temporelles est très prometteuse mais elle ne résout pas le problème du choix initial du modèle paramétrique pour $\{Z_i\}$.

Gérard Collomb envisageait le problème de la prédiction de Z_{N+1} à partir de $\{Z_i\}_{i=1}^N$ par une approche non paramétrique. Cette approche a l'avantage, pour un grand nombre d'observations, de fournir des informations très détaillées, et peut donc permettre de définir un modèle paramétrique raisonnable.

Dans une série de publications, [1], [2], [3], [4], [5], [6], il a obtenu les vitesses de convergence de prédicteurs non paramétriques construits par la méthode du noyau ou par celle des k -points les plus proches. Sa revue bibliographique [7] sur l'estimation non paramétrique de la régression et sur la prédiction récapitule l'ensemble des travaux effectués dans ce domaine et apporte un point de vue général sur les méthodes non paramétriques. Dans les articles [8], [9], [10], [11] il a étudié la convergence du prédicteur à noyau et défini le prédicteogramme.

(*) Je tiens à remercier chaleureusement Philippe Vieu pour l'aide qu'il a apportée en fournissant des détails intéressants sur les travaux de Gérard Collomb.

Ce prédictogramme est très utile en analyse exploratoire des données comme il l'a montré dans [9], un article où sont exposées des réalisations de calculs pour des techniques exploratoires en régression et prédiction. Pour établir des résultats de convergence uniforme (sur des compacts) l'outil probabiliste essentiel est une inégalité du type Bernstein concernant les processus Φ -mélangeants qu'il démontre dans [13]. Cette inégalité est présentée dans le paragraphe 2.

Il s'est aussi intéressé aux estimateurs non paramétriques de la densité de la loi marginale du processus stationnaire $\{Z_i\}$, [19]. Dans des applications particulières, ces estimateurs de densité peuvent indiquer si une hypothèse Gaussienne est justifiée et le cas échéant quel type d'hypothèse non Gaussienne choisir. Ainsi, à partir d'estimateurs non paramétriques, on peut aboutir à des conclusions sur le type de modèle paramétrique approprié pour de futures analyses. Une structure non linéaire peut être détectée ou l'indication qu'une approche Gaussienne classique est justifiée.

La méthode qu'employait le plus souvent Gérard Collomb est la méthode du noyau. Les estimateurs à noyau de la densité ont été introduits par Rosenblatt (1956) et les estimateurs à noyau de la régression par Nadaraya (1964) et Watson (1964). Il est intéressant de noter que pour ce dernier auteur, l'introduction de tels estimateurs était motivée par l'analyse non paramétrique de données météorologiques. L'approche non paramétrique en liaison avec les idées de robustesse a été proposée par Brillinger dans une discussion au sujet de l'article de Stone (1974). Ce propos a été repris plus en détail par Härdle (1984) et Härdle et Gasser (1984). Les M-estimateurs robustes qui y sont étudiés ont été employés pour la prédiction non paramétrique pour des séries temporelles dans un travail commun de Gérard Collomb et moi-même [19]. Ici aussi une étape importante dans les preuves est constituée par l'inégalité, déjà mentionnée, du type Bernstein pour des variables aléatoires Φ -mélangeantes [13].

Comme tous les autres estimateurs non paramétriques, l'estimateur à noyau de Nadaraya-Watson dépend d'un paramètre de lissage. La vitesse de convergence est fonction de ce paramètre de lissage et de la taille de l'échantillon. Elle est d'autant meilleure que la fonction à estimer (généralement la fonction de régression ou la densité) est lisse. Or, dans les applications, ce degré de régularité est inconnu et pourtant le paramètre se doit d'être choisi convenablement. Une mesure possible de la valeur d'un prédicteur est la moyenne des erreurs quadratiques (ASE) ou la moyenne intégrée de ces erreurs quadratiques (MISE).

Comment ces mesures peuvent-elles être optimisées sur une classe de paramètres de lissage ? Dans un travail commun avec moi-même, Gérard Collomb étendit l'idée de validation croisée (Härdle et Marron (1985)) au cas de la prédiction optimale de Z_{N+1} . Ceci constituait le dernier projet sur lequel il travaillait avant sa mort bien trop soudaine pour nous tous.

2. LA METHODE DU NOYAU EN PREDICTION NON PARAMETRIQUE

La fonction d'autorégression $r^* : \mathbb{R}^d \rightarrow \mathbb{R}$ est définie par

$$r^*(u) = E \left\{ Z_{i+1} / (Z_{i-d+1}, \dots, Z_i) = u \right\}, \quad i \geq d.$$

Nous voulons prédire Z_{N+1} à partir des données $\{Z_i\}_{i=1}^N$. Pour une fonction de perte quadratique le meilleur prédicteur est $r^*(Z_{N-d+1}, \dots, Z_N)$. L'estimateur de Nadaraya-Watson de r^* est

(1986) Härdle, W. Quelques aspects de la prédiction non paramétrique: travaux de Gérard Collomb (1951-1985) en analyse non paramétrique des séries temporelles

défini à l'aide d'un noyau K qui est une fonction réelle, bornée, symétrique ($K(x) = K(-x)$), définie sur \mathbb{R}^d et telle que

$$|u|^d K(u) \rightarrow 0 \quad \text{quand} \quad |u| \rightarrow \infty,$$

$$\int K(u) \, du = 1,$$

$$|K(u) - K(v)| \leq C_K |u-v|^\gamma, \quad 0 \leq \gamma < 1.$$

En posant $n=N-d$, cet estimateur est défini par

$$r_n^*(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x - X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right)}, \quad (2.1)$$

où $(h_n)_{n \in \mathbb{N}}$ est une suite de nombres réels strictement positifs de limite nulle, et où

$$X_i = (Z_i, \dots, Z_{i+d-1}) \text{ et } Y_i = Z_{i+d}. \quad (2.2)$$

Ce fut une importante contribution de Gérard Collomb de voir que le problème de la prédiction pouvait être traité dans un cadre plus général en considérant un processus $\{(X_i, Y_i)\}$ à valeurs dans \mathbb{R}^{d+1} satisfaisant certaines conditions de mélange et de particulariser ensuite au cas (2.2). Ce point de vue général englobe aussi le cas où les variables (X_i, Y_i) sont indépendantes qui sera dans la suite appelé "le cas indépendant".

Dans les articles [13] et [16], Gérard Collomb obtient des résultats de convergence uniforme qui conduisent à la propriété importante

$$r_n^*(Z_{n-d+1}, \dots, Z_N) - r(Z_{N-d+1}, \dots, Z_N) \xrightarrow[N \rightarrow +\infty]{P.S.} 0.$$

Une étape fondamentale dans l'établissement des résultats de convergence uniforme sur des compacts est l'application d'inégalités sur les moments. Ceci ne pose pas de problème dans le cas indépendant, mais dans le cas de variables dépendantes, de telles inégalités n'existaient pas et devaient donc être établies. Dans le cas de variables aléatoires ϕ -mélangeantes il a obtenu dans [13] une inégalité du type Bernstein que nous allons énoncer après avoir rappelé la définition de la condition de ϕ -mélange.

Définition

Un processus $\{\epsilon_n, n \in \mathbb{N}\}$ est dit ϕ -mélangeant si pour une suite $\{\rho_n, n \in \mathbb{N}\}$ de réels positifs telle que

$$\rho_n \xrightarrow[n \rightarrow \infty]{} 0,$$

on a pour tout entier $k > 0$,

$$|P(A \cap B) - P(A)P(B)| \leq \rho_k P(A),$$

pour tout entier $n > 0$ et pour tout ensemble A (resp. B) qui soit $\sigma(\epsilon_1, \dots, \epsilon_n)$ (resp. $\sigma(\epsilon_{n+k}, \epsilon_{n+k+1}, \dots)$)-mesurable.

Théorème 1 (Inégalité de Collomb)

Soit $\Delta_i = \Delta_{n_i}$, $i \in \mathbb{N}$ une suite de variables aléatoires Φ -mélangeantes telles que

$$E \Delta_i = 0, |\Delta_i| \leq d, E |\Delta_i| \leq \delta, E \Delta_i^2 \leq D,$$

et supposons que la suite $\{\rho_k, k \in \mathbb{N}\}$ des coefficients de mélange soit indépendante de n .

En posant $\tilde{\rho}_m = \sum_{i=1}^m \rho_i$ on a pour tout $\epsilon > 0$

$$P \left(\left| \sum_{i=1}^n \Delta_i \right| > \epsilon \right) \leq C e^{(-d\epsilon + d^2 n C)},$$

où

$$C = \sigma(D + 4 \delta d \tilde{\rho}_m), c = e^{3\sqrt{C} n \rho_m/m}$$

et où m et α sont respectivement un entier et un réel positif tels que

$$1 \leq m \leq n, \quad \alpha m d \leq 1/4.$$

Cette inégalité joue un rôle essentiel dans les preuves des résultats de convergence uniforme des estimateurs à noyau de la densité marginale des $\{X_i\}$. Elle est aussi très utile pour l'estimation de r^* si l'on applique une technique de troncature comme celle utilisée par les étudiants de Gérard Collomb, Sarda et Vieu (1985).

Dans [13] la convergence, uniforme presque complète sur un compact, de r_n^* vers r^* était établie dans le cas de variables $\{Y_i\}$ uniformément bornées.

3. PREDICTION NON PARAMETRIQUE ROBUSTE

Nous observons que l'estimateur de Nadaraya-Watson $r_n^*(x)$ défini par (2.1) peut être considéré comme un estimateur des moindres carrés dans ce sens qu'il est solution (pour $K \geq 0$) du problème de minimisation en t de la fonction suivante

$$\sum_{i=1}^n K[(x - X_i)/h_n] (Y_i - t)^2.$$

Il est clair que r_n^* doit être fortement sensible aux grandes variations des données puisqu'il est une moyenne locale d'observations de Y . Cette difficulté est extrêmement gênante dans le cas de petits échantillons. Un remède consiste à remplacer la perte quadratique par une fonction de perte qui donne moins de poids aux valeurs extrêmes.

Ainsi nous considérerons un estimateur $r_n(x)$ qui est implicitement défini comme un zéro de la fonction suivante

$$t \rightarrow \sum_{i=1}^n K((x - X_i)/h_n) \psi_x(Y_i - t),$$

où ψ_x est une fonction bornée pour tout x qui satisfait certaines conditions de régularité que nous donnerons plus loin. Plus généralement nous définissons un prédicteur $r(x)$ qui est un zéro de la fonction suivante

$$t \rightarrow E \{ \psi_x(Y_1 - t/X_1 = x) \}.$$

Lorsque le processus $\{Z_n\}_{n \in \mathbb{N}}$ est markovien d'ordre d , la v.d.r.

$$r(Z_{N-d+1}, \dots, Z_N)$$

est le meilleur prédicteur pour la fonction de perte

$$g(\tilde{v}) = \int_{-\infty}^{\tilde{v}} \psi(s) ds \quad \text{où} \quad \psi \equiv \psi_X.$$

Le fait que ψ_X soit bornée garantit une faible sensibilité aux valeurs aberrantes.

L'estimateur de Nadaraya-Watson correspond au cas particulier

$$\psi_X(\bar{x}) \equiv \bar{x}.$$

Dans [19] les vitesses de convergence forte uniforme sont obtenues pour les deux estimateurs r_n et r_n^* . Dans le cas indépendant $r_n(x)$ avait été étudié par Tsybakov (1983), Robinson (1984) et Härdle (1984). Ici à nouveau on fait l'hypothèse de Φ -mélange.

Les conditions supplémentaires suivantes sont nécessaires :

$$h_n \rightarrow 0, \quad nh_n^d \rightarrow \infty, \quad h_n > 0 \quad \forall n \in \mathbb{N};$$

$$\left| \frac{\partial \psi_X(\bar{x})}{\partial (\bar{x})} \right| \leq C\psi.$$

En introduisant une suite croissante d'entiers $(m_n)_{n \in \mathbb{N}}$ satisfaisant

$$\exists A < \infty, \quad n \Phi_{m_n/m_n} \leq A, \quad 1 \leq m_n \leq n, \quad \forall n \in \mathbb{N},$$

on a sous les hypothèses précédentes le résultat suivant.

Théorème 2

Soit C un compact de \mathbb{R}^d et G un voisinage compact de 0 dans \mathbb{R} . Nous supposons que K est positif et que

$$\inf_{t \in G} \inf_{x \in C} E [\psi'_X(Y - r(x) - t) / X = x] f(x) \geq C_0 > 0$$

où f est la densité marginale de X , et que

$$\sup_{t \in G} \sup_{x \in C} \sup_{\bar{x} \in \mathbb{R}^d} \left| \frac{\partial^2 E(\psi_X(Y - r(x) - t) / X = \bar{x}) f(\bar{x})}{\partial^2(\bar{x})} \right| \leq C_1 < \infty.$$

Si la suite $\{h_n\}$ vérifie

$$\theta_n = (m_n \log n / (n h_n^d))^{1/2} \rightarrow 0$$

ainsi que

$$\exists B, \quad 0 < B < +\infty, \quad \theta_n^{-1} h_n^2 \leq B \quad \forall n \in \mathbb{N},$$

alors

$$\theta_n^{-1} \sup_{x \in C} |r_n(x) - r(x)| = o(1) \text{ p.s. .}$$

L'application de ce résultat à la prédiction d'un processus markovien est discutée plus en détail dans [19]. Nous voudrions simplement remarquer que l'on peut choisir $m_n = \lfloor c \log n \rfloor$ dès que le processus est géométriquement ϕ -mélangeant, ce qui amène comme vitesse de convergence $\theta_n = \log n (n h_n^d)^{-1/2}$ dans le cas d'un processus $\{Z_i\}$ qui est markovien et qui satisfait la condition de Doebelin.

Ces résultats appliqués au cas indépendant généralisent ceux de Mack et Silvermann (1982) et Härdle et Lückauš (1984).

4. PREDICTION OPTIMALE POUR L'ERREUR QUADRATIQUE

Le problème qui se pose au praticien optant pour la méthode de Nadaraya-Watson est celui du choix de la largeur de fenêtre h_n . Une façon de sélectionner h consiste à minimiser l'erreur quadratique moyenne intégrée (MISE) définie par

$$d_M(h) = \int E(r_n^*(x) - r(x))^2 f(x) dx .$$

Dans [11] Gérard Collomb a calculé $d_M(h)$ et montré que si r était 2 fois continûment différentiable, on avait pour $d = 1$

$$d_M(h) = n^{-1} h^{-1} C_1 + h^4 C_2 \tag{4.1}$$

où les constantes C_1 et C_2 dépendent respectivement de la variance conditionnelle $\text{Var}(Y/X = x)$ et de $r''(x)$. L'approximation (4.1) doit être comprise au sens que tous les termes d'ordre inférieur à $n^{-1} h^{-1} + h^4$ ont été supprimés. A la lumière de cette approximation, il semble désirable de choisir h_n proportionnel à $n^{-1/5}$ mais dans la pratique les constantes C_1 et C_2 sont généralement inconnues. Pour surmonter cet obstacle, une méthode de sélection de h_n entièrement basée sur les données est nécessaire. Pour des raisons de simplicité, nous supposons dorénavant que $d = 1$.

Pour fixer les idées, considérons la définition suivante.

Définition

Une méthode de sélection \hat{h} est dite asymptotiquement optimale lorsque

$$\frac{d_M(\hat{h})}{\inf_{h \in H_n} d_M(h)} \xrightarrow[n \rightarrow \infty]{p.} 1,$$

où H_n est un ensemble (éventuellement fini) de valeurs pour h_n .

Cette définition dit que le risque relatif lorsqu'on sélectionne \hat{h} à partir des données tend vers 1. En utilisant la convexité de $d_M(h)$, voir formule (4.1), il est clair qu'une sélection \hat{h} asymptotiquement optimale résout le problème de l'estimation de C_1 et C_2 .

Comment trouver une sélection \hat{h} asymptotiquement optimale ? Regardons tout d'abord certains travaux récents concernant le cas indépendant. Dans ce cas la technique du "leave-one-out" peut être employée pour construire l'estimateur suivant de l'erreur de prédiction :

(1986) Härdle, W. Quelques aspects de la prédiction non paramétrique: travaux de Gerard Collomb (1951-1985) en analyse non paramétrique des séries temporelles

$$CV(h) = n^{-1} \sum_{i=1}^n (Y_i - r_{n,i}^*(X_i))^2, \quad (4.2)$$

où $r_{n,i}^*$ est l'estimateur de Nadaraya-Watson basé sur l'échantillon privé de la ième observation.

En insérant $r(X_i) - r(X_i)$ à l'intérieur des parenthèses et en développant on obtient

$$CV(h) = n^{-1} \sum_{i=1}^n \varepsilon_i^2 + d_A^i(h) + 2 C(h),$$

où

$$d_A^i(h) = n^{-1} \sum_{i=1}^n (r(X_i) - r_{n,i}^*(X_i))^2$$

est une mesure quadratique de la valeur de l'estimateur r_n^* , où

$$n^{-1} \sum_{i=1}^n \varepsilon_i^2 = n^{-1} \sum_{i=1}^n (Y_i - r(X_i))^2$$

est un terme indépendant de h, et où

$$C_n(h) = n^{-1} \sum_{i=1}^n \varepsilon_i (r(X_i) - r_{n,i}^*(X_i)).$$

Si le terme croisé $C_n(h)$ s'annule quand $n \rightarrow \infty$ uniformément sur H_n , alors (4.2) donne une possibilité de sélection de h. Dans un article récent, Härdle et Marron (1985) ont prouvé que la méthode consistant à choisir \hat{h} minimisant $CV(h)$ est asymptotiquement optimale. Leur preuve peut se décomposer en deux étapes :

$$\sup_{h \in H_n} \left| \frac{d_A^i(h) - d_M(h)}{d_M(h)} \right| = Op(1), \quad (4.3)$$

$$\sup_{h \in H_n} \left| \frac{C_n(h)}{d_M(h)} \right| \xrightarrow{p.s.} 0. \quad (4.4)$$

Une approche similaire peut être envisagée dans notre cas, mais on ne doit pas s'attendre à voir le terme croisé s'annuler asymptotiquement, à moins de modifier la technique du "leave-one-out". Dans le cas indépendant, cette technique introduit une structure d'indépendance spécifique dans $C_n(h)$ qui permet de conclure. Pour utiliser la même idée dans notre cas de processus Φ -mélangeant, il est nécessaire d'écartier plus d'une observation à la fois. Ainsi, nous définissons l'estimateur "leave-not-too-many-out" par

$$r_{n,i}^*(x) = (n - \rho_n)^{-1} h^{-1} \sum_{|i-j| \geq \rho_n} K\left(\frac{x - X_j}{h_n}\right) Y_j / \hat{f}_{n,i}(x), \quad (4.5)$$

où

$$\hat{f}_{n,i}(x) = (n - \rho_n)^{-1} h^{-1} \sum_{|i-j| \geq \rho_n} K\left(\frac{x - X_j}{h_n}\right),$$

où $\{\rho_n, n \in \mathbb{N}\}$ est une suite qui croît lentement.

Nous notons que si $\rho_n = 1$, cet estimateur est l'estimateur "leave-one-out" utilisé en (4.2). Définissons, de manière similaire à (4.2),

$$CV(h) = n^{-1} \sum_{i=1}^n (Y_i - r_{n,i}^*(X_i))^2. \quad (4.6)$$

Les détails de cette analyse se trouvent dans l'article de Collomb, Härdle et Vieu (1985). Nous voulons simplement décrire ce qui dans une partie du terme équivalent à $C_n(h)$ nécessite des considérations supplémentaires. Cette partie est (voir (4.5))

$$T(h) = n^{-1} \sum_{i=1}^n (n - \rho_n)^{-1} \sum_{|i-j| \geq \rho_n} h^{-1} K((X_i - X_j)/h_n) \varepsilon_i \varepsilon_j$$

où

$$\varepsilon_i = Y_i - m(X_i) \quad \forall i \in \mathbb{N}.$$

Nous devons montrer que pour $H_n \subset [a_n^{-1/s}, b_n^{-1/s}]$, $0 < a < b$, nous avons

$$P \left[\sup_{h \in H_n} \left| \frac{T(h)}{d_M(h)} \right| > \tau \right] \rightarrow 0, \quad \forall \tau > 0.$$

En utilisant l'inégalité de Bonferroni et celle de Tchebicheff, cette probabilité est bornée par

$$\tau^{-4} \# H_n n^4 \sup_{h \in H_n} h^4 E[T(h)^4]. \quad (4.7)$$

Le problème revient donc à trouver une borne convenable pour $E[T^4(h)]$. Ceci se trouve dans Collomb, Härdle et Vieu (1984). Il faut remarquer que des techniques de calcul du type de celles de Doukhan et Portal (1983) ne peuvent pas être utilisées ici puisque $T(h)$ est formé d'une double somme. En fait, un argument, que nous appelons "big block-small block" permet de montrer que

$$E[T(h)^4] \leq C_3 F(\rho_n) n^{-4} h^{-2}, \quad (4.8)$$

ce qui avec (4.7) prouve que la probabilité pour que $T(h)/d_M(h)$ s'annule uniformément sur H_n est majorée par

$$C_4 \# H_n F(\rho_n) h^2.$$

En supposant que $\# H_n$ est d'ordre algébrique et que $F(\rho_n)$ ne croît pas trop vite un résultat analogue à (4.4) est établi.

Après avoir établi (4.3) dans le cas de variables Φ -mélangeantes, nous obtenons l'optimalité asymptotique de h :

Théorème 3

Si l'on choisit \hat{h} minimisant $CV(h)$ défini en (4.6), alors \hat{h} est asymptotiquement optimal, i.e.,

$$\frac{d_M(\hat{h})}{\inf_{h \in H_n} d_M(h)} \xrightarrow[n \rightarrow \infty]{P.} 1.$$

Nous renvoyons à Collomb, Härdle et Vieu (1985) pour une démonstration intégrale de ce résultat. Cet article était en fait le dernier projet sur lequel travaillait Gérard Collomb. Il mourut peu de temps après avoir terminé la première rédaction de la démonstration (4.8).

(1986) Härdle, W. Quelques aspects de la prédiction non paramétrique: travaux de Gerard Collomb (1951-1985) en analyse non paramétrique des séries temporelles

BIBLIOGRAPHIE DE GÉRARD COLLOMB

1977

- [1] Quelques propriétés de la méthode du noyau pour l'estimation non paramétrique de la régression en un point fixé. *C.R. Acad. Sci. Paris*, t. 285, Série A, 289-292.
- [2] Estimation non paramétrique de la régression par la méthode du noyau : propriété de convergence asymptotiquement normale indépendante. *Ann. Scientifiques de l'Université de Clermont*, 15, 24-46.

1979

- [3] Conditions nécessaires et suffisantes de convergence uniforme d'un estimateur de la régression, estimation des dérivées de la régression. *C.R. Acad. Sci. Paris*, t. 288, Série A, 161-164.
- [4] Estimation de la régression par la méthode des k points les plus proches : propriétés de convergence ponctuelle. *C.R. Acad. Sci. Paris*, t. 289, Série A, 245-247.

1980

- [5] Estimation non paramétrique de probabilités conditionnelles. *C.R. Acad. Sci. Paris*, t. 291, Série A, 427-430.

1981

- [6] Estimation de la régression par la méthode des k points les plus proches avec noyau. *Lectures Notes in Mathematics*, 821, 159-175.
- [7] Estimation non paramétrique de la régression : revue bibliographique. *International Statistical Review*, 49, 75-93.

1982

- [8] Prédiction non paramétrique : étude de l'erreur quadratique du predictogramme. *C.R. Acad. Sci. Paris*, t. 294, Série I, 59-62.
- [9] From Data Analysis to Non Parametric Statistics : Second Developments and a Computer Realization for Exploratory Techniques in Regression or Prediction. *Compstat 1982, Proc. Computational Statistics*, Physica Verlag, Wien, 173-178.

1983

- [10] From Non Parametric Regression to Non Parametric Prediction : Survey on the Mean Square Error and Original Results on the Predictogram. *Lectures Notes in Statistics*, 16, 182-204.

1984

- [11] Prédiction non paramétrique : étude de l'erreur quadratique du prédicogramme. *Statistique et Analyse des Données*, 9, 1, 1-34.

(1986) Härdle, W. Quelque aspects de la prediction non parametrique: travaux de Gerard Collomb (1951-1985) en analyse non parametrique des series temporelles

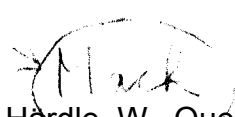
- [12] Estimation non paramétrique de la fonction d'autorégression d'un processus stationnaire et φ -mélangeant : risques quadratiques pour la méthode du noyau (avec P. Doukhan). *C.R. Acad. Sci. Paris, t. 296, Série I, 859-862.*
- [13] Propriétés de convergence presque complète du prédicteur à noyau. *Z. Wahrsch. verw. Gebiete, 66, 441-460.*
- [14] Robustness in parametric and non parametric regression estimation : an investigation by computer simulations (avec J. Antoch, S. Hassani). *Compstat 1984, Proc. Computational Statistics, Physica Verlag, Wien, 49-54.*

1985

- [15] Nonparametric regression : An up-to-date bibliography. *Math. Oper. Stat. Series Statistics, 16, 2, 309-324.*
- [16] Nonparametric time series analysis and prediction : uniform almost sure convergence of the window and k-NN autoregression estimates. *Math. Oper. Stat. Series Statistics, 16, 2, 297-307.*
- [17] Contribution to the discussion of "Some aspects of the spline smoothing approach to non-parametric regression curve fitting" of B.W. Silverman. *J. Royal Stat. Soc., à paraître*
- [18] A Note on prediction via estimation of the conditional mode function (avec W. Härdle et S. Hassani), soumis pour publication.
- [19] Strong uniform convergence rates in robust nonparametric time series analysis and prediction : kernel regression estimation from dependent observations (avec W. Härdle) *Stoch. Proc. and its Appl., à paraître.*
- [20] Optimal Nonparametric Time Series Prediction (avec W. Härdle et P. Vieu), manuscrit.

REFERENCES

- HÄRDLE, W. (1984). Robust regression function estimation. *J. Mult. Analysis, 14, 169-180.*
- HÄRDLE, W. et GASSER, Th. (1984). Robust nonparametric function fitting. *J. Royal Stat. Soc. (B), 46, 42-51.*
- HÄRDLE, W. et LUCKHAUS, S. (1984). Uniform consistency of a class of regression function estimators, *Ann. Statist. 12, 612-623.*
- HÄRDLE, W. et MARRON, S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist., 13,*
- LAWRENCE, A.J. et LEWIS, P.A.W. (1980). The exponential autoregressive-moving average EARMA(p,q) Process. *J. Royal Stat. Soc. (B) 42, 150-161.*
- MACH, Y.P. et SILVERMAN, B.W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. verw. gebiete, 61, 405-415.*
- MARTIN, R.D. et YOHAÏ, V. (1984). Gross-Error sensitivities of GM and RA estimates. *Lecture Notes in Statistics (ed. Franke, Härdle, Martin), 26, 198-217.*
- NADARAYA, E.A. (1964). On estimating regression. *Theor. Prob. and its Appl., 9, 141-142.*
- ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat. 27, 832-837.*
- ROBINSON, P. (1984). Robust nonparametric autoregression. *Lecture Notes in Statistics (ed. Franke, Härdle, Martin), 26, 247-255.*
- ROBINSON, P. (1984). Robust and nonlinear time series analysis. Proceedings of a workshop held at the University of Heidelberg 1983. *Lecture Notes in Statistics, 26 (ed. Franke, Härdle, Martin).*


 (1986) Härdle, W. Quelques aspects de la prédiction non paramétrique : travaux de Gérard Collomb (1951-1985) en analyse non paramétrique des séries temporelles

- SARDA, P. et VIEU, P. (1985). Estimation non paramétrique de la régression pour des variables dépendantes, application à la prédiction pour un processus markovien. Manuscrit.
- STONE, C. (1977). Consistent nonparametric regression. *Ann. Stat.* 5, 595-620.
- TONG, H. et LIN, K.S. (1980). Threshold autoregression, limit cycles and cyclical data (with Discussion). *J. Royal. Stat. Soc. (B)*, 42, 245-292.
- TSYBAHOV, A.B. (1983). Robust estimates of a function. *Prob. Inf. Theory*, 18, 39-52.
- WATSON, G.S. (1964). Smooth regression analysis. *Sankhya* 26, ser. A, 359-372.

TSYBAHOV

Approximations to the Mean Integrated Squared Error with Applications to Optimal Bandwidth Selection for Nonparametric Regression Function Estimators

WOLFGANG HÄRDLE*

Universität Heidelberg,
Sonderforschungsbereich 123, Im Neuenheimer Feld 293,
D-6900 Heidelberg 1, West Germany, and
University of North Carolina, Department of Statistics,
321 Phillips Hall 039A, Chapel Hill, North Carolina 27514

Communicated by G. Kallianpur

Discrete versions of the mean integrated squared error (MISE) provide stochastic measures of accuracy to compare different estimators of regression functions. These measures of accuracy have been used in Monte Carlo trials and have been employed for the optimal bandwidth selection for kernel regression function estimators, as shown in Härdle and Marron (1983), *Optimal Bandwidth Selection in Nonparametric Regression Function Estimation*. Inst. of Statistics Mimeo Series No. 1530, Univ. of North Carolina, Chapel Hill). In the present paper it is shown that these stochastic measures of accuracy converge to a weighted version of the MISE of kernel regression function estimators, extending a result of Hall (1982, *Biometrika* **69**, 383–390) and Marron (1983, *J. Multivariate Anal.* **18**, No. 2) to regression function estimation. © 1986 Academic Press, Inc.

1. INTRODUCTION AND BACKGROUND

Let $(X_1, Y_1), (X_2, Y_2), \dots$, be independent random vectors distributed as (X, Y) with common joint probability density function $f(x, y)$ and let $m(x) = E(Y | X = x) = \int yf(x, y) dy / f_X(x)$, f_X the marginal density of X , be

* Research partially supported by the "Deutsche Forschungsgemeinschaft" SFB123, "Stochastische Mathematische Modelle," partially supported by the Air Force of Scientific Research Contract AFOSR-F49620 82 c 0009.

Received October 21, 1983; revised November 29, 1983.

AMS 1980 subject classifications: Primary 60F05; Secondary 62G05.

Keywords and phrases: stochastic measure of accuracy, nonparametric regression function estimation, optimal bandwidth selection, limit theorems, mean square error.

the regression curve of Y on X . Let $m_n^*(x)$ denote the nonparametric kernel estimate of $m(x)$, as introduced by Nadaraya [12] and Watson [21],

$$m_n^*(x) = \hat{m}_n(x)/f_n(x) \quad (1.1)$$

where

$$\hat{m}_n(x) = n^{-1}h^{-1} \sum_{i=1}^n K((x - X_i)/h) Y_i$$

and

$$f_n(x) = n^{-1}h^{-1} \sum_{i=1}^n K((x - X_i)/h).$$

Here K is a kernel function and $h = h(n)$ is a sequence of "bandwidths" converging to zero as n tends to infinity.

This estimator was studied by Rosenblatt [15] who derived bias, variance, and asymptotic normality; Schuster [17] demonstrated multivariate normality at a finite number of distinct points. For further results we refer to the bibliography of Collomb [3].

In the present paper we show that

$$A_n^*(h) = n^{-1} \sum_{j \in \mathcal{J}} [m_n^*(X_j) - m(X_j)]^2, \quad \mathcal{J} = \{j : X_j \in [0, 1]\}, \quad (1.2)$$

a stochastic measure of accuracy on the interval $[0, 1]$ for the estimate m_n^* , exhibits the same limiting behaviour as the deterministic measure

$$\text{MISE} = \int_0^1 \text{MSE}(t) f_X(t) dt \quad (1.3)$$

where $\text{MSE}(t)$ is the mean squared error (MSE) of $m_n^*(t)$. The proper definition of the MSE for m_n^* will be delayed to Section 2.

The result of this paper addresses two problems. First, in a survey paper, Wegman [22] was interested in comparing the mean integrated squared error (MISE) of several different density estimators. As Wegman pointed out, the computation of the actual MISE can be quite tedious. Hence, Wegman used an empirical measure of accuracy of the structure as in formula (1.2) and gave some heuristic justification. Now, since the bias/variance decomposition of regression function estimators is rather similar to that of density estimators [15, 16] it may be argued that Wegman's heuristics hold also in the regression function estimation setting. The answer is positive: It is shown here that, as $n \rightarrow \infty$, uniformly over an interval $[h \bar{h}]$,

$$A_n^*(h) = \text{MISE} + o_p(\text{MISE}), \quad h \in [h \bar{h}]. \quad (1.4)$$

The appealing feature of this approximation is, that it holds uniformly in $h \in [h, A]$. A Monte Carlo trial comparing different estimators of $m(x)$ (w.r.t. MISE) at different sequences of bandwidths can thus be based on $A_n^*(h)$ which is faster to compute than MISE as defined in (1.3).

Second, the approximation (1.4) contributes to the solution of the "optimal bandwidth selection" problem. As the optimal bandwidth h^* we understand that sequence $h = h(n)$ which minimizes the MISE for each n . Hardle and Marron [5] demonstrated by a crossvalidation argument that minimization (with respect to h) of $A_n^*(h)$ is asymptotically equivalent to minimization of

$$n^{-1} \sum_{j \in \mathcal{J}} [Y_j - m_n^{*(j)}(X_j)]^2, \tag{1.5}$$

where

$$m_n^{*(j)}(x) = n^{-1} h^{-1} \sum_{i \neq j} K((x - X_i)/h) Y_i / f_n(x)$$

is the "leave-one-out" estimator. So the result of this paper, as stated in (1.4), ensures that the minimization of (1.5) with respect to h yields the (MISE)-optimal sequence of bandwidth h^* and solves, as is shown in Hardle and Marron, a problem raised by Stone [19, Qestion 3, p. 10541.

We will not only analyze $m_n^*(x)$, as defined in (1.1), but also

$$\hat{m}_n(x) / f_X(x) \tag{1.6}$$

where f_X denotes the marginal density of X . This estimator of $m(x)$ is reasonable if we know the marginal density and is somewhat more tractable than m_n^* . The estimator (1.6) was studied by Johnston [8], who also observed that \hat{m}_n / f_X has in general a higher asymptotic variance than m_n^* .

The stochastic measure of accuracy (1.2) was defined only on the interval $[0, 1]$. It will later be assumed that the support of f_X properly contains this interval. This is due to "boundary effects," more precisely, the bias at the endpoints of the support of f_X inflates and has a slower rate than in the interior [4, 13]. Thus, defining the MISE over the whole support off,, would ultimately lead to the unappealing situation that the optimal bandwidth with respect to MISE would be determined in such a way that it minimizes the mean square error at the boundaries, since that is of lower order. The estimate in the interior would thus exhibit suboptimal behaviour.

The results of this paper are improvements over some previous work for several reasons. First, we do not need such strong smoothness assumptions on f_X as in Hall [6], who proves similar results in the density estimation setting. Second, our assumptions on the variance curve $V^2(t) =$

$\text{var}(Y|X=t)$ and the range of allowable bandwidths are considerably weaker than those in Johnston [8] who demonstrates a Gaussian approximation to $(nh)^{1/2} [m, -Em]$ along the same lines as Bickel and Rosenblatt [1]. Third, our work extends the result of Wong [23] who deals only with the fixed design case, i.e., X , are nonrandom. Finally, we may note that Hall's proof would simplify if one uses the approximation provided by the Bickel and Rosenblatt paper and the outline of the proof given here for regression function estimators.

Note that although only the two-dimensional case is considered here, the proof can probably be extended to the higher dimensional case where we observe a $(d+1)$ -dimensional random vector (X_1, \dots, X_d, Y) , $d > 1$. The assumptions will be different in that case, since it is still unknown whether the multivariate empirical process can be strongly approximated by Brownian bridges with rates comparable to those in the univariate or bivariate case. This approximation technique by Brownian bridges, as carried out in the Appendix, is vital to our results. A similar technique, exploiting the idea of invariance principles in nonparametric regression, was used by Mack and Silverman [9] who showed weak and strong uniform consistency (in sup-norm) of m_n^* .

The outline of the paper is organized as follows. First, we prove that $\hat{m}_n(t) - E\hat{m}_n(t)$ can be uniformly (in t and h) approximated by a Gaussian process similar to that occurring in Bickel and Rosenblatt [1, p. 1974, formula (2.5)]. Second, we plug this approximating process into the formula (1.2), which defined the discrete version of MISE, and by evaluation of covariances and higher moments we finally arrive at the deterministic measure (1.3).

2. RESULTS

We will make use of the following definition.

DEFINITION. A function w is called Lipschitz-continuous of order α ($LC(\alpha)$) iff with a constant L ,

$$|w(t) - w(t')| \leq L_w |t - t'|^\alpha, \quad 0 < \alpha \leq 1.$$

The following assumptions fix the range of allowable bandwidths $[h, \bar{h}]$, determine the kernel function K and describe some smoothness of $m(t)$, $\text{var}(Y|X=t)$, and $f_X(t)$:

(A1) Let $\{h_n\}$ denote a sequence for which there is an $\varepsilon > 0$ so that

$$\lim_{n \rightarrow \infty} h_n n^{1/3 - \varepsilon} / \log n = 0, \quad \lim_{n \rightarrow \infty} h_n n^{1/2 - \varepsilon} = \infty$$

and let $\{\bar{h}_n\}$ denote a sequence for which

$$\lim_{n \rightarrow \infty} \bar{h}_n = 0, \quad \lim_{n \rightarrow \infty} \bar{h}_n \log n = \infty.$$

Assume from $h = h(n)$ that it satisfies

$$\underline{h} \leq h \leq \bar{h}.$$

(A2) There exists a sequence of positive constants $\{a_n\} \uparrow \infty$ and a $c < \infty$ such that

$$\begin{aligned} \sup_{\underline{h} \leq h \leq \bar{h}} h^{-3} \int_{|y| > a_n} y^2 f_Y(y) dy &\leq c, & f_Y \text{ the marginal density of } Y \\ \lim_{n \rightarrow \infty} \sup_{0 \leq x \leq 1} \int_{|y| > a_n} y^2 f(x, y) dy &= 0 \\ \lim_{n \rightarrow \infty} \sup_{\underline{h} \leq h \leq \bar{h}} n^{-1/2} h^{-1/2} a_n (\log n)^2 &= 0 \\ |g_n(x)| = \left| \int_{-a_n}^{a_n} y^2 f(x, y) dy \right| &\geq \eta > 0 \quad \text{for all } 0 \leq x \leq 1, n \geq 1. \\ \int |d_u[g_n(uh)]| &= o(\{\log(1/h)\}^{1/2}). \end{aligned}$$

(A3) The functions $S^2(t) = E[Y^2 | X = t]$, $f_X(t)$ and $m(t)$ are $LC(\alpha)$ with $\alpha > \frac{1}{2}$ and are all of bounded variation. The marginal density of X is bounded from below:

$$\inf_{0 \leq t \leq 1} f_X(t) \geq \gamma > 0.$$

(A4) The kernel function K is differentiable with K' of bounded variation and fulfills

$$\int K(u) du = 1 \quad \text{support}\{K\} \subset [-A, A].$$

K is not assumed to be positive.

By straightforward computations it can be shown that g_n is $LC(\alpha)$, $\alpha > \frac{1}{2}$ and of bounded variation by assumption (A3) on $S^2(t)$ and $f_X(t)$. It is also not hard to see that if g_n is $LC(1)$ then the last condition in (A3) follows. Note that the set of assumptions in (A2) holds if Y is bounded ($a_n = \log \log n$), an assumption that is often made in other papers, to avoid conditions on moments of Y as in (A2). (A2) also holds, if $a_n = n^\beta$, β small, while (X, Y) are jointly normally distributed. For simplicity of notation, we will not explicitly write the indices of $\bar{h}_n, \underline{h}_n$.

The following results show that the approximation (1.4) holds for both \hat{m}_n/f_X and m_n^* . Only the proof of Theorem 1 (dealing with \hat{m}_n/f_X) will be given in full detail since the result for m_n^* can be obtained quite analogously. Let us define

$$\beta_k = \int_{-A}^A K^2(u) du$$

and

$$\hat{b}_n(t) = f_X^{-1}(t) \int_{-A}^A K(u) [m(t-uh) f_X(t-uh) - m(t) f_X(t)] du,$$

the bias of \hat{m}_n/f_X .

THEOREM 1. Assume that (A1) to (A4) hold and $\hat{b}_n(t)$ is of bounded variation. Then uniformly over $h \in [h \hat{h}]$

$$\begin{aligned} \hat{A}_n(h) &= n^{-1} \sum_{j \in \mathcal{J}} [\hat{m}_n(X_j)/f_X(X_j) - m(X_j)]^2 \\ &= (nh)^{-1} \beta_k \int_0^1 S^2(t) dt \\ &\quad + \int_0^1 [\hat{b}_n(t)]^2 f_X(t) dt \\ &\quad + o_p\left((nh)^{-1} + \int_0^1 [\hat{b}_n(t)]^2 dt\right) \\ &= \text{MISE}[\hat{m}_n/f_X] + o_p(\text{MISE}). \end{aligned}$$

Assume that f_X is d_1 -times continuously differentiable and m is d_2 -times continuously differentiable. Then, as in Rosenblatt [16], the bias $\hat{b}_n(t)$ would read as

$$\hat{b}_n(t) \simeq h^d \Lambda_d p^{(d)}(t)/f_X(t), \quad p = m f_X, \quad d = d_1 \wedge d_2$$

provided that K satisfies $\int u^j K(u) du = 0, j = 1, \dots, d-1$, and $\int u^d K(u) du = d! \Lambda_d$. Many papers in nonparametric regression function estimation assume such a kind of differentiability as above and are dealing with methods to balance the contribution from the variance and the bias (see [3] for a review).

In a similar manner define $b_n^*(t)$, the bias of $m_n^*(t)$, as follows

$$b_n^*(t) = f_X^{-1}(t) \int_{-A}^A K(u) [m(t-uh) - m(t)] f_X(t-uh) du.$$

Where the expression "bias" has to be understood as the expected value of $f_x^{-1}[\hat{m}_n - mf_n]$, $f_n(t) = n^{-1}h^{-1} \sum_{i=1}^n K((t - X_i)/h)$ a density estimate of the marginal density f_x . This is justified by the observation that

$$m_n^* - m = [m_n - mf_n]/f_x + o_p(\hat{m}_n - mf_n)$$

(see [5]) and that moments of m_n^* need not exist in general [15].

The next theorem shows how $A_n^*(h)$ approximates the MISE.

THEOREM 2. Assume that (A1) to (A4) hold and that $b_n^*(t)$ is of bounded variation. Then uniformly over $h \in [h, h]$,

$$\begin{aligned} A_n^*(h) &= n^{-1} \sum_{j \in \mathcal{J}} [m_n^*(X_j) - m(X_j)]^2 \\ &= (nh)^{-1} \beta_k \int_0^1 V^2(t) dt \\ &\quad + \int_0^1 [b_n^*(t)]^2 f_x(t) dt \\ &\quad + o_p\left((nh)^{-1} + \int_0^1 [b_n^*(t)]^2 dt\right) \\ &= \text{MISE}[m_n^*] + o_p(\text{MISE}), \end{aligned}$$

where $V^2(t) = S^2(t) - m^2(t)$.

Note that the variance terms and the bias terms of the two estimators \hat{m}_n/f_x and m_n^* are completely different. Since $V^2(t) \leq S^2(t)$, the Nadaraya-Watson estimator $m_n^*(t)$ attains in general a smaller (asymptotic) variance than \hat{m}_n/f_x . This was also observed by Johnston [8]. The condition " $nh^5 \rightarrow 0$ ", appearing in the work of the latter, implies that the bias vanishes asymptotically faster than the variance. Therefore, any difference in bias terms does not show up in that work. It would be interesting to find a similar comparison of bias terms, but this would lead to complicated and rather unnatural assumptions on derivatives of m and f_x , as can be seen from the formula for \hat{b}_n , following Theorem 1.

3. THE PROOFS

We shall prove Theorem 1 in full detail, the proof of Theorem 2 will only be sketched since the technical details are similar to the proof of Theorem 1. $F(x, y)$ will denote the joint cumulative distribution function (df) of (X, Y) and $F_n(x, y)$ will denote the two-dimensional empirical df,

defined as usual. It is understood throughout these proofs that $o, 0$ in remainder terms are uniform over $h \in [\underline{h}, \bar{h}]$.

Proof of Theorem 1. The basic decomposition is

$$\hat{m}_n(t)/f_X(t) - m(t) = \hat{Y}_n(t) + \hat{b}_n(t) \tag{3.1}$$

where

$$\hat{Y}_n(t) = f_X^{-1}(t) h^{-1} \left[\int_{-\infty}^{\infty} y K((t-x)/h) d[F_n(x, y) - F(x, y)] \right].$$

In the Appendix it is shown that

$$Y_{o,n}(t) = [S^2(t)/f_X(t)]^{-1/2} \hat{Y}_n(t) - n^{-1/2} h^{-1} \int_{-\infty}^{\infty} K((t-x)/h) dW(x) + o_p(n^{-1/2} h^{-1/2}),$$

where the remainder term is uniform in t . The basic decomposition (3.1) now reads

$$\hat{m}_n(t)/f_X(t) - m(t) = n^{-1/2} h^{-1/2} V_n(t) + \hat{b}_n(t) + \rho_n \tag{3.2}$$

where $\rho_n = o_p(n^{-1/2} h^{-1/2})$ is uniformly in t and

$$V_n(t) = [S^2(t)/f_X(t)]^{1/2} h^{-1/2} \int_{-\infty}^{\infty} K((t-x)/h) dW(x). \tag{3.3}$$

Using (3.2) and (3.3) the stochastic measure of accuracy is then

$$\begin{aligned} \hat{A}_n(h) = & \int_0^1 [\hat{b}_n(t)]^2 dF_{X,n}(t) \\ & + n^{-1} h^{-1} \int_0^1 V_n^2(t) dF_{X,n}(t) \\ & + 2n^{-1/2} h^{-1/2} \int_0^1 \hat{b}_n(t) V_n(t) dF_{X,n}(t) \\ & + \rho_n \left\{ 2 \left[\int_0^1 \hat{b}_n(t) dF_{X,n}(t) + n^{-1/2} h^{-1/2} \int_0^1 V_n(t) dF_{X,n}(t) \right] + \rho_n \right\}, \end{aligned}$$

where $F_{X,n}$ denotes the empirical distribution function of $\{X_i\}_{i=1}^n$. This can be rewritten as

$$\begin{aligned} \hat{A}_n(h) = & n^{-1} \sum_{j \in \mathcal{J}} [\hat{b}_n(X_j)]^2 \\ & + n^{-1} h^{-1} [U_{n1} + U_{n2}] \\ & + 2n^{-1/2} h^{-1/2} [U_{n3} + U_{n4}] \\ & + \rho_n \left\{ 2 \left[\int_0^1 \hat{b}_n(t) dF_{X,n}(t) + n^{-1/2} h^{-1/2} \int_0^1 V_n(t) dF_{X,n}(t) \right] + \rho_n \right\} \end{aligned}$$

where

$$\begin{aligned} U_{n1} &= \int_0^1 V_n^2(t) f_X(t) dt \\ U_{n2} &= \int_0^1 V_n^2(t) d[F_{X,n}(t) - F_X(t)] \\ U_{n3} &= \int_0^1 V_n(t) \hat{b}_n(t) f_X(t) dt \\ U_{n4} &= \int_0^1 V_n(t) \hat{b}_n(t) d[F_{X,n}(t) - F_X(t)]. \end{aligned}$$

We now show that the limits of U_{ni} , $i=1, 2, 3, 4$ give us the desired limit behaviour of $\hat{A}_n(h)$. We may note that the approximations, as carried out in Bickel and Rosenblatt [1], would have led to a process similar to $V_n(t)$ when estimating a density. So the technique developed here, would be useful in density estimation also and would provide an alternative proof of Hall's [6] result on stochastic measures of accuracy for density estimators.

Let us begin with the limit behaviour of U_{n1} . Note first that

$$\begin{aligned} EU_{n1} &= \int_0^1 E \left\{ h^{-1/2} \int_{-\infty}^{\infty} K((t-x)/h) dW(x) \right\}^2 S^2(t) dt \\ &= \int_0^1 h^{-1} \int_{-\infty}^{\infty} K^2((t-x)/h) dx S^2(t) dt \\ &= \int_0^1 \int_{-A}^A K^2(u) S^2(t-uh) du dt \\ &= \beta_k \int_0^1 S^2(t) dt + o(1). \end{aligned}$$

where the remainder term is uniform in h , since $S^2(t)$ is $LC(\alpha)$, $\alpha > \frac{1}{2}$ by assumption (A3). To show that

$$U_{n1} \xrightarrow{p} \int_{-A}^A K^2(u) du \int_0^1 S^2(t) dt \tag{3.4}$$

we demonstrate $E(U_{n1}^2) \sim (EU_{n1})^2$. The statement (3.4) will then follow from Chebyshev's inequality.

Since $Z(t) = h^{-1/2} \int_{-\infty}^{\infty} K((t-x)/h) dW(x)$ is a Gaussian process we conclude by the Isserlis [7] formula

$$\begin{aligned} EU_{n1}^2 &= \int_0^1 \int_0^1 \{EZ^2(t_1)EZ^2(t_2) + 2[E[Z(t_1)Z(t_2)]]^2\} \\ &\quad \times S^2(t_1)S^2(t_2) dt_1 dt_2 \\ &= \int_0^1 \int_0^1 S^2(t_1)S^2(t_2) \\ &\quad \times \left\{ h^{-2} \int K^2((t_1-x_1)/h) dx_1 \int K^2((t_2-x_2)/h) dx_2 \right. \\ &\quad \left. + 2h^{-2} \left[\int K((t_1-x)/h)K((t_2-x)/h) dx \right]^2 \right\} dt_1 dt_2. \end{aligned}$$

The first summand satisfies

$$\begin{aligned} &\int_0^1 \int_0^1 S^2(t_1)S^2(t_2)h^{-2} \int K^2((t_1-x_1)/h) dx_1 \int K^2((t_2-x_2)/h) dx_2 dt_1 dt_2 \\ &= \left[\beta_k \int_0^1 S^2(t) dt \right]^2 + O(h) \end{aligned}$$

by assumption (A4) on the kernel K .

The second summand satisfies

$$\int_0^1 \int_0^1 S^2(t_1)S^2(t_2)2h^{-2} \left[\int K((t_1-x)/h)K((t_2-x)/h) dx \right]^2 dt_1 dt_2 = O(h)$$

by evaluation of the integral inside the [.] -brackets. This shows that

$$U_{n1} = \beta_k \int_0^1 S^2(t) dt + o_p(1).$$

Next we show that

$$U_{n2} = O_p(n^{-1/2}h^{-1}) \tag{3.5}$$

Define $H_n(t) = F_{X,n}(t) - F_X(t)$ and $Z_n(t) = \int_{-\infty}^{\infty} K((t-x)/h) dW(x)$. We obtain by partial integration,

$$\begin{aligned}
 hU_{n2} = & -2 \int_0^1 H_n(t) q(t) Z_n(t) \left[h^{-1} q(t) \int_{-\infty}^{\infty} K'((t-x)/h) dW(x) \right] dt \\
 & -2 \int_0^1 H_n(t) q(t) Z_n^2(t) dq(t) \\
 & + hH_n(t) V_n^2(t)|_0^1,
 \end{aligned}$$

where $q(t) = S^2(t)/f_X(t)$.

Now since $H_n(t) = O_p(n^{-1/2})$ uniformly in t and $V_n^2(t_0) = O_p(1)$, $t_0 = 0, 1$, as is easily verified by Chebyshev's inequality, we only have to consider the first two summands in the equality above.

These are further estimated by Schwarz's inequality, which shows that the absolute value of the sum of both is dominated by

$$\begin{aligned}
 & n^{-1/2} \sup_{0 \leq t \leq 1} |n^{1/2} H_n(t)| \times \left\{ S_1 \left[\int_0^1 [h^{-1/2} Z_n(t)]^2 dt \right]^{1/2} \right. \\
 & \quad \times \left. \left[\int_0^1 \left[h^{-1/2} \int_{-\infty}^{\infty} K'((t-x)/h) dW(x) \right]^2 dt \right]^{1/2} \right. \\
 & \quad \left. + S_2 \sup_{0 \leq t \leq 1} |Z_n^2(t)| \int_0^1 |dq(t)| \right\},
 \end{aligned}$$

where $S_1 = \sup_{0 \leq t \leq 1} q^2(t)$ and $S_2 = \sup_{0 \leq t \leq 1} q(t)$.

By Chebyshev's inequality we have

$$\int_0^1 \left[h^{-1/2} \int_{-\infty}^{\infty} L((t-x)/h) dW(x) \right]^2 dt = O_p(1)$$

where L is either K or K' . Integration by parts applied to $Z_n^2(t)$ show immediately that $\sup_{0 \leq t \leq 1} Z_n^2(t) = O_p(1)$, therefore (3.5) holds. Now, since

$$\begin{aligned}
 EU_{n3}^2 = & \int_0^1 \int_0^1 \left\{ h^{-1} \int_{-\infty}^{\infty} K((t_1-x)/h) K((t_2-x)/h) dx \right\} \\
 & \times \hat{b}_n(t_1) \hat{b}_n(t_2) q(t_1) q(t_2) dt_1 dt_2 \\
 \leq & o \left(\int_0^1 [\hat{b}_n(t)]^2 dt \right)
 \end{aligned}$$

by an application of Schwarz's inequality, we conclude that

$$U_{n3} = o_p \left(\left[\int_0^1 [\hat{b}_n(t)]^2 dt \right]^{1/2} \right). \tag{3.6}$$

The term U_{n4} is estimated again by a partial integration argument as follows,

$$\begin{aligned} U_{n4} &= h^{-1/2} \int_0^1 H_n(t) \hat{b}_n(t) h^{-1} q(t) \int_{-\infty}^{\infty} K'((t-x)/h) dW(x) dt \\ &\quad + h^{-1/2} \int_0^1 H_n(t) \hat{b}_n(t) Z_n(t) dq(t) \\ &\quad + h^{-1/2} \int_0^1 H_n(t) q(t) Z_n(t) d\hat{b}_n(t) \\ &\quad + H_n(t) V_n(t) \hat{b}_n(t)|_0^1 = T_{1n} + T_{2n} + T_{3n} + T_{4n}, \end{aligned}$$

where, as for the computations for U_{n2} , $H_n(t) = F_{X,n}(t) - F_X(t)$, and $Z_n(t) = \int_{-\infty}^{\infty} K((t-x)/h) dW(x)$. The last summand T_{4n} is obviously $O_p(n^{-1/2}) = o_p(n^{-1}h^{-1})$ by (A1).

The first term, T_{1n} , can be estimated as follows:

$$\begin{aligned} |T_{1n}| &\leq n^{-1/2} h^{-1} \sup_{t \in [0,1]} |n^{1/2} H_n(t)| \left[\int_0^1 [\hat{b}_n(t)]^2 dt \right]^{1/2} \\ &\quad \times S_2 \left[\int_0^1 \left[h^{-1/2} \int_{-\infty}^{\infty} K'((t-x)/h) dW(x) \right]^2 dt \right]^{1/2} \end{aligned}$$

Now, since $\int_0^1 [h^{-1/2} \int_{-\infty}^{\infty} K'((t-x)/h) dW(x)]^2 dt = O_p(1)$ and $n^{1/2} \sup_{t \in [0,1]} |H_n(t)| = O_p(1)$, we conclude that

$$T_{1n} = O_p \left(n^{-1/2} h^{-1} \left[\int_0^1 [\hat{b}_n(t)]^2 dt \right]^{1/2} \right).$$

The terms T_{2n} and T_{3n} are estimated in a similar fashion as we did estimate the terms of U_{n2} employing the Lipschitz continuity of $\hat{b}_n(t)$ and $q(t)$ and we thus obtain

$$\begin{aligned} T_{2n} &= O_p(n^{-1/2}) = o_p(n^{-1}h^{-1}), \\ T_{3n} &= O_p(n^{-1/2}) = o_p(n^{-1}h^{-1}). \end{aligned}$$

This shows finally that

$$U_{n4} = O_p \left(n^{-1/2} h^{-1} \left[\int_0^1 [\hat{b}_n(t)]^2 dt \right]^{1/2} \right) + o_p(n^{-1}h^{-1}). \tag{3.8}$$

It remains to show that

$$\int_0^1 [\hat{b}_n(t)]^2 d[H_n(t)] = O_p(n^{-1/2}) = o_p(n^{-1}h^{-1}). \tag{3.9}$$

Again by partial integration we have that the LHS of (3.9) is

$$-2 \int_0^1 H_n(t) \hat{b}_n(t) d\hat{b}_n(t) + H_n(t) \hat{b}_n^2(t)|_0^1.$$

As before the last summand is $O_p(n^{-1/2})$ and so is the first summand. Now, putting together (3.5) to (3.9) we finally have that

$$\begin{aligned} \hat{A}_n(h) &= \int_0^1 [\hat{b}_n(t)]^2 f_X(t) dt + n^{-1}h^{-1}\beta_k \int_0^1 S^2(t) dt \\ &+ o_p\left(n^{-1}h^{-1} + \int_0^1 [\hat{b}_n(t)]^2 dt\right) \end{aligned}$$

which proves the theorem.

Proof of Theorem 2. This proof goes mainly along the lines of the proof of Theorem 1. From Hardle and Marron [5, formula (2.4)], we have

$$m_n^*(t) - m(t) = Y_n^*(t) + b_n^*(t) + o_p\left(n^{-1/2}h^{-1/2} + \int_0^1 [b_n^*(t)]^2 dt\right) \quad (3.10)$$

where

$$b_n^*(t) = f_X^{-1}(t)h^{-1} \int_{-\infty}^{\infty} K((t-u)/h)[m(u) - m(t)]f_X(u) du$$

and

$$Y_n^*(t) = f_X^{-1}(t) h^{-1} \iint_{-\infty}^{\infty} [y - m(t)] K((t-x)/h) d[F_n(x, y) - F(x, y)].$$

This process can now be approximated as $\hat{Y}_n(t)$ (see the Appendix) but with $V^2(t) = S^2(t) - m^2(t)$ in the place of $S^2(t)$. So we obtain that

$$\begin{aligned} Y_{o,n}^*(t) &= [V^2(t)/f_X(t)]^{-1/2} Y_n^*(t) \\ &= n^{-1/2}h^{-1} \int_{-\infty}^{\infty} K((t-x)/h) dW(x) + o_p(n^{-1/2}h^{-1/2}) \end{aligned}$$

uniformly in t . The decomposition (3.10) then reads as

$$m_n^*(t) - m(t) = b_n^*(t) + n^{-1/2}h^{-1/2}V_n^*(t) + \rho_n^* \quad (3.11)$$

where

$$\rho_n^* = o_p\left(n^{-1/2}h^{-1/2} + \int_0^1 [bf(t)]^2 dt\right)$$

and

$$V_n^*(t) = [V^2(t)/f_X(t)]^{1/2} h^{-1/2} \int_{-\infty}^{\infty} K((t-x)/h) dW(x).$$

We then carry out the same procedures as for $V_n(t)$ in the proof of Theorem 1.

APPENDIX

It is shown here that the variance terms in (3.1) can be approximated by a sequence of Gaussian processes. The crucial step in these approximations is provided by the following lemma, due to Tusnady [20].

LEMMA 1. *Let $T(x, y) = (F_X, F_{Y|X})(x, y)$ be the Rosenblatt transformation [14]. Then on a suitable probability space there exists a sequence of Brownian bridges $B_n(x', y')$ on $[0, 1] \times [0, 1]$ such that*

$$\sup_{x,y} |[F_n(x, y) - F(x, y)] - n^{-1/2} B_n(T(x, y))| = O_p(n^{-1} [\log n]^2).$$

It is next shown that $\hat{Y}_n(t)$ can be approximated (uniformly in t) by Gaussian processes. For this define

$$Y_{0,n}(t) = [S^2(t)/f_X(t)]^{-1/2}$$

$$Y_{1,n}(t) = [S^2(t)f_X(t)]^{-1/2} h^{-1} \iint_{\Gamma_n} yK((t-x)/h) d[F_n(x, y) - F(x, y)]$$

where $\Gamma_n = \{|y| \leq a_n\}$,

$$Y_{2,n}(t) = [S_n^2(t)/S^2(t)]^{-1/2} Y_{1,n}(t)$$

where $S_n^2(t) = E[Y^2 I(|y| \leq a_n) | X = t]$,

$$Y_{3,n}(t) = [S_n^2(t)f_X(t)]^{-1/2} h^{-1} n^{-1/2} \iint_{\Gamma_n} yK((t-x)/h) dB_n(T(x, y))$$

where $\{B_n\}$ is the sequence of Brownian bridges as in Lemma 1.

$$Y_{4,n}(t) = [S_n^2(t)f_X(t)]^{-1/2} h^{-1} n^{-1/2} \iint_{\Gamma_n} yK((t-x)/h) dW_n(T(x, y))$$

where $\{W_n\}$ is a sequence of Wiener processes used in constructing $\{B_n\}$ as

$$B_n(x', y') = W_n(x', y') - x'y'W_n(1, 1) \quad [20]$$

$$Y_{5,n}(t) = [S_n^2(t)f_X(t)]^{-1/2} h^{-1}n^{-1/2} \times \int_{-\infty}^{\infty} [S_n^2(x)f_X(x)]^{1/2} K((t-x)/h) dW(x)$$

$$Y_{6,n}(t) = n^{-1/2}h^{-1} \int_{-\infty}^{\infty} K((t-x)/h) dW(x)$$

where $W(x)$ is a standard Wiener process on $(-\infty, \infty)$.

For the following lemmas $\|Y\|$ will denote $\sup_{0 \leq t \leq 1} |Y(t)|$.

LEMMA 2. $\|Y_{0,n} - Y_{1,n}\| = o_p(n^{-1/2}h^{-1/2})$.

Proof: We have to show that $\|U_n\| \rightarrow^p 0$, where

$$U_n(t) = n^{1/2}h^{-1/2} \iint_{|y| > a_n} yK((t-x)/h) d[F_n(x, y) - F(x, y)] = \sum_{i=1}^n X_{n,i}(t)$$

and

$$X_{n,i}(t) = (nh)^{-1/2} \{ Y_i K((t - X_i)/h) \cdot I(|Y_i| > a_n) - E[Y \cdot I(|Y| > a_n) K((t - X)/h)] \}.$$

Note that $EX_{n,i}(t) = 0$ for all t and that $X_{n,i}(\cdot)$ are independent, identically distributed for each n . Therefore

$$EX_{n,i}^2(t) \leq n^{-1}h^{-1} \sup |K|^2 \int_{|y| > a_n} y^2 f_Y(y) dy \quad (4.2)$$

establishes $U_n(t) \rightarrow^p 0$ for each t by assumption (A2). By (A4) and the Cauchy-Schwarz inequality we have

$$E |U_n(t) - U_n(t_1)| |U_n(t_2) - U_n(t)| \leq M_0 h^{-3} |t_1 - t| |t_2 + t| \int_{|y| > a_n} y^2 f_Y(y) dy,$$

establishing by (A2) tightness of $U_n(t)$ [2, Theorem 15.6]. ■

Note that the proof of this lemma was done as in Johnston's paper, but note also that our assumption is somewhat weaker than his, since we are employing Lemma 1, due to Tusnady [20], establishing a faster rate for the two-dimensional empirical process.

LEMMA 3. $\|Y_{1,n} - Y_{2,n}\| = o_p(n^{-1/2}h^{-1/2})$.

Proof Define $g(t) = S^2(t)f_X(t)$, $g_n(t) = S_n^2(t)f_X(t)$. We must show that

$$\begin{aligned} & \sup_{0 \leq t \leq 1} \left\{ |g(t)^{-1/2} - g_n(t)^{-1/2}| \right. \\ & \quad \cdot \left. \left| h^{-1} \iint_{\Gamma_n} yK((t-x)/h) d[F_n(x, y) - F(x, y)] \right| \right\} \\ & = o_p(n^{-1/2}h^{-1/2}). \end{aligned}$$

Now, from Johnston [8] we have that the second factor inside the curly brackets is $O_p(n^{-1/2}h^{-1/2})$ and from the mean value theorem

$$|g_n^{-1/2} - g^{-1/2}| = |g_n - g| \cdot \left| \frac{1}{2} \xi_n^{-3/2} \right|,$$

where ξ_n is between g_n and g . Since g_n, g are bounded away from zero by assumption (A3), $\|\xi_n^{-3/2}\|$ is a bounded sequence. Finally, from (A2) it follows that $\|g_n - g\| \rightarrow 0$ and thus the lemma follows.

LEMMA 4. $\|Y_{2,n} - Y_{3,n}\| = o_p(n^{-1/2}h^{-1/2})$.

Proof. Using integration by parts (see [8, Lemma A.5] for details), we obtain

$$\begin{aligned} & n^{1/2}h^{1/2} |g_n(t)|^{1/2} |Y_{2,n}(t) - Y_{3,n}(t)| \\ & = O_p(n^{-1/2}(\log n)^2) h^{-1/2} \left\{ 4a_n \int_{-A}^A |K'(u)| du + 4a_n[|K(A)| + |K(-A)|] \right\} \\ & = O_p(n^{-1/2}h^{-1/2}a_n(\log n)^2) \end{aligned}$$

uniformly in t . The proof thus follows using assumption (A2).

LEMMA 5. $\|Y_{3,n} - Y_{4,n}\| = o_p(n^{-1/2}h^{-1/2})$.

Proof. Since the Jacobian of the transformation T , introduced in Lemma 1, is $f(x, y)$, we have by Masani [11, Theorem 5.19],

$$\begin{aligned} & n^{1/2} |Y_{3,n}(t) - Y_{4,n}(t)| \\ & \leq |g_n(t)^{-1/2} h^{-1} \iint_{\Gamma_n} yK((t-x)/h) f(x, y) dx dy| \cdot |W_n(1, 1)|. \end{aligned}$$

So we finally have

$$n^{1/2} \| Y_{3,n} - Y_{4,n} \| \leq |W_n(1, 1)| \lambda_1 h^{-1} \int |K((t-x)/h)| dx$$

where λ_1 is a constant ($\lambda_1 = \sup_{0 \leq t \leq 1} |m(t)f_X(t)|$). This proves the lemma. Note that $Y_{4,n}(t)$ is a zero mean Gaussian process with covariance

$$\begin{aligned} & \text{cov}\{Y_{4,n}(t_1), Y_{4,n}(t_2)\} \\ &= [S_n^2(t_1)f_X(t_1)]^{-1/2} [S_n^2(t_2)f_X(t_2)]^{-1/2} \\ & \quad \times n^{-1}h^{-2} \iint_{\Gamma_n} y^2 K((t_1-x)/h) K((t_2-x)/h) f(x, y) dx dy \\ &= \text{cov}\{Y_{5,n}(t_1), Y_{5,n}(t_2)\}. \end{aligned}$$

So both $Y_{4,n}$ and $Y_{5,n}$ are Gaussian processes with the same covariance structure and can thus be identified.

LEMMA 6. $\| Y_{5,n} - Y_{6,n} \| = o_p(n^{-1/2}h^{-1/2})$.

Proof: Note that by assumption (A3) on $g_n(t) = S_n^2(t)f_X(t)$,

$$G_{n,t}(u) = [g_n(t)]^{-1/2} \{ [g_n(t-uh)]^{1/2} - [g_n(t)]^{1/2} \}$$

is also LC(α), $\alpha > \frac{1}{2}$, i.e.,

$$|G_{n,t}(u) - G_{n,t}(u')| \leq L_G h^\alpha |u - u'|^\alpha, \quad \alpha > \frac{1}{2},$$

where L_G is independent of t by (A3).

The difference of interest is now

$$\begin{aligned} & (nh)^{1/2} |Y_{5,n}(t) - Y_{6,n}(t)| \\ &= h^{-1/2} \left| \int \{ [g_n(x)/g_n(t)]^{1/2} - 1 \} K((t-x)/h) dW(x) \right| \\ &= |R_n(t)|. \end{aligned}$$

We will now show that $\sup_{t \in \mathcal{A}} |R_n(t)| = o_p(1)$. By partial integration we have for all n and t ,

$$\begin{aligned} |R_n(t)| &\leq \left| h^{-1/2} \int_{-A}^A W(t-uh) G_{n,t}(u) K'(u) du \right| \\ & \quad + \left| h^{-1/2} \int_{-A}^A [W(t-uh) - W(t)] K(u) d[G_{n,t}(u)] \right| \\ & \quad + \left| h^{-1/2} \int_{-A}^A W(t) G_{n,t}(u) K'(u) du \right| + O_p(h^{1/2}). \\ &= R_{1,n}(t) + R_{2,n}(t) + R_{3,n}(t) + R_{4,n}, \end{aligned}$$

where $R_{4,n}$ is independent of t . The term $R_{1,n}(t)$ is estimated as in Johnston [8, Lemma 4.6, p. 411] to obtain

$$\sup_{0 \leq t \leq 1} |R_{1,n}(t)| = o_p(1).$$

We now show that

$$\sup_{0 \leq t \leq 1} |R_{2,n}(t)| = o_p(1).$$

Let $w_0(s)$ denote the modulus of continuity of $W(t)$ and let $\bar{K} = \sup_{-A \leq u \leq A} |K(u)|$, we then have with Silberman [18, formula (7), (8), and his definitions of p, q, B],

$$\begin{aligned} |R_{2,n}(t)| &\leq h^{-1/2} \bar{K} \int w_0(|u|h) |dG_{n,t}(u)| \\ &\leq h^{-1/2} 16 \bar{K} 2^{1/2} \int_{-A}^A q(|u|h) dG_{n,t}(u) \\ &\quad + h^{-1/2} 16 \bar{K} (\log B)^{1/2} \int_{-A}^A p(|u|h) |dG_{n,t}(u)| \end{aligned}$$

Now following the proof of Silverman [18, Proposition 4] we see that the both summands are by assumption (A3) on $|dg_n(u)|$ of the order $o_p(1)$ uniformly in t . It remains to show that $\sup_{0 \leq t \leq 1} |R_{3,n}(t)| = o_p(1)$. This follows again from assumption (A3) on the LC(α), $\alpha > \frac{1}{2}$ condition $g_n(\cdot)$, and the following inequality:

$$\sup_{0 \leq t \leq 1} |R_{3,n}(t)| \leq \eta^{-2} \sup_{0 \leq t \leq 1} |W(t)| h^{-1/2} L_G h^\alpha \int_{-A}^A |u|^\alpha |K'(u)| du = o_p(1).$$

ACKNOWLEDGMENT

I am grateful to Steve Marron for helpful discussions. Ray Carroll contributed much to the approximations of the Appendix.

REFERENCES

- [1] BICKEL, P., AND ROSENBLATT, M. (1973). On some global measures of the deviation of density function estimators. *Ann. Statist.* 1 1071-1095.
- [2] BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

(1986) Härdle, W. Approximations to the Mean Squared Error with Applications to Optimal Bandwidth Selection for Nonparametric Regression Function Estimators

- [3] COLLOMB, G. (1981). Estimation non-parametrique de la regression: Revue bibliographique. *Internat. Statist. Rev.* **49** 75–93.
- [4] GASSER, T. AND MÜLLER, G. H. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation* (T. Gasser and M. Rosenblatt, Ed.), Lecture Notes in Mathematics Vol. 757, Springer-Verlag Heidelberg.
- [5] HARDLE, W., AND MARRON, S. (1983). *Optimal Bandwidth Selection in Nonparametric Regression Function Estimation*. Institute of Statistics Mimeo Series No. 1530, University of North Carolina, Chapel Hill.
- [6] HALL, P. (1982). Cross-validation in density estimation. *Biometrika* **69** 383–390.
- [7] ISSERLIS, L. (1918). On a formula for the product moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika* **12** 134–139.
- [8] JOHNSTON, G. (1982). Probabilities of maximal deviations of nonparametric regression function estimation. *J. Multivariate Anal.* **12** 402–414.
- [9] MACK, Y. P., AND SILVERMAN, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. Verw. Gebiete* **61** 405–415.
- [10] MARRON, J. S. (1986). Convergence properties of an empirical error criterion for multivariate density estimation. *J. Multivariate Anal.* **18**, No. 2.
- [11] MASANI, P. (1968). Orthogonally scattered measures. *Adv. in Math.* **2** 61–117.
- [12] NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.
- [13] RICE, T., AND ROSENBLATT, M. (1983). Smoothing splines: Regression, derivatives and deconvolution. *Ann. Statist.* **11** 141–156.
- [14] ROSENBLATT, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Statist.* **23** 470–472.
- [15] ROSENBLATT, M. (1969). Conditional probability density and regression estimation. In *Multivariate Analysis II* (P. R. Krishnaiah, Ed.), pp. 25–31. Academic Press, New York.
- [16] ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Stat.*, **42**, 1815–1842.
- [17] SCHUSTER, E. F. (1972). Joint asymptotic distribution of the estimated regression function at a finite number of district points. *Ann. Math. Stat.*, **43**, 84–88.
- [18] SILVERMAN, B. (1982). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Stat.*, **6**, 177–184.
- [19] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Stat.*, **10**, 1040–1053.
- [20] TUSNÁDY, G. (1977). A remark on the approximation of the sample distribution function in the multidimensional case. *Period. Math. Hung.*, **8**, 53–55.
- [21] WATSON, G. S. (1964). Smooth regression analysis. *Sankhya, Series A*, Vol. **26**, 359–372.
- [22] WEGMAN, E. J. (1972). Nonparametric probability density estimation: A comparison of density estimation methods. *J. Statist. Comput. Simulation*, **1**, 225–245.
- [23] WONG, W. H. (1983). On the consistency of cross-validation in kernel nonparametric regression. *Ann. Stat.*, to appear.

**STRONG UNIFORM CONVERGENCE RATES IN ROBUST
NONPARAMETRIC TIME SERIES ANALYSIS AND
PREDICTION: KERNEL REGRESSION ESTIMATION
FROM DEPENDENT OBSERVATIONS**

G rard COLLOMB*

*Universit  Paul Sabatier, Laboratoire de Statistique et Probabilit s, 118, route de Narbonne,
31062 Toulouse, France*

Wolfgang H RDLE**

Johann Wolfgang Goethe-Universit t, FB Mathematik, 6000 Frankfurt/M, FRG

Received 3 April 1985

Revised 10 February 1986

Let $\{Z_i; i \in \mathbb{N}\}$ be a strictly stationary real valued time series. We predict Z_{N+1} from $\{Z_1, \dots, Z_N\}$ by a robust nonparametric method. The predictor is defined by the kernel method and constructed as a functional M -estimate connected with the conditional law of Z_{p+1} on Z_1, \dots, Z_p , when $\{Z_i; i \in \mathbb{N}\}$ is Markovian of order p . Strong uniform convergence rates of this estimate are given together with some new results concerning robust regression kernel estimates from a sequence of $\mathbb{R}^p \times \mathbb{R}$ valued, identically distributed and ϕ -mixing random pairs $\{(X_i, Y_i); i = 1, \dots, n\}$. As a special case we obtain strong uniform convergence rates for estimators of the regression curve $E(Y_1|X_1 = \cdot)$ and of the density of the law of X_1 .

AMS 1980 Subject Classifications: Primary 62F15; Secondary 62G05.

robust time series analysis * robust prediction * robust nonparametric regression * M -estimation
* rate of convergence kernel estimate * nonparametric regression and density estimation

1. Introduction

Let $\{Z_i; i \in \mathbb{N}\}$ be a strictly stationary, real-valued process and let p be a positive integer. The autoregression function $r^*: \mathbb{R}^p \rightarrow \mathbb{R}$ is defined through

$$r^*(\cdot) = E(Z_{i+1} | (Z_{i-p+1}, \dots, Z_i) = \cdot), \quad i \geq p.$$

The Nadaraya-Watson method [16, 30] for estimating $r^*(\cdot)$ from $\{Z_i; i = 1, \dots, N\}$ has been studied by a number of authors. Watson [30] applied the kernel estimator

$$r_n^*(x) = \sum_{i=1}^n K((x - X_i)/h_n) Y_i / \sum_{i=1}^n K((x - X_i)/h_n), \quad (1.1)$$

* Unit  Associ e No. 745, Centre National de la Recherche Scientifique (France).

** Research partially supported by Deutsche Forschungsgemeinschaft Sonderforschungsbereich 123, "Stochastische Mathematische Modelle". Present address: Inst. Wirtschaftsth. II, Universit t Bonn, Adenauerallee 24-26, 5300 Bonn 1.

with bandwidth $h_n > 0$, kernel K and

$$n = N - p, \quad X_i = (Z_i, \dots, Z_{i+p-1}), \quad Y_i = Z_{i+p}, \quad i = 1, 2, \dots, n,$$

to some time series data. Pointwise asymptotic properties of the above kernel estimate have been investigated in [24, 2, 20, 21, 7, 31]. A recursive version of (1.1) was discussed in [17, 19]. Strong uniform convergence of r_n^* on a compact of \mathbb{R}^p was derived in [5, 6], leading to the strong convergence of the kernel predictor of Z_{n+1} from $\{Z_1, \dots, Z_n\}$ (when Z_1 is valued in a compact) in the following sense

$$r_n^*(Z_{n-p+1}, \dots, Z_n) - E(Z_{n+1} | Z_{n-p+1}, \dots, Z_n) \rightarrow_{n \rightarrow \infty} 0 \quad \text{w.p. 1.} \quad (1.2)$$

We here consider a more general nonparametric estimator $r_n(\cdot)$ which is implicitly defined as a zero with respect to (w.r.t.) t of

$$t \rightarrow \sum_{i=1}^n K((x - X_i)/h_n) \psi_x(Y_i - t) \quad (1.3)$$

where ψ_x is a bounded function for all x , satisfying some regularity conditions to be stated below. We denote by $r(x)$ a zero w.r.t. t of

$$t \rightarrow E(\psi_x(Y_1 - t) | X_1 = x).$$

In the special case of a process $\{Z_n; n \in \mathbb{N}\}$ which is markovian of order p , with $\psi_x = \psi, \forall x \in \mathbb{R}^p$, we can associate a loss function $\rho(u) = \int_{-\infty}^u \psi(s) ds$. The equality

$$E(\psi(Z_{N+1} - t) | Z_1, \dots, Z_N) = E(\psi(Z_{N+1} - t) | Z_{N-p+1}, \dots, Z_N)$$

then shows that the real random variable $r(Z_{N-p+1}, \dots, Z_N)$ is the best predictor of Z_{N+1} from $\{Z_1, \dots, Z_N\}$ with respect to the loss ρ .

We prove that r_n is uniformly convergent to r in some compact set and compute rates for this convergence under mild conditions on the process $\{Z_n; n \in \mathbb{N}\}$. The results will be stated in a more general setting for a process $\{(X_i, Y_i); i \in \mathbb{N}\}$, including the case of i.i.d. random pairs. The main application concerns the problem of prediction for a Markov process (considered after the statement of our Theorem 2) and leads to a result in the spirit of (1.2).

The estimator r_n enjoys some robustness properties. The Nadaraya-Watson estimate $r_n^*(\cdot)$ defined in (1.1) can be viewed as a least squares estimator, since $r_n^*(\cdot)$ is a solution to

$$\sum_{i=1}^n K((x - X_i)/h_n) (Y_i - t)^2 = \min_{t \in \mathbb{R}} \quad \text{if } K \geq 0.$$

Evidently, $r_n^*(x)$ is a weighted average of the $\{Y_i; i = 1, \dots, n\}$ and is therefore highly sensitive to occasionally occurring large fluctuations in the data which entails a high variation of the predictor r_n^* . The choice of a family of bounded functions ψ_x in (1.3) guarantees bounded influence and suggests more stable prediction properties. The unbounded influence function $\psi_x(u) \equiv u, x \in \mathbb{R}^p, u \in \mathbb{R}$ reproduces the classical Nadaraya-Watson estimator $r_n^*(x)$. Robustness properties of $r_n(x)$ in

the case of independent pairs $\{(X_i, Y_i); i = 1, \dots, n\}$ along with pointwise asymptotic properties are discussed in [28, 21, 10]. In the case of independent (X_i, Y_i) observations uniform convergence rates for $r_n^*(\cdot)$ were derived in [15, 29] and more recently, for $r_n(x)$, in [11]. It will be discussed below how our results apply to the case of independent (X_i, Y_i) .

2. Results

Let $\{(X_i, Y_i); i \in \mathbb{N}\}$ be a strictly stationary process valued in $(\mathbb{R}^p \times \mathbb{R}, \mathcal{B}_{\mathbb{R}} \otimes \mathcal{B}_{\mathbb{R}})$ and uniformly strongly mixing, i.e. (see [1])

there exists a sequence $\{\phi_i; i \in \mathbb{N}\}$ of positive numbers, tending to zero, such that for every integer $k > 0$, $|P(A \cap B) - P(A)P(B)| \leq \phi_k P(A)$ for all integers $n > 0$ and all $\sigma((X_1, Y_1), \dots, (X_n, Y_n))$ -measurable sets A and all $\sigma((X_{n+k}, Y_{n+k}), \dots)$ -measurable sets B .

The kernel $K : \mathbb{R}^p \rightarrow \mathbb{R}$ is a symmetric (i.e. $K(u) = K(-u)$, $u \in \mathbb{R}^p$) bounded function, satisfying

$$uK(u) \rightarrow 0 \text{ as } |u| \rightarrow \infty, \quad \int K(u) \, du = 1$$

and, in addition, is submitted to the following Lipschitz condition:

$$\exists \gamma > 0, c_k < \infty: |K(u) - K(v)| \leq c_k |u - v|^\gamma \quad \forall u, v \in \mathbb{R}^p. \tag{2.1}$$

The sequence $\{h_n; n \in \mathbb{N}\}$ is such that

$$h_n \rightarrow_{n \rightarrow \infty} 0, \quad nh_n^p \rightarrow_{n \rightarrow \infty} \infty, \quad h_n > 0, \quad \forall n \in \mathbb{N}. \tag{2.2}$$

The functions ψ_x are assumed to satisfy the following conditions, involving the density f of the marginal law of X and the regression function r (we set $X = X_1$ and $Y = Y_1$)

for all x , $\psi_x : \mathbb{R} \rightarrow \mathbb{R}$ is uniformly bounded, strictly monotone, continuously differentiable with

$$\left| \frac{\partial \psi_x(u)}{\partial u} \right| \leq c_\psi, \quad c_\psi \text{ independent of } x \text{ and } u, \tag{2.3}$$

and for all $x \in \mathbb{R}^p$, $r(x)$ is the unique zero with respect to of

$$t \rightarrow H(x, t) = E(\psi_x(y - t) | X = x) f(x), \tag{2.4}$$

the density of f being uniformly bounded on \mathbb{R}^p .

The strict monotony of ψ_x , for all x in \mathbb{R}^p , is assumed here to simplify the proofs. The proofs generalize to the case of functions ψ_x that are piecewise differentiable with monotonicity at the origin. Analogous arguments as in classical robust theory

would apply, but would introduce additional complications. The family of ψ -functions is indexed by x , in order to allow for general M -estimates. The situation that one has in mind is $\psi_x(\cdot) = \psi(\cdot/\sigma(x))$, where $\sigma(x)$ is a measure of spread for the conditional distribution of $(Y|X = x)$. It is also worth noting that the above condition on $\psi_x, x \in \mathbb{R}^p$, could be simplified by introduction of symmetry conditions (Huber [12, Chapter 4, p. 95], "Symmetry is an investistic assumption"). In this case $r = r^*$ and $r_n(x)$ provides a robust estimate of the conditional mean $r^*(x)$.

In the following condition it is assumed that there is an increasing sequence $\{m_n; n \in \mathbb{N}\}$ of positive integers such that

$$\exists A < \infty: \quad n\phi_{m_n}/m_n \leq A, \quad 1 \leq m_n \leq n, \quad \forall n \in \mathbb{N}. \tag{2.5}$$

We first present a uniform convergence result for the estimator $r_n, r_n(x)$ being defined for all x in \mathbb{R}^p as a zero with respect to t of the function (1.3). The existence and unicity of $r_n(x)$ are a consequence of the proof of the following theorem.

Theorem 1. *We suppose that the kernel K is positive, the density f is strictly positive on a compact C of \mathbb{R}^p and that the uniform equicontinuity condition $\forall \varepsilon > 0 \exists \alpha > 0$:*

$$\sup_{x \in C} \sup_{u: |u-x| \leq \alpha} |E(\psi_x(Y - r(x) - t | X = u)f(u) - H(x, r(x) + t))| \leq \varepsilon \tag{2.6}$$

is satisfied for all fixed t . If

$$nh_n^p / (m_n \text{Log } n) \rightarrow_{n \rightarrow \infty} \infty$$

then $r_n(x)$ exists and is unique w.p. 1. for all x in C and sufficiently large n , and we have

$$\sup_{x \in C} |r_n(x) - r(x)| \rightarrow_{n \rightarrow \infty} 0 \quad \text{w.p. 1.}$$

We now make precise the rate of this uniform convergence, and only assume the condition (2.2) for the sequence $\{h_n; n \in \mathbb{N}\}$.

Theorem 2. *Let C be a compact set in \mathbb{R}^p and G be a compact neighborhood of 0 in \mathbb{R} . We suppose that K is positive and that*

$$\inf_{t \in G} \inf_{x \in C} E(\psi'_x(Y - r(x) - t | X = x)f(x)) \geq C_0 > 0 \tag{2.7}$$

and

$$\sup_{t \in G} \sup_{x \in C} \sup_{u \in \mathbb{R}^p} \left| \frac{\partial^2 E(\psi_x(Y - r(x) - t | X = u)f(u))}{\partial^2 u} \right| \leq C_1 < \infty. \tag{2.8}$$

If the sequence $\{h_n; n \in \mathbb{N}\}$ is such that

$$\theta_n = (m_n \text{Log } n / (nh_n^p))^{1/2} \tag{2.9}$$

satisfies $\theta_n \rightarrow_{n \rightarrow \infty} 0$ and

$$\exists B > 0, B < \infty: \quad \theta_n^{-1} h_n^2 \leq B, \quad \forall n \in \mathbb{N}, \tag{2.10}$$

then we have

$$\theta_n^{-1} \sup_{x \in C} |r_n(x) - r(x)| = O(1) \quad \text{w.p. 1.}$$

In the following applications of Theorem 2 we discuss the choice of $\{m_n; n \in \mathbb{N}\}$ (see (2.5) for various applications).

Prediction for a Markov or a m-dependent process

The principal application of our result concerns the problem of time series analysis and prediction in the markovian case that we mentioned in the introduction. If the process $\{Z_n; n \in \mathbb{N}\}$ is markovian of order p , then the associated process $\{X_n = (Z_n, \dots, Z_{n+p-1}), Y_n = Z_{n+p}; n \in \mathbb{N}\}$ is also markovian (of order 1). If in addition Doeblin's condition (see [9, page 209], and also the L_p -norm condition in [23, page 206]) is fulfilled this markovian process is geometrically ϕ -mixing (i.e. $\exists \alpha \in]0, \infty[$ and $\exists \beta \in]0, 1[$: $\phi_m \leq \alpha \beta^m, m \in \mathbb{N}$) so that one can choose $m_n = c \text{Log } n (c > -1/\text{Log } \beta)$ in (2.5). This choice leads to the rate

$$\theta_n = \text{Log } n / (nh_n^p)^{1/2}$$

in (2.11), so that for such a Doeblin markovian process $\{Z_n; n \in \mathbb{N}\}$ the robust predictor $r_N(Z_{N-p+1}, \dots, Z_N)$ of Z_{N+1} satisfies

$$\theta_n^{-1} [r_N(Z_{N-p+1}, \dots, Z_N) - E(Z_{N+1} | Z_1, \dots, Z_N)] 1_{\{(Z_{N-p+1}, \dots, Z_N) \in C\}} = O(1)$$

w.p. 1.

If $\{Z_n; n \in \mathbb{N}\}$ is a m -dependent time series we can choose $m_n = 1 + m$ in (2.5) so that

$$\theta_n = (\text{Log } n / (nh_n^p))^{1/2}$$

is the rate of strong uniform consistency of r_n . It is interesting to note that the ϕ -mixing condition is rather restrictive when we consider Gaussian autoregressive processes: a stationary Gaussian process is ϕ -mixing if and only if it is m -dependent (see [13, Theorem 17.3.2]). It seems therefore reasonable to direct future research in the nonparametric analysis of time series towards weaker mixing conditions such as the strong mixing condition.

The case of independent $\{(X_i, Y_i)\}_{i=1}^n$

If we consider the problem of the robust estimation of r from a sequence of i.i.d. random pairs $(X_i, Y_i), i = 1, \dots, n$, we can see that our Theorem 2 extends the results of [11] who consider the robust estimate r_n but work under different assumptions on r . Theorem 2 also generalizes the results of [15] who obtain the rate (2.11) but only deal with the case $p = 1$ and the classical Nadaraya-Watson estimate r_n^* , which will be considered in our following Theorem 3. We note here that the undermentioned works involve proof techniques using strong approximations of the empirical process,

leading to the restriction (besides independence) $p = 1$ (see also [24] who give a limit law of a uniform norm associated with r_n^*). However let us note that [15] considered the case of a r.r.v. Y which is not necessarily bounded; an extension of our results to the case of unbounded Y is possible by a suitable truncation technique (see [25]). Our condition $nh_n^p N \text{Log } n \rightarrow_{n \rightarrow \infty} \infty$ cannot be improved upon since it can be shown [3, 4] that it is a necessary and sufficient condition for the uniform almost sure convergence (but also in probability) on a compact set of \mathbb{R}^p , see [8] for similar results on the pointwise convergence.

Lastly we give a theorem concerning the estimate r_n^* defined by (1.1) and the classical [18, 22] density estimate

$$f_n(x) = (nh_n^p)^{-1} \sum_{i=1}^n K((x - X_i)/h_n), \quad \forall x \in \mathbb{R}^p,$$

which also plays an important role in the analysis of the time series $\{Z_n; n \in \mathbb{N}\}$. These last results extend the results of [26] on f_n and of [15] on r_n^* to the case of ϕ -mixing random pairs $(X_i, Y_i), i = 1, \dots, n$ and to the case $p \geq 1$. A related result in density estimation is shown in [27].

Theorem 3. *If the sequence $\{h_n; n \in \mathbb{N}\}$ is such that*

$$\exists \lambda > 0, C < \infty: \theta_n^{-1} h_n^\lambda \leq C, \quad \forall n \in \mathbb{N}, \tag{2.12}$$

holds, then we have, for all compact C ,

$$\theta_n^{-1} \sup_{x \in C} |f_n(x) - Ef_n(x)| = O(1) \quad \text{w.p. 1} \tag{2.13}$$

and, if Y is bounded,

$$\theta_n^{-1} \sup_{x \in C} |r_n^* f_n(x) - Er_n^*(x) f_n(x)| = O(1) \quad \text{w.p. 1.} \tag{2.14}$$

If in addition the second derivative of f [resp of $r^(\cdot)f(\cdot)$] is uniformly bounded on an ε -neighborhood of C and the above assumption on h_n is satisfied for $\lambda = 2$, then*

$$\theta_n^{-1} \sup_{x \in C} |f_n(x) - f(x)| = O(1) \quad \text{w.p. 1,} \tag{2.15}$$

and, if f is bounded below on C by a strictly positive number,

$$\theta_n^{-1} \sup_{x \in C} |r_n^*(x) - r^*(x)| = O(1) \quad \text{w.p. 1.} \tag{2.16}$$

It is interesting to note that all convergence results we presented here are not holding “w.p. 1.” but in fact hold “almost complete” (see [5]).

3. Proof.

The proofs involve mainly an extension of the Bernstein inequality to ϕ -mixing real random variables (Lemma 1 of [5]), an argument using the Lipschitz condition

on K and methods introduced in [10] for the M -estimation of regression curves. Define

$$H_n(x, t) = \sum_{i=1}^n \alpha_i(x) \psi_x(Y_i - t) \tag{3.1}$$

where as a shorthand,

$$\alpha_i(x) = \alpha_{ni}(x) = (nh_n^p)^{-1} K((x - X_i)/h_n),$$

so that $r_n(x)$ satisfies

$$H_n(x, r_n(x)) = 0. \tag{3.2}$$

3.1. Preliminary lemmas

We shall use the following positive constants

$$\Gamma = \sup_{x \in \mathbb{R}^p} f(x), \quad \tilde{K} = \sup_{x \in \mathbb{R}^p} K(x) \quad \text{and} \quad \bar{K} = \int |K(u)| du$$

and we shall omit the index n of h_n from now on.

Lemma 1. *Let*

$$R_n(x, t) = \sum_{i=1}^n \alpha_i(x) \eta(Y_i, x, t), \quad \forall (x, t) \in \mathbb{R}^p \times \mathbb{R} \tag{3.3}$$

where η is a measurable function defined on $\mathbb{R} \times \mathbb{R}^p \times \mathbb{R}$ satisfying

$$\eta(Y, x, t) \leq \tilde{M} < \infty, \quad \forall (x, t) \in \mathbb{R}^p \times \mathbb{R}, \tag{3.4}$$

then there exist $B > 0, n_0 \in \mathbb{N}$:

$\forall \varepsilon \in (0, B) \forall n \in \mathbb{N}, n \geq n_0,$

$$\sup_{x \in C} \sup_{t \in \mathbb{R}} P\{|R_n(x, t) - ER_n(x, t)| \geq \varepsilon\} \leq a e^{-b\varepsilon^2 nh_n^p / m_n}, \tag{3.5}$$

where n_0, a and b are positive constants, which depend only on $\tilde{M}, \Gamma, \tilde{K}$ and the sequence $\{\phi_n; n \in \mathbb{N}\}$.

Proof. Write

$$R_n(x, t) - ER_n(x, t) = \sum_{i=1}^n \Delta_i$$

where

$$\Delta_i = \eta(Y_i, x, t) \alpha_i(x) - E[\eta(Y_i, x, t) \alpha_i(x)].$$

Define

$$d = n^{-1} h^{-p} 2\tilde{M}\tilde{K}, \quad \delta = n^{-1} 2\tilde{M}\tilde{K}\Gamma, \quad D = n^{-2} h^{-p} \tilde{M}^2 \tilde{K}\bar{K}\Gamma, \\ \beta = (8\tilde{M}\tilde{K})^{-1}, \quad B = 6\beta\tilde{M}^2 \tilde{K}\bar{K}\Gamma,$$

and note that, with (3.4),

$$|\eta(Y_i, x, t)\sigma_i(x)| \leq d/2,$$

$$E|\eta(Y_i, x, t)\alpha_i(x)| \leq (n^{-1}\tilde{M})^l(\tilde{K}h^{-p})^{l-1}\Gamma\tilde{K}, \quad l = 1, 2.$$

The choice $\alpha = \varepsilon\beta nh^p/m, m \in \{1, \dots, n\}$ satisfies condition (4.5) of Lemma 1 of [5] and this gives

$$P\{|R_n(x, t) - ER_n(x, t)| \geq \varepsilon\} \leq c_m e^{-t(\varepsilon, m)nh^p/m} \tag{3.6}$$

where, $\tilde{\phi}_m = \sum_{i=1}^m \phi_i, m \in \mathbb{N}, t(\varepsilon, m) = \varepsilon^2\beta[1 - B(m^{-1} + 16\phi_m/m)]$ and $c_m = 2e^{3\sqrt{\varepsilon n}\tilde{\phi}_m/m}$ do not depend on x or t . There exists m'_0 such that

$$B(1/m + 16\tilde{\phi}_m/m) \leq \frac{1}{2} \quad \text{for } m \geq m'_0.$$

Put $m = m'_n = \max\{m_n, m'_0\}$ in (3.6), then (3.5) follows with $a = 2e^{3A\sqrt{\varepsilon}}, A$ as in (2.5) and $b = (\beta/2) \inf_n\{m_n/m'_n\}$.

Lemma 2. *If the sequence $\{\theta_n; n \in \mathbb{N}\}$ defined by (2.9) satisfies (2.12), then there exists an $\varepsilon_0 > 0$ such that*

$$\sup_{t \in \mathbb{R}} \sum_{n=1}^{\infty} P\{\theta_n^{-1} \sup_{x \in C} |R_n(x, t) - ER_n(x, t)| > \varepsilon_0\} \leq D < \infty, \tag{3.7}$$

where D is a constant which depends only on $\tilde{M}, \Gamma, \varepsilon_0, K, C$ and $\{\phi_n; n \in \mathbb{N}\}$.

Proof. The proof follows closely Lemma 3 in [5], we therefore omit it.

Lemma 3. *Put $\eta(y, x, t) = \psi_x(y - r(x) - t)$ with ψ_x as in (2.3) and assume (2.12), then for any compact $G \subset \mathbb{R}$ there exists $\varepsilon_0 > 0$ such that*

$$\sum_{n=1}^{\infty} P\{\theta_n^{-1} \sup_{t \in G} \sup_{x \in C} |R_n(x, t) - ER_n(x, t)| \geq \varepsilon_0\} \leq D < \infty,$$

where D is a constant depending only on $\tilde{M}, \Gamma, \varepsilon_0, K, C, G$ and $\{\phi_n; n \in \mathbb{N}\}$.

Proof. We consider without loss of generality only the case $G = [-0.5, 0.5]$. Divide G into M disjoint subintervals, each of length M^{-1} , define $t_i = (i-1)/M + 1/(2M) - 0.5$ and put $U_n(x, t) = R_n(x, t) - ER_n(x, t)$. For each $t \in G$, if t_k denotes the nearest neighbor of t in $\{t_j; j = 1, \dots, M\}$, we have

$$|U_n(x, t)| = |U_n(x, t_k) + \tilde{U}_n(x, t)| \tag{3.8}$$

with

$$|U_n(x, t)| \leq |U_n(x, t_k)| + |\tilde{U}_n(x, t)|.$$

The condition (2.3) implies

$$|R_n(x, t) - R_n(x, t_k)| \leq C_\psi |t - t_k| \sum_{i=1}^n \alpha_i(x) \leq C_\psi \tilde{K} / (Mh^p)$$

because of the definitions of t_k and $\alpha_i(x)$, so that we have

$$\theta_n^{-1} \sup_{x \in C} \sup_{t \in G} |\tilde{U}_n(u, t)| \leq 2C_\psi \tilde{K} / (\theta_n M h^p)$$

and therefore from (3.8)

$$\theta_n^{-1} \sup_{x \in C} \sup_{t \in G} |U_n(x, t)| \leq 2C_\psi \tilde{K} / (\theta_n M h^p) + W_n \tag{3.9}$$

with

$$W_n = \theta_n^{-1} \max_{k=1, \dots, M} \sup_{x \in C} |U_n(x, t_k)|.$$

The trivial inequality

$$P(W_n > \varepsilon) \leq \sum_{k=1}^M P(\theta_n^{-1} \sup_{x \in C} |U_n(x, t_k)| > \varepsilon)$$

and an argument as in Lemma 3 of [5] shows that there is a constant β_1

$$P(W_n > \varepsilon) \leq \beta_1 M n^{-3}, \quad \forall n \in \mathbb{N}, n > n_1.$$

Now, if we choose $M = n$ we obtain

$$\sum_{n=1}^{\infty} P(W_n > \varepsilon) < \infty$$

and, from (3.9),

$$\theta_n^{-1} \sup_{x \in C} \sup_{t \in G} |U_n(x, t)| \leq 2C_\psi / (m_n \text{Log } n(nh^p)^{1/2}) + W_n$$

so that, since $nh^p \rightarrow_{n \rightarrow \infty} \infty$, the result (3.8) is proved.

Lemma 4. Under the assumptions of Theorem 1 we have, for all fixed real t ,

$$\sup_{x \in C} |EH_n(x, r(x) + t) - H(x, r(x) + t)| \rightarrow_{n \rightarrow \infty} 0$$

and under the assumptions of Theorem 2 we have

$$\theta_n^{-1} \sup_{x \in C} \sup_{t \in G} |EH_n(x, r(x) + t) - H(x, r(x) + t)| = O(1).$$

Proof. The equidistribution of the couples (X_i, Y_i) implies

$$EH_n(x, r(x) + t) = h_n^{-1} E(E^X \psi_x(Y - r(x) - t)) K((x - X)/h_n)$$

so that we have, since $\int K(u) \, du = 1$,

$$EH_n(x, r(x) + t) - H(x, r(x) + t) = h_n^{-1} \int \{E(\psi_x(Y - r(x) - t) | X = u) f(u) - H(x, r(x) + t)\} K((x - u)h_n^{-1}) \, du.$$

A slight modification of Bochner's Theorem used in [18] gives immediately the first part of the lemma from the condition (2.6).

A Taylor expansion of the function

$$u \rightarrow E(\psi_x(Y - r(x) - t) | X = u)f(u)$$

up to the order two, the symmetry of K (implying $\int uK(u) du = 0$) and the condition (2.8) give

$$|E(H_n(x, r(x) + t) - H(x, r(x) + t))| \leq h_n^2 \frac{C_1}{2} \int u^2 K(u) du$$

uniformly for x in C and t in G . The condition (2.10) implies immediately the second part of the lemma.

3.2. Proof of theorems

We first remark that $H_n(\cdot, \cdot)$ defined by (3.1) satisfies

$$H_n(x, r(x) + t) = R_n(x, t)$$

where $R_n(\cdot, \cdot)$ is defined by (3.3) for the choice of η given in the Lemma 3. The Lemma 2 and the first part of the Lemma 4 imply that under the condition of Theorem 1 we have

$$\sup_{x \in C} |H_n(x, r(x) + t) - H(x, r(x) + t)| \rightarrow_{n \rightarrow \infty} 0 \quad \text{w.p. 1.} \tag{3.10}$$

for all fixed real t .

The Lemma 3 and the second part of the Lemma 4 show that under the conditions of Theorem 2 (note that (2.10) implies (2.12), with $\lambda \geq 2$) show that

$$\theta_n^{-1} \sup_{x \in C} \sup_{t \in G} |H_n(x, r(x) + t) - H(x, r(x) + t)| = O(1) \quad \text{w.p. 1.} \tag{3.11}$$

Proof of Theorem 1. We use a classical approach for proving consistency of M -estimates (see [12]): this technique is extended here to the uniform consistency case. Fix $\varepsilon > 0$. The strict monotony of ψ_x and positivity of f on C imply

$$\forall x \in C \quad H(x, r(x) + \varepsilon) < 0 < H(x, r(x) - \varepsilon).$$

The result (3.10) entails, for all sufficiently large n ,

$$\forall x \in C \quad H_n(x, r(x) + \varepsilon) < 0 < H_n(x, r(x) - \varepsilon) \quad \text{w.p. 1}$$

and therefore

$$\forall x \in C \quad r(x) - \varepsilon < r_n(x) < r(x) + \varepsilon \quad \text{w.p. 1}$$

because of (3.2) and the positivity of K . This last result can also be written as in the conclusion of Theorem 1.

Lastly we show the existence and the unicity of $r_n(x)$ defined by (3.2): the positivity of K and the strict monotony of ψ_x imply the unicity of $r_n(x)$ when

$$\exists t_0 \in \mathbb{R}: H_n(x, r(x) + t_0) \neq 0;$$

then, since $r(x)$ is supposed to be the unique zero with respect to t of (2.4)

$$\exists t_0 \in \mathbb{R}: H(x, r(x) + t_0) \neq 0,$$

so that the result of (3.10) implies that the above condition on H_n is satisfied w.p. 1. for all x in C and sufficiently large n .

Proof of Theorem 2. The definitions (2.4) of r and (3.2) of r_n show that for all $x \in \mathbb{R}^p$ we have

$$H(x, r(x)) = H(x, r_n(x)) + H(x, r_n(x)) - H_n(x, r_n(x)) = 0$$

so that a Taylor expansion of $H(x, \cdot)$ leads to

$$(r_n(x) - r(x))E(\psi'(Y - \xi_n(x)) | X = x)f(x) = H_n(x, r_n(x)) - H(x, r_n(x)) \tag{3.12}$$

where $\xi_n(x)$ is between $r(x)$ and $r_n(x)$. The result of the Theorem 1 shows that for a sufficiently large n_0

$$\sup_{x \in C} |r_n(x) - r(x)| \in G \text{ w.p. 1., } \forall n \geq n_0,$$

so that we have w.p. 1. for such integers n

$$\inf_{x \in C} E|(\psi'_x(Y_1 - \xi_n(x)) | X_1 = x)f(x)| > 0 \text{ w.p. 1.}$$

because of (2.7) and

$$\begin{aligned} &\sup_{x \in C} |H_n(x, r_n(x)) - H(x, r_n(x))| \\ &\leq \sup_{x \in C} \sup_{t \in G} |H_n(x, r(x) + t) - H(x, r(x) + t)| \text{ w.p. 1.} \end{aligned}$$

The formulas (3.11) and (3.12) immediately give the result of the second theorem.

Proof of Theorem 3. The result (2.13), resp. (2.14), follows immediately from the Lemma 2 applied to the case

$$\eta(\cdot, \cdot, \cdot) = 1, \text{ resp. } \eta(Y_i, \cdot, \cdot) = Y_i, i \in \mathbb{N}.$$

The other results are obtained by Taylor expansion, e.g. [18] giving

$$\sup_{x \in C} |ER_n(x) - R(x)| = O(h_n^2) \text{ for } R_S = f \text{ and } R_S = r^*f, \text{ with } S = "n" \text{ or } "n",$$

with, for (2.16), the inequality (the argument x is omitted)

$$|r_n^* - r^*| \leq \{ |r_n^*f_n - Er_n^*f_n| - |r^*|f_n - Ef_n| + |r_n^*f_n - r^*Ef_n| \} / f_n$$

leading, because of the last assumption on f , to a convenient majorization of $\sup_{x \in C} |r_n^*(x) - r^*(x)|$.

Acknowledgement

We would like to thank the referee for various suggestions leading to substantial improvements of presentation.

References

- [1] P. Billingsley, *Convergence of Probability Measures* (Wiley, New York, 1967).
- [2] D. Bosq, Sur la prédiction non paramétrique des variables aléatoires et de mesures aléatoires, *Z. Wahr. Verw. Geb.* 64 (1983) 541-553.
- [3] G. Collomb, *Estimation non paramétrique de la régression par la méthode du noyau*, Thèse, Université Paul Sabatier, Toulouse, 1976.
- [4] G. Collomb, Condition nécessaires et suffisantes de convergence uniforme d'un estimateur de la régression, estimation des dérivées de la régression, *C.R.A.S. Paris Série A* 288 (1979) 161-164.
- [5] G. Collomb, Propriétés de convergence presque complète du prédicteur à noyau, *Z. Wahr. Verw. Geb.* 66 (1984) 441-460.
- [6] G. Collomb, Nonparametric time series analysis and prediction: uniform almost sure convergence of the window and $k - N$, N , autoregression estimates, *Math. Oper. and Stat. Ser. Statistics* 16 (1985) 297-307.
- [7] G. Collomb and P. Doukhan, Estimation non paramétrique de la fonction d'autorégression d'un processus stationnaire et ϕ -mélangeant: risques quadratiques pour la méthode de noyau, *C.R.A.S. Paris Série I* (296 (1983) 1983) 859-862.
- [8] L. Devroye, On the almost everywhere convergence of nonparametric regression function estimates, *Annals of Statistics* 9 (1981) 1310-1319.
- [9] J. Doob, *Stochastic Processes* (Wiley, New York, 1953).
- [10] W. Härdle, Robust regression function estimation, *J. Mult. Analysis* 14 (1984) 169-180.
- [11] W. Härdle and S. Luckhaus, Uniform consistency of a class of regression estimators, *Ann. Stat.* 12 (1984) 612-623.
- [12] P.J. Huber, *Robust Statistics* (Wiley, New York, 1981).
- [13] I.A. Ibragimov and Y.V. Linnik, *Independent and Stationary Sequences of Random Variables* (Wolters-Noordhoff, Groningen, 1971).
- [14] H. Liero, On the maximal deviation of the kernel regression function estimate. *Math. Oper. and Stat. Ser. Statistics* 13 (1982) 171-182.
- [15] Y.P. Mack and B.W. Silverman, Weak and Strong Uniform Consistency of Kernel Regression Estimates. *Z. Wahr. Verw. Geb.* 61 (1982) 405-415.
- [16] E.A. Nadaraya, On estimating regression, *Theory of Probability and its Applications* 9 (1964) 141-142.
- [17] H.T. Nguyen and D.T. Pham, Nonparametric estimation in diffusion model by discrete sampling, *Publications de l'Institut de Statistique de l'Université de Paris XX VI* 2 (1981) 89-109.
- [18] E. Parzen, On estimation of a probability density function, *Ann. Math. Stat.* 31 (1962) 1065-1079.
- [19] D.T. Pham, Nonparametric estimation of the drift coefficient in the diffusion equation, *Math. Operationsforsch. Statist.* 12(1) (1981) 61-74.
- [20] P. Robinson, Nonparametric estimators for time series, *J. Time Series Anal.* 4(3) (1983) 185-207.
- [21] P. Robinson, Robust nonparametric autoregression, in: J. Franke, W. Härdle, D. Martin, eds., *Lecture Notes in Statistics* (Springer Verlag, Berlin).
- [22] M. Rosenblatt, Remarks on some nonparametric estimates of a density function, *Annals of Mathematical Statistics* 27 (1956) 832-837.

- [23] M. Rosenblatt, Markov Processes. Structure and Asymptotic Behavior (Springer, Berlin).
- [24] G. Roussas, Nonparametric estimation of the transition distribution function of a Markov process, *Annals of Mathematical Statistics* 40 (1969) 1386-1400.
- [25] P. Sarda and P. Vieu, Estimation non paramétrique de la régression pour des variables dépendantes, application à la prédiction par un processus markovian, *Manuscript*, 1985.
- [26] R.J. Serfling, Property and applications of metrics on nonparametric density estimators, *Colloquia Mathematica Societatis Janos Bolyai, Budapest (Hungary)* 32 (1980) 859-873.
- [27] W. Stute, A law of the logarithm for kernel density estimators, *Annals of Probability* 10 (1982) 414-422.
- [28] A.B. Tsybakov, Robust estimates of a function, *Problems of Inf. Theory*, 18 (1983) 39-52.
- [29] H. Wandl, On kernel estimation of regression functions, *Wiss. Sit. z. Stoch. WSS* 03 (1980) 1-25.
- [30] G.S. Watson, Smooth regression analysis, *Sankhya ser. A* 26 (1964) 359-372.
- [31] S.J. Yakowitz, Nonparametric estimation of Markov transition functions, *Annals of Statistics* 7 (1979) 671-679.

Random Approximations to Some Measures of Accuracy in Nonparametric Curve Estimation

JAMES STEPHEN MARRON*

*Department of Statistics, University of North Carolina,
Chapel Hill, North Carolina 27514*

AND

WOLFGANG HÄRDLE†

Universität Heidelberg, Sonderforschungsbereich 123, Heidelberg, West Germany

Communicated by M. Rosenblatt

This paper deals with a quite general nonparametric statistical curve estimation setting. Special cases include estimation of probability density functions, regression functions, and hazard functions. The class of "fractional delta sequence estimators" is defined and treated here. This class includes the familiar kernel, orthogonal series, and histogram methods. It is seen that, under some mild assumptions, both the *average square error* and *integrated square error* provide reasonable (random) approximations to the *mean integrated square error*. This is important for two reasons. First, it provides theoretical backing to a practice that has been employed in several simulation studies. Second, it provides a vital tool for proving theorems about selecting smoothing parameters for several different nonparametric curve estimators. © 1986 Academic Press, Inc.

1. INTRODUCTION

Let X, X_1, \dots, X_n be a random sample of d -dimensional random vectors having density function $f(x)$ and cumulative distribution function $F(x)$.

Received December 6, 1984; revised December 15, 1985.

AMS 1980 subject classifications: Primary 62H12, 62G05.

Key words and phrases: Hazard functions, mean integrated square error, nonparametric estimation, regression function.

* Research partially supported by Office of Naval Research, Contract N00014-75-C-0809, and the Deutsche Forschungsgemeinschaft.

† Research partially supported by the Deutsche Forschungsgemeinschaft, SFB 123, "Stochastische Mathematische Modelle," and Air Force Office of Scientific Research Contract AFOSR-F49620 82 C009

Suppose we are interested in a certain functional $g(x)$, $x \in \mathbb{R}^d$ of the distribution of X . The problem of estimating the curve $g(x)$ from the random sample is called *nonparametric curve estimation*.

Some special cases of nonparametric curve estimation are:

D—*density estimation*: where g is taken to be f .

H—*hazard function estimation*: where g is given by

$$g(x) = \frac{f(x)}{1 - F(x)}.$$

R—*Regression estimation*: where g is the regression curve of Y on Z' ,

$$g(x) = g(z) = E[Y | Z = z],$$

using the notation

$$d' = d - 1,$$

$$z = (z^{(1)}, \dots, z^{(d')}), \tag{1.1}$$

$$x = (z^{(1)}, \dots, z^{(d')}, y),$$

$$X = (Z^{(1)}, \dots, Z^{(d')}, Y).$$

This list of examples is meant to be representative, not exhaustive. See Prakasa-Rao [26] for other possibilities.

Quite a number of different estimators have been proposed for each of the curves given above. For comparison of these estimators, several measures of accuracy have been considered. A very common measure is the mean integrated square error,

$$\text{MISE} = E \int [\hat{g}(x) - g(x)]^2 w(x) dF(x),$$

with some nonnegative weight function $w(x)$ (depending only on z in the regression setting).

While MISE is theoretically pleasing as a distance between \hat{g} and g , it is often hard to compute. The literature contains two different ways of overcoming this difficulty. The first is to study the asymptotic (as $n \rightarrow \infty$) behavior of MISE. The second is to consider Monte Carlo (and hence random) approximations to MISE. In this paper it is seen that, for many estimators, these two approaches give quite similar results for large values of n .

Stochastic (i.e., random) distances that have been considered include the integrated square error (ISE) given by

$$\text{ISE} = \int [\hat{g}(x) - g(x)]^2 w(x) dF(x),$$

and the average square error (ASE) given by

$$\text{ASE} = n^{-1} \sum_{i=1}^n [\hat{g}(X_i) - g(X_i)]^2 w(X_i).$$

Wegman [48] argued in the setting of density estimation that, for n large, ASE should be a good approximation of MISE.

He then used ASE as a distance measure for a Monte Carlo comparison of several density estimators. ASE has also been employed for this purpose by Fryer [11] and Wahba [42]. Breiman, Meisel, and Purcell [5] and Raatgever and Duin [27] used a "normalized version" of ASE in their Monte Carlo studies. The distance ISE also has been attractive to several authors, see, for example, Rust and Tsokos [32], Scott and Factor [33], Bean and Tsokos [1], and Bowman [4]. In the regression setting, Stone [36] has used a "leave-one-out" version of ASE and Engle, Granger, Rice, and Weiss [9] and Silverman [34] have used ASE to study cross-validated estimators. In the hazard function setting, Tanner and Wong [40] have compared two estimators by computing the difference of their ASEs.

The use of ASE and ISE as measures of accuracy was criticized by Steele [35], who gave an example in which, asymptotically as $n \rightarrow \infty$, ASE behaved very differently from ISE (hence, at least one is a poor approximation to MISE). In reply to this objection, Hall [13] showed that Steele's example was somewhat artificial by showing that, in the case $d=1$, if $\hat{g}(x)$ is a kernel density estimator, then under some reasonable assumptions, as $n \rightarrow \infty$,

$$\text{ASE} = \text{MISE} + o_p(\text{MISE}), \quad (1.2)$$

$$\text{ISE} = \text{MISE} + o_p(\text{MISE}), \quad (1.3)$$

and if $\hat{g}(x)$ is a trigonometric series density estimator (1.3) holds.

The object of this paper is twofold. First, Hall's results are extended to a wider class of estimators and to a variety of nonparametric curve estimation settings. This demonstrates that the objections of Steele [35] need cause no concern in the case of many commonly considered estimators. Second, the results of this paper provide an important tool for use in analyzing curve estimators with data-based smoothing parameter selection. In particular, asymptotic optimality results can be derived from suitable uniform versions of (1.2) and (1.3). Special cases of this may be seen, either explicitly or implicitly, in the results of Hall [14], Stone [37, 38], Burman [2], and Marron [22, 23] in the density estimation setting, and in the results of Rice [28], Härdle and Marron [19, 20], and Burman and Chen [3] in the regression setting.

Section 2 introduces the class of "fractional delta sequence estimators" and makes evident that many of the most widely studied nonparametric

estimators are contained in this class. Section 3 contains theorems which give sufficient conditions for (1.2) and (1.3) for a subset of these estimators. Section 4 contains theorems which extend the results of Sections 3 to all fractional delta sequence estimators. Section 5 contains examples for illustration of these theorems. The proofs of the theorems are in Section 6.

2. FRACTIONAL DELTA SEQUENCE ESTIMATORS

The class of *fractional delta sequence estimators* is defined to consist of all estimators of the form

$$\hat{g}_\lambda(x) = \frac{\sum_{i=1}^n \delta_\lambda(x, X_i)}{\sum_{i=1}^n \delta'_\lambda(x, X_i)}, \quad (2.1)$$

where δ_λ and δ'_λ are measurable functions on $\mathbb{R}^d \times \mathbb{R}^d$, which are indexed by a "smoothing parameter" $\lambda = \lambda(n) \in \mathbb{R}^+$. The special case $\delta'_\lambda(x, X_i) \equiv 1$ gives the delta sequence estimators studied by Watson and Leadbetter [47], Földes and Révész [10], and Walter and Blum [44], among others.

In the setting of density estimation, some well-known estimators of this type are:

D-1. *Kernel estimators.* Introduced by Rosenblatt [29] and Parzen [25], given a "kernel function," $K: \mathbb{R}^d \rightarrow \mathbb{R}$, and the smoothing parameter, $\lambda \in \mathbb{R}^+$, define

$$\begin{aligned} \delta_\lambda(x, X_i) &= \lambda K(\lambda^{1/d}(x - X_i)), \\ \delta'_\lambda(x, X_i) &\equiv 1. \end{aligned} \quad (2.2)$$

D-2. *Histogram estimators.* Write $\mathbb{R}^d = \bigcup_{l=1}^\infty A_l$, where the "bins" A_l are disjoint with Lebesgue measure λ^{-1} (where λ is the smoothing parameter). For $l = 1, 2, \dots$ let $1_l(x)$ denote the indicator of A_l . Define

$$\begin{aligned} \delta_\lambda(x, X_i) &= \sum_{l=1}^\infty \lambda 1_l(x) 1_l(X_i), \\ \delta'_\lambda(x, X_i) &\equiv 1. \end{aligned} \quad (2.3)$$

The extension to unequal bin sizes is straightforward, but requires more notation.

D-3. *Orthogonal series estimators.* Introduced by Cencov [6]. Suppose $\{\psi_l(x)\}$ is a sequence of functions which is orthonormal and complete with respect to the inner product

$$\int \psi_l(x) \psi_r(x) w(x) dF(x). \quad (2.4)$$

Given the smoothing parameter $\lambda \in \mathbb{Z}^+$, define

$$\begin{aligned}\delta_\lambda(x, X_i) &= \sum_{l=1}^{\lambda} \psi_l(x) \psi_l(X_i) w(X_i), \\ \delta'_\lambda(x, X_i) &\equiv 1.\end{aligned}\tag{2.5}$$

Further examples of delta sequence density estimators may be found in Walter and Blum [44] and Susarla and Walter [39]. Some examples of fractional delta sequence estimators in the regression setting are:

R-1. *Kernel estimators.* Introduced by Nadaraya [24] and Watson [45]. Given a kernel function, $K(x')$ and a smoothing parameter, λ , using the notation (1.1), define

$$\begin{aligned}\delta_\lambda(x, X_i) &= \lambda K(\lambda^{1/d'}(z - Z_i)) Y_i \\ \delta'_\lambda(x, X_i) &= \lambda K(\lambda^{1/d'}(z - Z_i)).\end{aligned}$$

Note that, $\hat{g}(x)$ is a weighted average of the Y_i .

R-2. *Known-marginal kernel estimators.* Studied by Johnston [21]. Let $f_M(z)$ denote the marginal density of Z_i and define

$$\begin{aligned}\delta_\lambda(x, X_i) &= \lambda K(\lambda^{1/d'}(z - Z_i)) Y_i \\ \delta'_\lambda(x, X_i) &= f_M(z).\end{aligned}$$

To see the idea behind this estimator, note that when the denominator of R-1 is properly normalized, it becomes the estimate D-1 of the marginal density, $f_M(z)$.

R-3. *Delta sequence estimators.* A generalization of R-1, discussed in Collomb [7]; define $\tilde{\delta}_\lambda(z, Z_i)$ as for any of the density estimators and let

$$\begin{aligned}\delta_\lambda(x, X_i) &= \tilde{\delta}_\lambda(z, Z_i) Y_i, \\ \delta'_\lambda(x, X_i) &= \tilde{\delta}_\lambda(z, Z_i).\end{aligned}$$

Note that the regressogram of Tukey [41] is a special case where $\tilde{\delta}_\lambda$ is defined as for D-2.

In the setting of hazard function estimation, Watson and Leadbetter [46] have introduced the following fractional delta sequence estimators:

H-1. *Kernel estimators.* Given a kernel function, $K(x)$, and a smoothing parameter, λ , define

$$\begin{aligned}\delta_\lambda(x, X_i) &= \lambda K(\lambda(x - X_i)), \\ \delta'_\lambda(x, X_i) &= 1 - \int_{-\infty}^x \lambda K(\lambda(t - X_i)) dt.\end{aligned}\tag{2.6}$$

H-2. *Delta sequence estimators.* A straightforward generalization of H-1; define $\delta_\lambda(x, X_i)$ as in any of the density estimators and let

$$\delta'_\lambda(x, X_i) = 1 - \int_{-\infty}^x \delta_\lambda(t, X_i) dt.$$

3. APPROXIMATION THEOREMS FOR DELTA SEQUENCE ESTIMATORS

This section gives sufficient conditions for (1.2) and (1.3) in the special case of delta sequence estimators, which are of the form

$$\hat{g}_\lambda(x) = n^{-1} \sum_{i=1}^n \delta_\lambda(x, X_i) = \int \delta_\lambda(x, x_1) dF_n(x_1). \quad (3.1)$$

Assume that λ ranges over a finite set A_n , whose cardinality is bounded by

$$\#(A_n) \leq \mathcal{C}n^\rho, \quad \rho > 0 \quad (3.2)$$

(i.e., is increasing at most algebraically fast). For estimators with a continuous smoothing parameter, such as the kernel estimators, the result of this paper can be easily extended to A_n an interval, by a continuity argument (compare Marron [22] and Härdle and Marron [19]).

For ease of presentation, it will be assumed that there are constants \mathcal{C} and $\varepsilon > 0$ so that, for each n , and for all $\lambda \in A_n$,

$$\mathcal{C}^{-1}n^\varepsilon \leq \lambda \leq \mathcal{C}n^{1-\varepsilon}. \quad (3.3)$$

The next assumptions are rather technical in nature, but are stated in this form because these are the common properties which make all of the diverse estimators of Section 2 satisfy (1.2) and (1.3). Implicit in these assumptions are conditions on w and f , e.g., boundedness of f or integrability of $w \cdot f$. Precise conditions (on w and f) depend on which estimator is being considered. These conditions are given in Section 5, where it is seen that quite different methods of verification of these assumptions are needed for different estimators. The assumption (3.4) represents the most important property of delta sequence estimators. Intuition can be gained by considering the kernel density estimation case and performing integration by substitution.

For $k = 1, 2, \dots$ assume there is a constant \mathcal{C}_k so that for any $m = 2, \dots, 2k$ and $\lambda \geq 1$,

$$\left| \int \cdots \int \left[\prod_{i,i'=1}^m \delta_\lambda(x_i, x_{i'})^{\alpha_{ii'}} \right] \times \left[\prod_{i=1}^m w(x_i)^{\beta_i} \right] dF(x_1) \cdots dF(x_m) \right| \leq \mathcal{C}_k \lambda^{k-m/2}, \quad (3.4)$$

where $\alpha_{ii'} = 0, \dots, k$ subject to

$$\sum_{i,i'=1}^m \alpha_{ii'} = k$$

and the restriction that for each $i = 1, \dots, m$, there is an $i' \neq i$ so that either $\alpha_{ii'}$ or $\alpha_{i'i}$ is nonzero, and where $\beta_i = 0, 1$ with $\beta_i = 1$ any time an $\alpha_{ii'} \geq 1$ (with $w(x_i)^{\beta_i}$ taken to be 1 when $w(x_i) = \beta_i = 0$).

Assume that the quantity

$$\delta_\lambda(x_1, x_2) = \int \delta_\lambda(x_3, x_1) \delta_\lambda(x_3, x_2) w(x_3) dF(x_3) \quad (3.5)$$

satisfies the assumption (3.4), with each $\beta_i = 0$, and that there is a constant \mathcal{C} so that

$$\iint \delta_\lambda(x_1, x_2) dF(x_1) dF(x_2) \leq \mathcal{C}. \quad (3.6)$$

Assume there is a constant \mathcal{C} so that

$$\int \delta_\lambda(x, x) dF(x) \geq \mathcal{C}\lambda. \quad (3.7)$$

Another assumption is that there is a constant $\xi > 0$, so that for $k = 1, 2, \dots$ there is a constant \mathcal{C}_k such that

$$\int B(x)^{2k} w(x) dF(x) \leq \mathcal{C}_k b(\lambda) \lambda^{(k-1)(1-\xi)}, \quad (3.8)$$

where $B(x)$ denotes the bias and $b(\lambda)$ denotes the integrated squared bias of the estimator \hat{g} given by

$$B(x) = E[\hat{g}(x)] - g(x) = \int \delta_\lambda(x, x_2) dF(x_2) - g(x), \quad (3.9)$$

$$b(\lambda) = \int B(x)^2 w(x) dF(x).$$

Finally assume that for $k = 1, 2, \dots$ there is a constant \mathcal{C}_k so that

$$\int [\delta_\lambda(x, x)]^{2k} w(x) dF(x) \leq \mathcal{C}_k \lambda^{2k}. \quad (3.10)$$

THEOREM 1. Under the assumptions (3.1)–(3.7),

$$\lim_{n \rightarrow \infty} \sup_{\lambda \in \mathcal{A}_n} \left| \frac{\text{ISE}(\lambda) - \text{MISE}(\lambda)}{\text{MISE}(\lambda)} \right| = 0 \quad \text{a.s.}$$

THEOREM 2. *Under the assumptions (3.1)–(3.10), and w bounded,*

$$\lim_{n \rightarrow \infty} \sup_{\lambda \in \mathcal{A}_n} \left| \frac{\text{ASE}(\lambda) - \text{MISE}(\lambda)}{\text{MISE}(\lambda)} \right| = 0 \quad \text{a.s.}$$

Remark 1. We believe that the proofs of these approximations can be extended to the case of λ , a vector, or even a matrix, but additional messy notation and tedious work are required for this.

Remark 2. In this case of kernel density estimation, under stronger conditions than those given here, the strong law of large numbers in Theorem 1 has been extended to a central limit theorem by Hall [15].

Remark 3. The supremum over λ is essential for analyzing curve estimators with a data-dependent smoothing parameter. Such estimators are of the form

$$\hat{g}_L(x) = n^{-1} \sum_{i=1}^n \delta_L(x, X_i),$$

where $L = L(X_1, \dots, X_n)$. Note that as long as $L \in \mathcal{A}$ a.s., we immediately have, under the above assumptions,

$$\lim_{n \rightarrow \infty} \left| \frac{\text{ISE}(L) - \text{MISE}(L)}{\text{MISE}(L)} \right| \rightarrow 0 \quad \text{a.s.}$$

and similarly for ASE.

4. APPROXIMATION THEOREMS FOR FRACTIONAL DELTA SEQUENCE ESTIMATORS

This section extends Theorems 1 and 2 to include fractional delta sequence estimators. Since these estimators have denominators containing random variables, they are technically more difficult to work with. In fact, for the estimator R-1, if the kernel function, K , is allowed to take on negative values, then the moments of $\hat{g}(x)$ may not exist (see Rosenblatt [30] and Härdle and Marron [18]) so MISE is not a reasonable distance. These difficulties are overcome using the same method as that employed in Chapter 6 of Cochran [8] for the study of ratio estimators. Assume there is a function $D(x)$ and a set $S \subset \mathbb{R}^d$ so that, uniformly over $x \in S$, $\lambda \in \mathcal{A}_n$,

$$n^{-1} \sum_i \delta'_\lambda(x, X_i) \rightarrow D(x) \quad \text{a.s.} \quad (4.1)$$

and assume that

$$\inf_{x \in S} D(x) > 0. \quad (4.2)$$

Then, uniformly over $x \in S$, $\lambda \in \Lambda_n$,

$$\begin{aligned} \hat{g}(x) - g(x) &= n^{-1} \sum_i \left[\frac{\delta_\lambda(x, X_i) - \delta'_\lambda(x, X_i) g(x)}{D(x)} \right] \\ &+ \frac{[D(x) - n^{-1} \sum_i \delta'_\lambda(x, X_i)] n^{-1} \sum_i [\delta_\lambda(x, X_i) - \delta'_\lambda(x, X_i) g(x)]}{D(x) n^{-1} \sum_i \delta'_\lambda(x, X_i)} \\ &= n^{-1} \sum_i \delta_\lambda^*(x, X_i) + o\left(n^{-1} \sum_i \delta_\lambda^*(x, X_i)\right), \end{aligned}$$

where

$$\delta_\lambda^*(x, X_i) = [\delta_\lambda(x, X_i) - \delta'_\lambda(x, X_i) g(x)]/D(x). \quad (4.3)$$

Thus, for $w(x)$ supported inside S , it makes sense to replace MISE by

$$\text{MISE}^* = E \int \left[n^{-1} \sum_{i=1}^n \delta_\lambda^*(x, X_i) \right]^2 w(x) dF(x). \quad (4.4)$$

Similarly, ISE and ASE may be replaced with

$$\begin{aligned} \text{ISE}^* &= \int \left[n^{-1} \sum_{i=1}^n \delta_\lambda^*(x, X_i) \right]^2 w(x) dF(x) \\ \text{ASE}^* &= n^{-1} \sum_{j=1}^n \left[n^{-1} \sum_{i=1}^n \delta_\lambda^*(X_j, X_i) \right]^2 w(X_j). \end{aligned} \quad (4.5)$$

Before the theorems are stated, note that MISE^* may be considered to be an assessment of how accurately the delta sequence estimator $\hat{g}^*(x)$, defined by

$$\hat{g}^*(x) = n^{-1} \sum_{i=1}^n \delta_\lambda^*(x, X_i),$$

estimates the function $g^*(x)$, defined by

$$g^*(x) \equiv 0.$$

Similarly ISE^* and ASE^* are the ISE and ASE for this new estimation problem. This observation allows immediate application of Theorems 1 and 2.

THEOREM 3. If δ_λ^* satisfies the assumptions (3.1)–(3.7) then

$$\limsup_{n \rightarrow \infty} \sup_{\lambda \in \mathcal{A}_n} \left| \frac{\text{ISE}^*(\lambda) - \text{MISE}^*(\lambda)}{\text{MISE}^*(\lambda)} \right| = 0 \quad a.s.$$

COROLLARY. If, in addition, (4.1) holds, then

$$\limsup_{n \rightarrow \infty} \sup_{\lambda \in \mathcal{A}_n} \left| \frac{\text{ISE}(\lambda) - \text{MISE}^*(\lambda)}{\text{MISE}^*(\lambda)} \right| = 0 \quad a.s.$$

THEOREM 4. If δ_λ^* satisfies the assumptions (3.1)–(3.10) and w is bounded, then

$$\limsup_{n \rightarrow \infty} \sup_{\lambda \in \mathcal{A}_n} \left| \frac{\text{ASE}^*(\lambda) - \text{MISE}^*(\lambda)}{\text{MISE}^*(\lambda)} \right| = 0 \quad a.s.$$

COROLLARY. If, in addition, (4.1) holds, then

$$\limsup_{n \rightarrow \infty} \sup_{\lambda \in \mathcal{A}_n} \left| \frac{\text{ASE}(\lambda) - \text{MISE}^*(\lambda)}{\text{MISE}^*(\lambda)} \right| = 0 \quad a.s.$$

To see how Theorem 1 and 2 are intimately related to Theorems 3 and 4, note that in the special case where $\hat{g}(x)$ is a delta sequence estimator (i.e., $\delta'_\lambda(x, X_i) \equiv 1$), conditions (4.1) and (4.2) hold trivially and the quantities MISE^* , ISE^* , and ASE^* are the same as their unasterisked counterparts. Thus Theorems 1 and 2 are special cases of Theorems 3 and 4. On the other hand, using the viewpoint given above, Theorems 3 and 4 are consequences of Theorems 1 and 2.

5. EXAMPLES

In this section it is seen how the fractional delta sequence estimators of Section 2 satisfy the conditions of Sections 3 and 4.

D-1. *Kernel estimators.* Conditions (3.4)–(3.7) follow easily from integration by substitution and the assumptions that f , $w \cdot f$, and K are bounded with $\int K(x) dx = 1$ and f , w not mutually singular. Condition (3.8) is also easily satisfied with $\xi = 1$. Condition (3.10) requires the additional assumption that $w \cdot f$ be integrable. Thus the results of Marron [23] and Theorems 1 and 2 of Hall [13] are special cases of the results of this paper.

D-2. *Histogram estimators.* Note that

$$\sup_{x_1, x_2} \delta_\lambda(x_1, x_2) = \lambda, \quad \sup_{x_2} \int \delta_\lambda(x_1, x_2) dx_1 = 1.$$

Hence, (3.4), (3.8), and (3.10) follow easily when it is assumed that f and $w \cdot f$ are bounded and integrable. Next observe that

$$\delta_\lambda(x_1, x_2) = \sum_{l=1}^{\infty} \lambda 1_l(x_1) 1_l(x_2) \left(\lambda \int_{A_l} w(x) dF(x) \right),$$

and so (3.4) with δ_λ replaced by δ_λ , (3.6), and (3.7) are satisfied under the above assumptions, together with (for (3.7)) the assumption that f and w are not mutually singular.

D-3. Orthogonal series estimators. The assumptions needed to verify (3.4) are summarized in

LEMMA 1. *If, for $k=1, 2, \dots$ there is a constant \mathcal{C}_k so that for $l_1, \dots, l_k = 1, 2, \dots$ and for $r = 1, \dots, k$,*

$$\int \psi_{l_1}^2(x) \cdots \psi_{l_k}^2(x) w(x)^r dF(x) \leq \mathcal{C}_k^2, \quad (5.1)$$

then (3.4) holds.

The proof of this lemma is in Section 7. Note that (5.1) is easily satisfied for either the familiar trigonometric or Hermite series. Next observe that

$$\iint \delta_\lambda(x_1, x_2)^2 w(x_1) dF(x_1) dF(x_2) = \int \sum_{l=1}^{\lambda} \psi_l(x_2)^2 w(x_2)^2 dF(x_2),$$

so (3.7) is easily satisfied. Condition (3.5) follows from

$$\delta_\lambda(x_1, x_2) = \delta_\lambda(x_1, x_2) w(x_1), \quad (5.2)$$

and the assumption that w is bounded. Condition (3.6) follows from (5.2) together with the Schwartz inequality. Verification of (3.8) follows easily from

$$\sup_{x_1} \left[\int \delta_\lambda(x_1, x_2) dF(x_2) - f(x_1) \right]^2 w(x_1) \leq \mathcal{C} \lambda^{(1-\xi)},$$

which is easy to check in the Hermite series case, but, using the computations of Hall [12], requires the additional assumption of f'' bounded in the case of trigonometric series. Condition (3.10) is obvious under the above assumptions for either the trigonometric or Hermite series. Theorem 3 of Hall [13] is a special case of this.

R-1, Kernel estimators. Conditions (3.4)–(3.10) are easily verified under the same assumptions as D-1, above, together with the assumption that for $k = 1, 2, \dots$ there is a constant \mathcal{C}_k so that, for z in the support of w ,

$$E[Y^k | Z = z] \leq \mathcal{C}_k.$$

The verification of (4.1) is easy, in view of Lemma 1 of Härdle and Marron [19], under the additional assumption that f_M is Hölder continuous.

R-2. *Known marginal kernel estimators.* This case is similar to R-1 except that (4.1) is not required (but (4.2) is still important). R-1 and R-2 contain the results of Härdle [17] and Hall [16] as special cases.

H-1. *Kernel estimators.* Conditions (3.4)–(3.10) are easily checked when it is assumed that

$$\int K(x) dx = 1,$$

and K , f , and $w \cdot f$ are bounded, together with the assumption that $1 - F$ is bounded above 0 on the support of w .

6. PROOFS OF THEOREMS 1 AND 2

Note that, by (3.2) and the Chebyshev inequality, for $\varepsilon > 0$, $k = 1, 2, \dots$,

$$P \left[\sup_{\lambda \in \mathcal{A}_n} \left| \frac{\text{ISE}(\lambda) - \text{MISE}(\lambda)}{\text{MISE}(\lambda)} \right| > \varepsilon \right] \leq \mathcal{C} n^\rho \sup_{\lambda \in \mathcal{A}_n} E \left[\frac{[\text{ISE}(\lambda) - \text{MISE}(\lambda)]^{2k}}{\text{MISE}(\lambda) \cdot \varepsilon} \right].$$

Thus, by the Borel–Cantelli lemma, the proof of Theorem 1 will be complete when it is seen that there is a constant $\gamma > 0$, so that for $k = 1, 2, \dots$, there are constants \mathcal{C}_k so that

$$E \left[\frac{[\text{ISE}(\lambda) - \text{MISE}(\lambda)]^{2k}}{\text{MISE}(\lambda)} \right] \leq \mathcal{C}_k n^{-\gamma k}. \quad (6.1)$$

Theorem 2 will be established by the same technique when it is shown that

$$E \left[\frac{[\text{ASE}(\lambda) - \text{MISE}(\lambda)]^{2k}}{\text{MISE}(\lambda)} \right] \leq \mathcal{C}_k n^{-\gamma k}. \quad (6.2)$$

The distance ISE can be decomposed as

$$\text{ISE} = R(\lambda) + 2S(\lambda) + b(\lambda),$$

where $b(\lambda)$ is defined in (3.9) and

$$\begin{aligned} R(\lambda) &= \iiint \delta_\lambda(x_1, x_2) \delta_\lambda(x_1, x_3) w(x_1) dF(x_1) \\ &\quad \times d(F_n - F)(x_2) d(F_n - F)(x_3), \end{aligned}$$

$$S(\lambda) = \iint \delta_\lambda(x_1, x_2) B(x_1) w(x_1) dF(x_1) d(F_n - F)(x_2).$$

The first term may be further split into

$$R(\lambda) = R_1(\lambda) + R_2(\lambda) + R_3(\lambda),$$

where, using the notation (3.5),

$$R_1(\lambda) = \iint_{\{x_2 \neq x_3\}} \delta_\lambda(x_2, x_3) d(F_n - F)(x_2) d(F_n - F)(x_3),$$

$$R_2(\lambda) = n^{-1} \int \delta_\lambda(x_2, x_2) d(F_n - F)(x_2),$$

$$R_3(\lambda) = n^{-1} \int \delta_\lambda(x_2, x_2) dF(x_2).$$

To finish the proof of (6.1) it is enough to show that

$$\left[\frac{R_3(\lambda) + b(\lambda) - \text{MISE}(\lambda)}{\text{MISE}(\lambda)} \right]^{2k} \leq \mathcal{C}_k n^{-\gamma k}, \quad (6.3)$$

and for "term" denoting R_1 , R_2 , or S ,

$$E \left[\frac{\text{term}}{\text{MISE}(\lambda)} \right]^{2k} \leq \mathcal{C}_k n^{-\gamma k}. \quad (6.4)$$

Write

$$\text{ASE} = \text{ISE} + T(\lambda).$$

As above, $T(\lambda)$ admits the decomposition

$$\begin{aligned} T &= T_1 + T_2 + T_3 + 2T_4 + 2T_5 + T_6 \\ &\quad + T_7 + 2U_1 + 2U_2 + 2U_3 + V, \end{aligned}$$

where

$$\begin{aligned} T_1(\lambda) &= \iint_{\{x_1 \neq x_2 \neq x_3 \neq x_1\}} \delta_\lambda(x_1, x_2) \delta_\lambda(x_1, x_3) w(x_1) d(F_n - F)(x_1) \\ &\quad \times d(F_n - F)(x_2) d(F_n - F)(x_3), \end{aligned}$$

$$T_2(\lambda) = n^{-1} \iint_{\{x_1 \neq x_2\}} \delta_\lambda(x_1, x_2)^2 w(x_1) d(F_n - F)(x_1) d(F_n - F)(x_2),$$

$$T_3(\lambda) = n^{-1} \iint \delta_\lambda(x_1, x_2)^2 w(x_1) d(F_n - F)(x_1) dF(x_2),$$

$$T_4(\lambda) = n^{-1} \iint_{\{x_1 \neq x_2\}} \delta_\lambda(x_1, x_2) \delta_\lambda(x_1, x_1) w(x_1) \\ \times d(F_n - F)(x_1) d(F_n - F)(x_2),$$

$$T_5(\lambda) = n^{-1} \iint \delta_\lambda(x_1, x_2) \delta_\lambda(x_1, x_1) w(x_1) dF(x_1) d(F_n - F)(x_2),$$

$$T_6(\lambda) = n^{-2} \int \delta_\lambda(x, x)^2 w(x) d(F_n - F)(x),$$

$$T_7(\lambda) = n^{-2} \int \delta_\lambda(x, x)^2 w(x) dF(x),$$

$$U_1(\lambda) = \iint_{\{x_1 \neq x_2\}} \delta_\lambda(x_1, x_2) B(x_1) w(x_1) d(F_n - F)(x_1) d(F_n - F)(x_2),$$

$$U_2(\lambda) = n^{-1} \int \delta_\lambda(x_1, x_1) B(x_1) w(x_1) d(F_n - F)(x_1),$$

$$U_3(\lambda) = n^{-1} \int \delta_\lambda(x_1, x_1) B(x_1) w(x_1) dF(x_1),$$

$$V(\lambda) = \int B(x_1)^2 w(x_1) d(F_n - F)(x_1).$$

Thus, (6.2) will be established when (6.4) is verified for each of the above terms as well.

To check (6.3), note that by the familiar variance-bias squared decomposition (see, e.g., Rosenblatt [31]), using the notation (3.9),

$$\text{MISE} = R_3(\lambda) - r(\lambda) + b(\lambda),$$

where, using the notation (3.5),

$$r(\lambda) = n^{-1} \iint \tilde{\delta}_\lambda(x_2, x_3) dF(x_2) dF(x_3).$$

The inequality (6.3) follows from this and from (3.3), (3.6), and (3.7).

The verification of (6.4) will now be done term by term, starting with those which do not involve $d(F_n - F)$:

Term T₇. Using (3.10),

$$\left[\frac{n^{-2} \int \delta_\lambda(x, x)^2 w(x) dF(x)}{\text{MISE}(\lambda)} \right]^{2k} \leq \mathcal{C}_k \left[\frac{n^{-2} \lambda^2}{n^{-1} \lambda} \right]^{2k} \leq \mathcal{C}'_k n^{-2k\xi}.$$

Term U₃. As above, using the Schwartz inequality,

$$\begin{aligned} & \left[\frac{n^{-1} \int \delta_\lambda(x_1, x_1) B(x_1) w(x_1) dF(x_1)}{\text{MISE}(\lambda)} \right]^{2k} \\ & \leq \left[\frac{n^{-1} [\int \delta_\lambda(x_1, x_1)^2 w(x_1) dF(x_1)]^{1/2} b(\lambda)^{1/2}}{\text{MISE}(\lambda)} \right]^{2k} \\ & \leq \mathcal{C}_k \left[\frac{n^{-1} \lambda \cdot b(\lambda)^{1/2}}{(n^{-1} \lambda)^{1/2} b(\lambda)^{1/2}} \right]^{2k} \leq \mathcal{C}_k (n^{-1} \lambda)^k \leq \mathcal{C}'_k n^{-k\varepsilon}. \end{aligned}$$

The remaining terms all have at least one $d(F_n - F)$, and so have mean 0. Thus to check (6.4), by the cumulant expansion of the $2k$ th moment, it is enough to check that, for $k = 2, 3, \dots$, there is a constant \mathcal{C}_k so that

$$\left| \text{cum}_k \left(\frac{\text{term}}{\text{MISE}} \right) \right| \leq \mathcal{C}_k n^{-\gamma k}, \quad (6.5)$$

where $\text{cum}_k(\cdot)$ denotes the k th order cumulant, for which each argument is the same.

To verify (6.5) in the case of those terms having only one $d(F_n - F)$, note that they may be written

$$n^{-1} \sum_{i=1}^n W(X_i).$$

Thus, using the independence property and linearity of cumulants, it is enough to show that

$$n^{-k+1} \text{MISE}^{-k} |E[W(X_1)]^k| \leq \mathcal{C}_k n^{-\gamma k}.$$

Term R₂. Note that here

$$\begin{aligned} W(X_2) = n^{-1} & \left[\int \delta_\lambda(x_1, X_2)^2 w(x_1) dF(x_1) \right. \\ & \left. - \iint \delta_\lambda(x_1, x_2)^2 w(x_1) dF(x_1) dF(x_2) \right]. \end{aligned}$$

So by the binomial theorem and repeated application of (3.4),

$$n^{-k+1} \text{MISE}^{-k} |E[W(X_2)]^k| \leq \mathcal{C}_k n^{-2k+1} (n^{-1} \lambda)^{-k} \lambda^{2k-(k+1)/2} \leq \mathcal{C}'_k n^{-k/4}.$$

Term T₃. Similar to R_2 .

Term T₅. Similar to R_2 .

Term T₆. Note that here

$$W(X_1) = n^{-2} \left[\delta_\lambda(X_1, X_1)^2 w(X_1) - \int \delta_\lambda(x_1, x_1)^2 w(x_1) dF(x_1) \right].$$

So by (3.10)

$$n^{-k+1} \text{MISE}^{-k} |E[W(X_1)]^k| \leq \mathcal{C}_k n^{-3k+1} (n^{-1}\lambda)^{-k} \lambda^{2k} \leq \mathcal{C}'_k n^{-k/2}.$$

Term V. Note that here

$$W(X_1) = B(X_1)^2 w(X_1) - b(\lambda).$$

Thus, by (3.8),

$$\begin{aligned} n^{-k+1} \text{MISE}^{-k} |E[W(X_1)]^k| &\leq \mathcal{C}_k n^{-k+1} \left(b(\lambda)^k [b(\lambda)]^{-k} \right. \\ &\quad \left. + \sum_{j=1}^k [b(\lambda) \lambda^{(j-1)(1-\epsilon)}] b(\lambda)^{k-j} [(n^{-1}\lambda)^{j-1} b(\lambda)^{k-j+1}]^{-1} \right) \\ &\leq \mathcal{C}'_k n^{-\gamma k}. \end{aligned}$$

Term U₂. Note that here

$$W(X_1) = n^{-1} \left[\delta_\lambda(X_1, X_1) B(X_1) w(X_1) - \int \delta_\lambda(x_1, x_1) B(x_1) w(x_1) dF(x_1) \right].$$

By (3.8), (3.10), and the Schwartz inequality, for $j = 1, 2, \dots$, there is \mathcal{C}_j so that

$$\begin{aligned} &\left| \int [\delta_\lambda(x_1, x_1) B(x_1) w(x_1)]^j dF(x_1) \right| \\ &\leq \left[\int \delta_\lambda(x_1, x_1)^{2j} w(x_1)^{2j-1} dF(x_1) \right]^{1/2} \left[\int B(x_1)^{2j} w(x_1) dF(x_1) \right]^{1/2} \\ &\leq \mathcal{C}_j \lambda^{(3j-1)/2} b(\lambda)^{1/2}. \end{aligned}$$

Hence,

$$\begin{aligned} n^{-k+1} \text{MISE}^{-k} |E[W(X_1)]^k| &\leq \mathcal{C}_k n^{-2k+1} [(n^{-1}\lambda)^{-k+1/2} b(\lambda)^{-1/2}] \lambda^{(3k-1)/2} b(\lambda)^{1/2} \\ &\leq \mathcal{C}_k^1 n^{-k/4}. \end{aligned}$$

Term S. Note that here

$$W(X_2) = \int \delta_\lambda(x_1, X_2) B(x_1) w(x_1) dF(x_1) \\ - \iint \delta_\lambda(x_1, x_2) B(x_1) w(x_1) dF(x_1) dF(x_2).$$

It follows from the Schwartz inequality that,

$$\left| \int \left[\int \delta_\lambda(x_1, x_2) B(x_1) w(x_1) dF(x_1) \right]^j dF(x_2) \right| \\ \leq \int \left[\int \delta_\lambda(x_1, x_2)^2 w(x_1) dF(x_1) \right]^{j/2} b(\lambda)^{j/2} dF(x_2). \quad (6.6)$$

So, by (3.4), for j even, there is a constant \mathcal{C}_j such that (6.6) is bounded by

$$\mathcal{C}_j b(\lambda)^{j/2} \lambda^{j - (j/2 + 1)/2} = \mathcal{C}_j \lambda^{3j/4 - 1/2} b(\lambda)^{j/2}.$$

And by the moment inequality, for j odd, there is a constant \mathcal{C}_j such that (6.6) is bounded by

$$b(\lambda)^{j/2} \left[\int \left(\int \delta_\lambda(x_1, x_2)^2 w(x_1) dF(x_1) \right)^{(j+1)/2} dF(x_2) \right]^{j/(j+1)} \\ \leq \mathcal{C}_j b(\lambda)^{j/2} \lambda^{3j/4 - j/2(j+1)}.$$

Thus, for $k = 3, 4, \dots$

$$n^{-k+1} \text{MISE}^{-k} |EW(X_2)^k| \leq \mathcal{C}_k n^{-k+1} (n^{-1}\lambda)^{-k/2} b(\lambda)^{-k/2} b(\lambda)^{k/2} \lambda^{3k/4 - 3/8} \\ = \mathcal{C}_k n^{-k/2+1} \lambda^{k/4 - 3/8} = \mathcal{C}_k (n^{-1}\lambda)^{k/2-1} \lambda^{-k/4+5/8} \\ \leq \mathcal{C}'_k n^{-\epsilon k/4}.$$

More precise computations are required in the case $k = 2$. By (3.5),

$$E(W(X_2)^2) \leq E \left[\int \delta_\lambda(x_1, X_2) B(x_1) w(x_1) dF(x_1) \right]^2 \\ = \int \left[\int \left(\int \delta_\lambda(x_1, x_2) \delta_\lambda(x'_1, x_2) dF(x_2) \right) \right. \\ \left. \times B(x'_1) w(x'_1) dF(x'_1) \right] B(x_1) w(x_1) dF(x_1) \\ \leq \left(\int \left[\int \int \delta_\lambda(x_1, x_2) \delta_\lambda(x'_1, x_2) dF(x_2) \right. \right. \\ \left. \left. \times B(x'_1) w(x'_1) dF(x'_1) \right]^2 w(x_1) dF(x_1) \right)^{1/2} b(\lambda)^{1/2}$$

$$\begin{aligned} &\leq \left(\int \left[\int \left(\int \left(\int \delta_\lambda(x_1, x_2) \delta_\lambda(x'_1, x_2) dF(x_2) \right)^2 w(x'_1) \right. \right. \right. \\ &\quad \left. \left. \left. \times dF(x'_1) \right)^{1/2} b(\lambda)^{1/2} \right]^2 w(x_1) dF(x_1) \right)^{1/2} b(\lambda)^{1/2} \\ &= b(\lambda) \left(\iint \delta_\lambda(x_2, x'_2)^2 dF(x_2) dF(x'_2) \right)^{1/2} \leq b(\lambda) \mathcal{C}(\lambda^2 - 2/2)^{1/2}. \end{aligned}$$

Thus,

$$n^{-1} \text{MISE}(\lambda)^{-2} EW(X_2)^2 \leq \mathcal{C} n^{-1} (n^{-1} \lambda b(\lambda))^{-1} b(\lambda) \lambda^{1/2} \leq \mathcal{C}' n^{-\varepsilon/2}.$$

It remains to verify (6.5) for the terms containing two or three $d(F_n - F)$'s. The terms containing 2 may all be written in the form

$$n^{-1} \sum_{\substack{i, i' = 1 \\ i \neq i'}}^n W(X_i, X_{i'}),$$

where

$$EW(X_i, X_{i'}) = 0, \quad i \neq i'.$$

So, using the linearity property of cumulants, (6.5) will be established in this case when it is seen that there is a constant $\gamma > 0$, so that for $k = 2, 3, \dots$, there are constants \mathcal{C}_k such that

$$\left| n^{-2k} \text{MISE}^{-k} \sum_{i_1, i'_1, \dots, i_k, i'_k} \text{cum}_k(W(X_{i_1}, X_{i'_1}), \dots, W(X_{i_k}, X_{i'_k})) \right| \leq \mathcal{C}_k n^{-\gamma k},$$

where, by a moment expansion of cum_k , it may be assumed that each of $i_1, i'_1, \dots, i_k, i'_k$ appears at least twice. In each case, it will be convenient to let m denote the number of $i_1, i'_1, \dots, i_k, i'_k$ that are unique. Note that, for $m = 2, 3, \dots, k$, the number of cum_k with m distinct indices is bounded by $\mathcal{C}_k n^m$.

Term T₂. Note that here

$$\begin{aligned} W(X_i, X_{i'}) &= n^{-1} \left[\delta_\lambda(X_i, X_{i'})^2 w(X_i) - \int \delta_\lambda(X_i, x_2)^2 w(X_i) dF(x_2) \right. \\ &\quad - \int \delta_\lambda(x_1, X_{i'})^2 w(x_1) dF(x_1) \\ &\quad \left. + \iint \delta_\lambda(x_1, x_2)^2 w(x_1) dF(x_1) dF(x_2) \right]. \end{aligned}$$

So, by (3.4)

$$\begin{aligned} & \left| n^{-2k} \text{MISE}^{-k} \sum \text{cum}_k(W(X_{i_1}, X_{i_1}), \dots, W(X_{i_k}, X_{i_k})) \right| \\ & \leq n^{-2k} (n^{-1}\lambda)^{-k} n^{-k} \mathcal{C}_k \sum_{m=2}^k n^m \lambda^{2k-m/2} \leq \mathcal{C}'_k n^{-k/2}. \end{aligned}$$

Term T₄. Similar to *T₂*.

Term R₁. Here

$$\begin{aligned} W(X_i, X'_i) &= \delta_\lambda(X_i, X'_i) - \int \delta_\lambda(X_i, x_2) dF(x_2) - \int \delta_\lambda(x_1, X'_i) dF(x_1) \\ &+ \iint \delta_\lambda(x_1, x_2) dF(x_1) dF(x_2). \end{aligned}$$

Thus,

$$\begin{aligned} & \left| n^{-2k} \text{MISE}^{-k} \sum \text{cum}_k(W(X_{i_1}, X_{i_1}), \dots, W(X_{i_k}, X_{i_k})) \right| \\ & \leq n^{-2k} (n^{-1}\lambda)^{-k} \mathcal{C}_k \sum_{m=2}^k n^m \lambda^{k-m/2} \leq \mathcal{C}'_k n^{-\epsilon k/2}. \end{aligned}$$

Term U₁. Here

$$\begin{aligned} W(X_i, X'_i) &= \delta_\lambda(X_i, X'_i) B(X_i) w(X_i) - \int \delta_\lambda(X_i, x_2) B(X_i) w(X_i) dF(x_2) \\ &- \int \delta_\lambda(x_1, X'_i) B(X_1) w(x_1) dF(x_1) \\ &+ \iint \delta_\lambda(x_1, x_2) B(x_1) w(x_1) dF(x_1) dF(x_2). \end{aligned}$$

This term is handled by means quite similar to those used on Term *T₂* above, except that (3.4) is augmented by the Schwartz inequality and (3.8). The result is, for $k = 2, 3, \dots$,

$$\begin{aligned} & \left| n^{-2k} \text{MISE}^{-k} \sum \text{cum}_k(W(X_{i_1}, X_{i_1}), \dots, W(X_{i_k}, X_{i_k})) \right| \\ & \leq n^{-2k} ((n^{-1}\lambda)^{k-1/2} b(\lambda)^{1/2})^{-1} \mathcal{C}_k \sum_{m=2}^k n^m \lambda^{(2k-m)/2} b(\lambda)^{1/2} \lambda^{(k-1)(1-\epsilon)/2} \\ & \leq \mathcal{C}'_k n^{-\epsilon^2 k/4}. \end{aligned}$$

It remains to verify (6.5) for

Term T₁. This term may be handled by methods similar to those used on term *T₂*.

This completes the proof of Theorems 1 and 2.

7. PROOF OF LEMMA 1

Using the definition of $\delta_\lambda(x, y)$, write

$$\begin{aligned} & \left| \int \cdots \int \left[\prod_{i,i'} \delta_\lambda(x_i, x_{i'})^{\alpha_{ii'}} \right] \left[\prod_i w(x_i)^{\beta_i} \right] dF(x_1) \cdots dF(x_m) \right| \\ &= \left| \sum_{l_1=1}^{\lambda} \cdots \sum_{l_k=1}^{\lambda} \int \cdots \int \psi_{l_1} \psi_{l_1} w \cdots \psi_{l_k} \psi_{l_k} w \left[\prod w^{\beta_i} \right] dF(x_1) \cdots dF(x_m) \right|. \quad (7.1) \end{aligned}$$

The multiple integral on the right-hand side may now be factored to give an expression of the form

$$\left| \sum_{l_1} \cdots \sum_{l_k} \left[\int dF(x_1) \right] \cdots \left[\int dF(x_m) \right] \right|. \quad (7.2)$$

Consider the set of $\alpha_{ii'}$ which have $i \neq i'$ and are positive. Find a subset, A , which has the property that each of $1, \dots, m$ appears at least once as an i or i' , and suppose that this subset is minimal in the sense that if any $\alpha_{ii'}$ is removed, then $1, \dots, m$ no longer all appear as the index of an α .

Group $1, \dots, m$ into two subsets, I and I' , by the following rules:

(1) Any of $1, \dots, m$ that appear twice (or more) as an index of an α in A goes into I .

(2) If i is in I , and $\alpha_{ii'}$ (or $\alpha_{i'i}$) is in A , put i' into I' .

(3) For the remaining $\alpha_{ii'}$ in A , put i in I and i' in I' .

The above rules partition $\{1, \dots, m\}$ into I and I' .

Observe that for each $\alpha_{ii'}$ in A , there is an l so that $\psi_l(x_i) \psi_l(x_{i'})$ appears in the integrand on the right side of (7.1). Suppose, without loss of generality, that l_1, \dots, l_L each correspond in this manner to a different element of A , where L denotes the cardinality of A . Also assume, without loss of generality, that l_1, \dots, l_b correspond to those α in A which have an index appearing more than once in A .

By the Schwartz inequality, (7.2) may be written as

$$\begin{aligned} & \left| \sum_{l_1} \cdots \sum_{l_k} \prod_{i=1}^m \int [\quad] dF(x_i) \right| \\ &= \left| \sum_{l_{L+1}} \cdots \sum_{l_k} \left[\sum_{l_1} \cdots \sum_{l_L} \left(\prod_{i \in I} \int [\quad] dF(x_i) \right) \left(\prod_{i \in I'} \int [\quad] dF(x_i) \right) \right] \right| \end{aligned}$$

$$\leq \sum_{l_{L+1}} \cdots \sum_{l_k} \left[\sum_{l_1} \cdots \sum_{l_L} \left(\prod_{i \in I'} \int [\] dF(x_i) \right)^2 \right]^{1/2} \\ \times \left[\sum_{l_1} \cdots \sum_{l_L} \left(\prod_{i \in I'} \int [\] dF(x_i) \right)^2 \right]^{1/2}.$$

Suppose, without loss of generality, that $I' = \{1, \dots, L\}$. Then,

$$\sum_{l_1} \cdots \sum_{l_L} \left(\prod_{i \in I'} \int [\] dF(x_i) \right)^2 = \prod_{i=1}^L \sum_{l_i} \left(\int [\] dF(x_i) \right)^2 \leq \prod_{i=1}^L \mathcal{C}_k,$$

where the last inequality follows from (5.1) and the Bessel inequality, because $\int [\] dF(x_i)$ is the l_i th Fourier coefficient of a function whose norm is bounded in (5.1). Similar techniques give

$$\sum_{l_1} \cdots \sum_{l_L} \left(\prod_{i \in I'} \int [\] dF(x_i) \right)^2 \leq \lambda^{b-1} \mathcal{C}_k^{L-b+1}.$$

It follows from the above that there is a constant \mathcal{C}_k so that (7.1) is bounded by

$$\mathcal{C}_k \lambda^{k-L} \lambda^{(b-1)/2}.$$

To put this in more useful terms, note that

$$2L - b \geq m - 1$$

and so

$$-L + b/2 - \frac{1}{2} \leq -m/2.$$

It follows that (7.1) is bounded by

$$\mathcal{C}_k \lambda^{k-m/2}.$$

This completes the proof of Lemma 1.

REFERENCES

- [1] BEAN, S., AND TSOKOS, C. P. (1982). Bandwidth selection procedures for kernel density estimates. *Comm. Statist. A* **11** 1045-1069.
- [2] BURMAN, P. (1984). A data dependent approach to density estimation. *Z. Wahrsch. Verw. Gebiete* **69** 609-628.
- [3] BURMAN, P., AND CHEN, K. W. (1984). Nonparametric estimation of a regression function, unpublished.
- [4] BOWMAN, A. W. (1982). A Comparative Study of Some Kernel-Based Non-Parametric Density Estimators. *J. Statist. Comput. Simulation* **21** 313-327.
- [5] BREIMAN, L., MEISEL, W., AND PURCELL, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics* **19** 135-144.

- [6] CENCOV, N. N. (1962). Evaluation of an unknown distribution density from observations. *Soviet Math.* **3** 1559-1562.
- [7] COLLOMB, G. (1981). Estimation non parametrique de la regression: Revue Bibliographique. *Internat. Statist. Rev.* **49** 75-93.
- [8] COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd. ed. Wiley, New York.
- [9] ENGLE, R. F., GRANGER, C. W. J., RICE, J., AND WEISS, A. (1983). *Non-Parametric Estimates of the Relation Between Weather and Elasticity of Demand*. Discussion paper #83-17, Department of Economics, University of California, San Diego.
- [10] FÖLDES, A., AND REVESZ, P. (1974). A general method for density estimation. *Studia Sci. Math. Appl. Hungar.* **9** 81-92.
- [11] FRYER, M. J. (1977). Review of some non-parametric methods of density estimation. *J. Inst. Math. Its Appl.* **20** 335-354.
- [12] HALL, P. (1981). On trigonometric series estimates of densities. *Ann. Statist.* **9** 683-685.
- [13] HALL, P. (1982). Limit theorems for stochastic measures of the accuracy of density estimators. *Stochastic Process. Appl.* **13** 11-25.
- [14] HALL, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11** 1156-1174.
- [15] HALL, P. (1984a). Central limit theorem for integrated square error of multivariate non-parametric density estimators. *J. Multivariate Anal.* **14** 1-16.
- [16] HALL, P. (1984b). Asymptotic properties of integrated square error and cross-validation for kernel estimation of a regression function. *Z. Wahrsch. Verw. Gebiete* **67** 175-196.
- [17] HÄRDLE, W. (1984). Approximations to the mean integrated squared error with applications to optimal bandwidth selection for nonparametric regression function estimators. *J. Multivariate Anal.* **18** 150-160.
- [18] HÄRDLE, W., AND MARRON, J. S. (1983). *The Nonexistence of Moments of Some Kernel Regression Estimators*. Mimeo Series No. 1537. (North Carolina Institute of Statistics.)
- [19] HÄRDLE, W., AND MARRON, J. S. (1985a). Optimal bandwidth selection in non-parametric regression function estimation. *Ann. Statist.* **13** 1465-1481.
- [20] HÄRDLE, W., AND MARRON, J. S. (1985b). Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression. *Biometrika* **72** 481-484.
- [21] JOHNSTON, G. J. (1982). Properties of maximal deviations for nonparametric regression function estimates. *J. Multivariate Anal.* **12** 402-414.
- [22] MARRON, J. S. (1984). An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *Ann. Statist.* **13** 1011-1023.
- [23] MARRON, J. S. (1986). Convergence properties of an empirical error criterion for multivariate density estimation. *J. Multivariate Anal.* **19** 1-13.
- [24] NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141-142.
- [25] PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Statist.* **33** 1056-1076.
- [26] PRAKASA RAO, B. L. S. (1983). *Nonparametric Functional Estimation*. Academic Press, New York.
- [27] RAATGEVER, J. W., AND DUIN, R. P. W. (1978). On the variable kernel model for multivariate nonparametric density estimation. In *COMPSTAT 1978: Proceedings* (L. C. A. Corsten and J. Hermans, Eds.). Birkhäuser, Basel.
- [28] RICE, J. (1982). Bandwidth choice for nonparametric kernel regression. *Ann. Statist.* **12** 1215-1230.
- [29] ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832-837.
- [30] ROSENBLATT, M. (1969). Conditional probability density and regression estimators. In *Multivariate Analysis-II* (P. R. Krishnaiah, Ed.) pp. 25-31. Academic Press, New York.
- [31] ROSENBLATT, M. (1971). Curve estimates, *Ann. Math. Statist.* **42** 1815-1842.

- [32] RUST, A. E., AND TSOKOS, C. P. (1981). On the convergence of kernel estimators of probability density functions. *Ann. Inst. Statist. Math.* **33** 233-246.
- [33] SCOTT, D. W., AND FACTOR, L. E. (1981). Monte Carlo study of three data-based non-parametric probability density estimators. *J. Amer. Statist. Assoc.* **76** 9-15.
- [34] SILVERMAN, B. W. (1984). A fast and efficient cross-validation method for smoothing parameter choice in spline regression. *J. Amer. Statist. Assoc.*, in press.
- [35] STEELE, J. M. (1978). Invalidity of average squared error criterion in density estimation. *Canad. J. Statist.* **6** 193-200.
- [36] STONE, C. J. (1976). Nearest neighbor estimators of a nonlinear regression function. In *Proceedings, Comput. Sci. Statist. 8th Annual Symposium on the Interface*. Health Sciences Computing Facility, U.C.L.A., pp. 413-418.
- [37] STONE, C. J. (1984a). An asymptotically efficient histogram selection rule. *Proceedings of the Neyman-Kiefer Meeting*, in press.
- [38] STONE, C. J. (1984b). An asymptotically optimal window selection rule for kernel density estimates *Ann. Statist.*, in press.
- [39] SUSARLA, V., AND WALTER, G. (1981). Estimation of a multivariate density function using delta sequences. *Ann. Statist.* **9** 347-355.
- [40] TANNER, M. A., AND WONG W. H. (1982). Data based nonparametric estimation of the hazard function with applications to model diagnostics and exploratory analysis. *J. Amer. Statist. Assoc.* **79** 174-182.
- [41] TUKEY, J. W. (1961). Curves as parameters, and touch estimation. *Proceedings, 4th Berkely Sympos.* 681-694.
- [42] WAHBA, G. (1977). Optimal smoothing of density estimates. In *Classification and Clustering*. (J. van Ryzin, Ed.), pp. 423-458.
- [43] WALTER, G. (1977). Properties of Hermite series estimation of probability density. *Ann. Statist.* **5** 1258-1264.
- [44] WALTER, G., AND BLUM, J. (1976). Probability density estimation using delta sequences. *Ann. Statist.* **7** 328-340.
- [45] WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A.* **26** 359-372.
- [46] WATSON, G. S., AND LEADBETTER, M. R. (1964a). Hazard analysis, I. *Biometrika* **51** 175-184.
- [47] WATSON, G. S., AND LEADBETTER, M. R. (1964b). Hazard Analysis, II. *Sankhyā Ser. A* **26** 101-116.
- [48] WEGMAN, E. J. [1972]. Nonparametric probability density estimation. II. A comparison of density estimation methods. *J. Statist. Comput. Simulation* **1** 225-245.

[2] P. Whittle, "The analysis of multiple stationary time series," *J. Roy. Statist. Soc.*, vol. 15, pp. 125-139, 1953.
 [3] R. H. Jones, "Spectral analysis with regularly missed observations," *Ann. Math. Statist.*, vol. 32, pp. 455-461, 1962.
 [4] E. Parzen, "On spectral analysis with missing observations and amplitude modulation," *Sankhya*, Ser. 4, vol. 25, pp. 383-392, 1963.
 [5] B. Friedlander and B. Porat, "A general lower bound for parametric spectrum estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 728-733, Aug. 1984.

A Note on Jackknifing Kernel Regression Function Estimators

WOLFGANG HÄRDLE

Abstract—Estimation of the value of a regression function at a point of continuity using a kernel-type estimator is considered and improvements by a jackknife technique are discussed. It is seen that a so-called generalized jackknife estimator asymptotically improves upon an ordinary kernel-type estimator. However, for a fixed sample size the generalized jackknife method may inflate the mean-square error.

I. INTRODUCTION AND BACKGROUND

Consider the observation model

$$Y_{ni} = m(t_{ni}) + \epsilon_{ni}, \quad 1 \leq i \leq n,$$

where $\epsilon_{n1}, \dots, \epsilon_{nn}$ are random errors, $t_{ni} = i/n$ are equispaced knot points in the interval $(0, 1]$, and m is an unknown regression function. The goal is to estimate m from the observations $\{(t_{ni}, Y_{ni})\}_{i=1}^n$. We consider here so-called kernel estimators

$$m_n(t) = \sum_{i=1}^n W_{ni}(t) Y_i, \quad 0 < t < 1,$$

where $\{W_{ni}\}_{i=1}^n$ is a sequence of weights generated by a continuous kernel function K , that is,

$$W_{ni}(t) = n^{-1} h^{-1} K\left(\frac{t - t_{ni}}{h}\right)$$

with a bandwidth $h = h_n$. Similar estimators have been considered by Georgiev [3] and Györfi [5] in signal processing and system identification.

Assume that the random errors $\{\epsilon_{ni}\}_{i=1}^n$ are independent and identically distributed zero-mean random variables with variance σ^2 , having a distribution independent of n . The mean-square error (mse) of m_n for fixed t can then be written as

$$\text{mse}(m_n) = \sigma^2 \sum_{i=1}^n W_{ni}^2(t) + \left(\sum_{i=1}^n W_{ni}(t) m(t_{ni}) - m(t) \right)^2$$

It has been shown (e.g., Priestley and Chao [7]) that under natural conditions on K and m the mse converges with a certain algebraic rate to zero if the sequence of bandwidths is suitably chosen. Schucany and Sommers [8] argued in a similar setting of kernel density estimation that a so-called generalized jackknife estimate might be helpful in improving this algebraic rate of convergence.

Manuscript received November 27, 1984; revised July 24, 1985. This work was supported in part by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 123 "Stochastische Mathematische Modelle," and Air Force Scientific Research Contract AFOSR-F49620 82 C 0009.

The author is with the Inst.f. Wirtschaftstheorie II, Universität Bonn, D-5300 Bonn, West Germany.
 IEEE Log Number 8406647.

In this correspondence we define a generalized jackknife estimate for $m(t)$ and show that, indeed, under certain assumptions the jackknife technique reduces the bias asymptotically. However, we will also see in an example that for a fixed sample size n , the variance of the generalized jackknife estimator may dominate the variance of $m_n(t)$ in such a drastic way that the mse of the ordinary kernel regression estimator $m_n(t)$ is smaller than the mse of the generalized jackknife estimator.

The generalized jackknife technique with the often useful bias reduction property should therefore be cautiously applied in this context. A proper inspection of the parameters involved (see Table II) seems to be necessary before using this sophisticated method. This observation has also been made by Efron [1] for the traditional jackknife. Also Huber [6, p. 16] points out that the jackknife may yield a variance that is worse than useless.

II. DOES THE JACKKNIFED ESTIMATE IMPROVE UPON m_n ?

Define the constants $\Lambda(K; j)$ by

$$j! \Lambda(K; j) = \int u^j K(u) du, \quad j \in \mathbf{N} \cup \{0\}$$

and consider only symmetric kernel functions with $\int K^2(u) du < \infty$. Note that $\Lambda(K; 0) = 1$ and that the symmetry entails $\Lambda(K; j) = 0$ for all odd integers $j \in \mathbf{N}$. A kernel K is said to be in the class \mathcal{R}_r , if for some even integer $r \geq 2$

$$\Lambda(K; j) = 0, \quad j \leq r - 1 \\ \neq 0, \quad j = r.$$

Let $m^{(s)}(t)$ denote the s th derivative of the regression function. The knot point $t \in (0, 1)$ is considered as fixed for the rest of this correspondence.

Proposition 1: Suppose that $h = h_n \rightarrow 0$ such that $nh \rightarrow \infty$ as $n \rightarrow \infty$. Let $m \in C^p[0, 1]$, $p = 2q$, $q \in \mathbf{N}$, and let $K \in \mathcal{R}_r$, $r = 2s \leq p$. Then

$$\text{mse}(m_n) = (nh)^{-1} \sigma^2 \int K^2(u) du \\ + \left[\sum_{j=s}^q h^{2j} m^{(2j)}(t) \Lambda(K; 2j) \right]^2 + o(n^{-1} h^{-1} + h^{4s}).$$

Proof: Use the fact that $K \in \mathcal{R}_r$, and use the Taylor expansion of m . Now consider two kernel functions K_1 and K_2 ; two sequences of bandwidths $h_1 = h_{1n}$ and $h_2 = h_{2n}$; two weight sequences

$$W_{ni}^{(l)}(t) = n^{-1} h_l^{-1} K_l\left(\frac{t - t_{ni}}{h_l}\right), \quad l = 1, 2,$$

and the estimators

$$m_n^{(l)}(t) = \sum_{i=1}^n W_{ni}^{(l)}(t) Y_i, \quad l = 1, 2.$$

The *generalized jackknife estimate* is then defined as

$$G[m_n^{(1)}, m_n^{(2)}](t) = (1 - R)^{-1} [m_n^{(1)}(t) - R m_n^{(2)}(t)]$$

with some constant $R \neq 1$. Note that the generalized jackknife is not based on pseudo-values as is the original jackknife. The relationship of the generalized jackknife to the traditional one is discussed in Gray and Schucany [4, ch. 3]. Since $G[m_n^{(1)}, m_n^{(2)}]$ is a linear combination of two ordinary kernel estimators, we obtain immediately the following proposition.

Proposition 2: Let $h_l \rightarrow 0$ such that $nh_l \rightarrow \infty$, $l = 1, 2$ as $n \rightarrow \infty$. Suppose that $m \in C^p[0, 1]$, $p = 2q$, and let $K_l \in \mathcal{R}_r$, $r = 2s \leq p$, $l = 1, 2$.

Then the bias term of $G[m_n^{(1)}, m_n^{(2)}]$ is

$$(1 - R)^{-1} \sum_{j=3}^q [h_1^2 \Lambda(K_1; 2j) - R h_2^2 \Lambda(K_2; 2j)] m^{(2j)}(t) + o(h^{4s}). \quad (2.1)$$

The bias reduction is now possible by a clever choice of the constant R . For simplicity we will consider for the remainder of this correspondence only the case $p = 4$ and $K_1, K_2 \in \mathcal{R}_2$. The following ideas carry over to the general case. Define

$$R = R_n = \frac{h_1^2 \Lambda(K_1; 2)}{h_2^2 \Lambda(K_2; 2)}$$

Then the coefficient of $m^{(2)}(t)$ in (2.1) is zero, and indeed the bias of $G[m_n^{(1)}, m_n^{(2)}]$ has been reduced compared with the bias of m_n in this situation. (Note that the $o(\cdot)$ term changes to $o(h^{2q})$.) Moreover, the following kernel (depending on n),

$$K^*(u) = \frac{[K_1(u) - rc^3 K_2(cu)]}{[1 - rc^2]}$$

with

$$r = \frac{\Lambda(K_1; 2)}{\Lambda(K_2; 2)}$$

and

$$c = c_n = \frac{h_{1n}}{h_{2n}}$$

could have been used to define the generalized jackknife estimate with $R = rc^2$; that is, in self-explaining notation,

$$m_n(K^*, t) = G[m_n^{(1)}, m_n^{(2)}](t).$$

At first sight the use of $G[m_n^{(1)}, m_n^{(2)}]$ looks like a good strategy. If the experimenter in a first attempt ascribes only a small amount of smoothness to m , i.e., the existence of the second derivative of the regression function, and uses, backed by Proposition 1, a kernel $K \in \mathcal{R}_2$, he might be leaning toward the generalized jackknife estimate for the following reason. If in fact the regression curve is smoother than expected, say, $m \in C^4[0, 1]$, then the estimate $G[m_n^{(1)}, m_n^{(2)}]$, being equivalent to m_n with $K^* \in \mathcal{R}_4$ yields a lower bias. However, a second look at the problem shows that the variance (for fixed n) may be drastically inflated. This is investigated in the following example.

TABLE I

c	$\int K^{*2}(u) du$	$\frac{\int K^{*2}(u) du}{\int K^2(u) du}$
0.1	0.6106	1.017
0.2	0.6383	1.063
0.3	0.6782	1.130
0.4	0.7273	1.212
0.5	0.7833	1.305
0.6	0.8446	1.407
0.7	0.9002	1.517
0.8	0.9792	1.632
0.9	1.05	1.751
0.91	1.058	1.764
0.92	1.065	1.776
0.93	1.073	1.788
0.94	1.08	1.800
0.95	1.087	1.812
0.96	1.095	1.825
0.97	1.1022	1.837
0.98	1.11	1.850
0.99	1.117	1.862

Example: Let $K = K_1 = K_2 \in \mathcal{R}_2$ with

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2), & |u| \leq 1 \\ 0, & |u| > 1, \end{cases}$$

a kernel function considered by Epanechnikov [2]. Straightforward computations show that

$$\int K^2(u) du = \frac{3}{5}$$

$$\Lambda(K; 2) = \frac{1}{10}$$

$$\Lambda(K; 4) = \frac{1}{280}$$

$$\int K^{*2}(u) du = \frac{\frac{9}{10} [c^3 + 2c^2 + \frac{4}{3}c + \frac{2}{3}]}{[c + 1]^2}$$

Table I shows the dependence of $\int K^{*2}(u) du$ on c together with the ratio $\int K^{*2}(u) du / \int K^2(u) du$.

It is apparent from these figures that some caution must be exercised in selecting c . Recall that the selection of c is the same as selecting R or selecting h_1 as a multiple of h_2 . To compare the mse of m_n with the mse of $G[m_n^{(1)}, m_n^{(2)}]$, we equalize the variances by setting $h_1 = (\int K^{*2}(u) du / \int K^2(u) du)h$. The mse of $G[m_n^{(1)}, m_n^{(2)}]$ is then considered as a function of c and h . Let $m^{(2)}(t)/10 = m^{(4)}(t)/280 = 1$ without loss of generality. Then by Propositions 1 and 2 we obtain that the leading bias term B_{1n}

TABLE II^a

h	h_1	$c = 0.1$			$c = 0.2$			$c = 0.3$			$c = 0.4$			$c = 0.5$		
		0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4
0.2	0.2	1.017	0.67	0.51	1.063	0.709	0.532	1.13	0.753	0.565	1.212	0.808	0.606	1.305	0.87	0.652
0.3	0.3	1.52	1.017	0.765	1.59	1.063	0.798	1.695	1.13	0.847	1.818	1.212	0.909	1.958	1.305	0.979
0.4	0.4	2.035	1.357	1.020	2.127	1.418	1.064	2.26	1.507	1.13	2.424	1.616	1.212	2.611	1.74	1.305

TABLE II (continued)

h	h_1	$c = 0.6$			$c = 0.7$			$c = 0.8$			$c = 0.9$		
		0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4
0.2	0.2	1.407	0.938	0.703	1.517	1.011	0.758	1.632	1.088	0.816	1.751	1.167	0.875
0.3	0.3	2.111	1.407	1.055	2.275	1.517	1.137	2.448	1.632	1.224	2.627	1.751	1.313
0.4	0.4	2.815	1.877	1.407	3.034	2.022	1.517	3.264	2.176	1.632	3.503	2.335	1.751

^a mse $\{G[m_n^{(1)}, m_n^{(2)}]\} / \text{mse}\{m_n\}$ for $n = 100$, $\sigma^2 = 1$, $m(t) = \sin t$, $t = \pi/4$.

of M_n and the leading bias term B_{2n} of $G[m_n^{(1)}, m_n^{(2)}]$ for $c = 0.99$ are

$$B_{1n} = h^2 + h^4, \quad B_{2n} = \sqrt{152.76} h^4.$$

This shows that if $h^2 > 1/(\sqrt{152.76} - 1)$ the mse of $G[m_n^{(1)}, m_n^{(2)}]$ dominates the mse of m_n . Similar conditions can be found by varying $m^{(2)}$ and $m^{(4)}$.

In a practical situation a choice of R that avoids a situation of this kind, described in the example, seems to be impossible. Such a selection of R has to take into account the unknown values $m^{(2)}(t)$ and $m^{(4)}(t)$. It is therefore impossible in a practical solution to compute the parameter regions where $G[m_n^{(1)}, m_n^{(2)}]$ actually improves ordinary kernel regression estimate m_n .

We also compared the leading terms of the mse $G[m_n^{(1)}, m_n^{(2)}](t)$ and of the mse $m_n(t)$ of a fixed regression curve in Table II. Shown are the ratios of the two leading terms for different values of h, h_1 , and c with $n = 100$ and $\sigma^2 = 1$. The regression curve $m(t) = \sin t$ was selected, and the mse at $t = \pi/4$ was evaluated with $K \in \mathcal{P}_2$ as before. A bandwidth h , being roughly about 0.3, would minimize the mse of $m_n(t)$; therefore only combinations are shown with $h, h_1 \in \{0.2, 0.3, 0.4\}$. The use of $G[m_n^{(1)}, m_n^{(2)}]$ may result in an mse nearly twice as high as the corresponding mse of m_n as can be seen from the entry $(h, h_1, c) = (0.3, 0.3, 0.9)$ in Table II.

REFERENCES

[1] B. Efron, "The jackknife, the bootstrap and other resampling plans," SIAM publication CBMS-NSF, 1982.
 [2] V. A. Epanechnikov, "Nonparametric estimation of a multivariate probability density," *Theory Prob. Appl.*, vol. 14, pp. 153-158, 1969.
 [3] A. A. Georgiev, "Nonparametric system identification by kernel methods," *IEEE Trans. Automat. Contr.*, vol. 29, pp. 356-358, 1984.
 [4] H. L. Gray and W. R. Schucany, *The Generalized Jackknife Statistic*. New York: Marcel Dekker, 1972.
 [5] L. Györfi, "The rate of convergence of k-NN regression estimate and classification," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 500-509, 1981.
 [6] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.
 [7] M. B. Priestley and M. T. Chao, "Non-parametric function fitting," *J. Roy. Statist. Soc. B*, vol. 34, pp. 385-392, 1972.
 [8] W. R. Schucany and J. P. Sommers, "Improvement of kernel-type density estimators," *J. Amer. Statist. Ass.*, vol. 72, pp. 420-423, 1977.

An On-Line Parameter Estimation Algorithm for Counting Process Observations

PETER SPREIJ

Abstract—The parameter estimation problem for counting process observation is considered. It is assumed that the intensity of the counting process is adapted to the family of σ -algebras generated by the counting process itself and that the intensity depends linearly on some deterministic constant parameters. An on-line parameter estimation algorithm is then presented for which convergence is proved by using a stochastic approximation type lemma.

I. INTRODUCTION

Counting processes frequently occur as observations in mathematical models for industrial processes and in biology, software engineering, and nuclear medicine. Usually, such a

Manuscript received August 1, 1984; revised July 12, 1985. This work was presented in part at the Fourteenth Conference on Stochastic Processes and Their Applications, Gothenburg, Sweden, June 12-16, 1984.

The author is with the Center for Mathematics and Computer Science, P.O. Box 4079, 1009 AB, Amsterdam, The Netherlands.

IEEE Log Number 8406623.

counting process can be considered as the output process of some stochastic system. The underlying state process then influences the counting process. A problem is then to estimate this state, given the observations. This is known as the filtering problem and has been investigated extensively [1].

The solution of this problem requires knowledge of all parameters needed to describe the stochastic system, which means that one can compute the solution to the filtering problem only if one knows the correct parameter values. Unfortunately, in many cases these are not known and therefore need to be estimated. This may happen before the processes start running, using related additional information and/or observations. In the former case some asymptotic results for off-line maximum likelihood estimation are available [3], [4].

The purpose of the present correspondence is to make a contribution to the on-line parameter estimation problem in a specific case. The approach has proven to be fruitful in discrete time ARMAX processes [7] or continuous time Gaussian AR processes [6].

The correspondence is organized as follows. In Section II we give some basic results for counting processes. In Section III we give a heuristic derivation of our parameter estimation algorithm. Section IV contains the convergence proof of the algorithm.

II. PRELIMINARY RESULTS

We assume that we are given a complete probability space (Ω, \mathcal{F}, P) , a time set $T = [0, \infty)$, and a filtration $\{\mathcal{F}_t\}_{t \geq 0}$ satisfying the usual conditions of [2]. All stochastic processes in the sequel are defined on $\Omega \times T$ and adapted to $\{\mathcal{F}_t\}_{t \geq 0}$. We study the case that we are given: an observed process, which is a counting process, that is a map $n: \Omega \times T \rightarrow \mathbf{N}_0$, which has only jumps of magnitude +1. Then it is known [1], [2] that n is a submartingale and therefore admits the so-called Doob-Meyer decomposition (with respect to $\{\mathcal{F}_t\}_{t \geq 0}$)

$$n_t = \Lambda_t + m_t \tag{2.1}$$

where $\Lambda: \Omega \times T \rightarrow \mathbf{R}$ is a predictable increasing process and m a local martingale. Now assume that Λ is an absolutely continuous process, say $\Lambda_t = \int_0^t \lambda_s ds$; then we can rewrite (2.1) as

$$dn_t = \lambda_t dt + dm_t. \tag{2.2}$$

The process λ is called the intensity process.

Often a major problem for counting process observations is to identify the intensity process λ . This problem can be set up in two stages. In the first stage we have to solve a filtering problem. To be precise we have to determine $\hat{\lambda}_t = E(\lambda_t | \mathcal{F}_t^n)$, where $\mathcal{F}_t^n = \sigma\{n_s, s \leq t\}$. Then $\hat{\lambda}_t$ is the optimal (in the sense of mean squared error) estimate given the observations during $[0, t] \subset T$ and given the values of deterministic parameters. We can then replace (2.2) by the minimal decomposition of n (i.e., with respect to $\{\mathcal{F}_t^n\}$)

$$dn_t = \hat{\lambda}_t dt + d\bar{m}_t, \tag{2.3}$$

where \bar{m} is a local martingale adapted to $\{\mathcal{F}_t^n\}_{t \geq 0}$. In the second stage one looks for estimates of remaining unknown deterministic parameters. If one adopts the maximum likelihood criterion, (2.3) and the computation of $\hat{\lambda}_t$ appear to be crucial. The likelihood functional in this case is known [1, p. 174] to be

$$L_t = \exp \left[- \int_0^t (\hat{\lambda}_s - 1) ds + \int_0^t \log \hat{\lambda}_s - dn_s \right]. \tag{2.4}$$

The Model

From here on we assume that $\hat{\lambda}$ has a special structure

$$\hat{\lambda}_t = p^T \phi_t, \tag{2.5}$$

where $p \in \mathbf{R}^m$ is the vector of unknown parameters and $\phi: \Omega \times T \rightarrow \mathbf{R}^m$ is a process adapted to $\{\mathcal{F}_t^n\}_{t \geq 0}$ and thus known. Indeed (2.5) imposes a restrictive condition on the intensity

SOME THEORY ON M -SMOOTHING OF TIME SERIES

BY WOLFGANG HÄRDLE

Universität Heidelberg, Heidelberg

AND

PHAM-DINH TUAN

Université de Grenoble, Grenoble

Abstract. In recent years many robust smoothing procedures for time series have been introduced. Their extreme nonlinearity made them mathematically untractable and their behaviour was mostly analysed by means of Monte Carlo studies. In this paper we develop some mathematical theory of a specific class of nonlinear smoothers. We investigate the asymptotics of so-called M -smoothers and discuss robustness of M -smoothers in some special cases.

Keywords. Resistant smoothing; robust time series analysis; M -estimation; robust filters.

1. INTRODUCTION

A time series $\{Y_i\}$, $-N \leq i \leq N$ with a deterministic trend $\{\mu_i\}$, $-N \leq i \leq N$ is a sequence of real data

$$Y_i = \mu_i + Z_i, \quad (1.1)$$

where $\{Z_i\}$, $-N \leq i \leq N$ represents a zero-mean noise process. A smoothing procedure is any algorithm operating on $\{Y_i\}$ to produce an estimate $\{S_i\}$ of the trend $\{\mu_i\}$. One approach to estimate the trend would be to assume that $\{\mu_i\}$ is of parametric form, i.e.,

$$\mu_i = f(i; \theta)$$

with some known function f and an unknown parameter θ . For instance $f(i; \theta)$ could be a polynomial with θ representing the unknown coefficients.

In contrast, we pursue in this paper a nonparametric approach: Beside smoothness assumptions on $\{\mu_i\}$ we do not assume any parametric form of the trend $\{\mu_i\}$. The trend function may be any nonlinear curve.

We consider here, robust M -filters (' M -smoothers') $\{S_i\}$, as described in Mallows (1980), p. 711, question (iii) which are solutions of

$$\sum_j \alpha_j \psi(Y_{i-j} - S_i) = 0 \quad (1.2)$$

where $\{\alpha_j\}$ is a sequence of weights (filter coefficients). The nonlinear function ψ makes the smoother $\{S_i\}$ robust.

A function ψ that is, beside other qualities, bounded yields a robust smoother since large observations are downweighted. The unbounded ψ -function $\psi(u) \equiv u$ yields the linear smoother $S_i = \sum_j \alpha_j Y_{i-j}$ whenever the weights $\{\alpha_j\}$ sum up to 1. The last smoother is extremely sensitive to the presence of occasional outliers since it operates like a local average ('running mean') on the original time series $\{Y_i\}$.

A well-known example of a ψ -function is Huber's ψ :

$$\psi(y) = \max(-\kappa, \min(y, \kappa)), \quad \kappa \geq 0. \quad (1.3)$$

A value of κ equal to zero yields the running median (Tukey, 1977)

$$S_i = \text{med} \{Y_{i-\delta}, \dots, Y_i, \dots, Y_{i+\delta}\},$$

whereas a large parameter κ , yields a smoother acting like a linear smoother.

Mallows (1980) investigated the properties of nonlinear smoothers under the following basic specification of the observed time series

$$Y_i = \mu + G_i + Z_i$$

when $\{G_i\}$ is a zero-mean stationary Gaussian process with a prescribed covariance function and where $\{Z_i\}$ is a sequence of independent random variables having an arbitrary common distribution. Here we take a different route: the basic idea is somewhat similar to nonparametric regression function estimation with equally spaced design points. We consider the observed time series $\{Y_i\}$ as embedded in a sequence of time series $\{Y_i^{(n)}\}$ that we sample finer and finer as n tends to infinity. Equivalently, we assume that the observed time series $\{Y_i\}$ is sampled from a continuous process on a compact interval in such a way that the sampling frequency *increases*, as the model index n tends to infinity. We also investigate the properties of M -smoothers in the situation where the trend $\{\mu_i\}$ tends to a constant, as $n \rightarrow \infty$. In this approach we obtain asymptotic results by sampling more and more with a *fixed* sampling frequency.

In the present framework we do not compare linear and nonlinear smoothers on the basis of their 'transfer functions' or similar means. In our setting, by embedding the time-series $\{Y_i\}$ into a sequence $\{Y_i^{(n)}\}$, the 'transfer functions' or impulse response coefficients $\{\alpha_j^{(n)}\}$ will become smooth continuous functions, as $n \rightarrow \infty$. Robustness properties and the degree of nonlinearity of the M -smoothers will then be seen from these asymptotic quantities. Once these indices have been derived, we can proceed to measure the degree of resistance of outliers and to find optimal robust smoothers among a class of possible smoothers.

The results of this paper address two questions raised by Mallows (1980). In his questions (iii) and (iv), p. 711, Mallows defined the M -smoothers analogously to the class of M -estimators, introduced by Huber (1964). He then asked if these M -smoothers, as defined in (1.2), have any merits such as robustness or asymptotic minimax optimality. It is seen here that the M -smoothers consistently estimate the trend and have some of the desired robustness properties. Moreover, in the simple situation where $\{Z_i\}$ is white noise, we obtain the same minimax result as Huber (1964). That is, if we model the existence of outliers in the noise process

by the following family of distribution functions

$$\mathcal{G}_\gamma = \{F: F = (1 - \gamma)\Phi + \gamma H\}$$

where Φ is the standard Gaussian distribution function, and H is arbitrary symmetric distribution, we can construct an optimal M -smoother which is minimax in a certain sense. In contrast to Mallows's setup of the robust smoothing problem, we are considering *replacing* noise whereas he is investigating *additive* noise (see his comment (iv), p. 711).

The outline of the paper is organized as follows:

- (a) Description of the robust smoothing problem and the introduced setup.
- (b) Consistency and asymptotic distribution of the M -smoothers.
- (c) Asymptotic optimality in a certain class of smoothers.
- (d) Remarks and open questions.

Some basic results are expressed in theorems 3.1-3.2 below. We find that when a M -smoother is applied to $\{Y_i^{(n)}\}$, then, as $n \rightarrow \infty$, the trend is consistently estimated. Furthermore, theorem 3.2, stating the asymptotic normality of M -smoothers, allows us to construct (pointwise) confidence intervals for the estimated trend function. We were also able to find optimal M -smoothers in a special case, but we are not completely happy with that optimality statement since it is based on a white noise assumption.

There are possibly ways to compute optimal M -smoothers in a more general setup, as the work of Portnoy (1977) indicates, but we were unable to find them. A discussion of this point together with some open questions is found in sections 4 and 5.

2. THE ROBUST SMOOTHING PROBLEM

The parametric approach to robust smoothing of time-series is to assume that the trend is of parametric form $\mu_i = f(i, \theta)$ with a parameter θ . In a next step, one would like to find robust estimates of θ on the basis of the observed time series $\{Y_i = f(i, \theta) + Z_i\}$. This approach was taken by Velleman (1980), who took $f(i, \theta)$ of sinusoidal form and added as the noise process $\{z_i\}$ independent Gaussian random variables with intermittent outliers. However, from Velleman's Monte Carlo study it is not clear how different smoothers perform in the presence of correlated noise. A disadvantage of the parametric setup is that we have to know the (parametric) form of f otherwise an estimation of the trend is not possible.

Mallows (1980) took a different approach, similar to that used in signal detection theory. In his basic specification of the observed time series he assumed a constant mean $\{\mu_i \equiv \mu\}$ and an additive noise structure, i.e., the noise process is a sum of a stationary zero-mean Gaussian process and a 'wild' sequence of independent random variables. He then studied nonlinear smoothers like '53H', '53H twice' and was interested in estimating the Gaussian signal. There are many examples where the assumption of a non-constant deterministic trend seems to

be more appropriate. For instance, the daily maximum 1-hour average ozone concentration as reported by Horowitz (1980, fig. 1), or the United Kingdom exports graphed in Brillinger (1975, fig. 1.1.4).

We take the following approach to the robust smoothing problem. We assume that the observed time series Y_i is sampled from some continuous process $X(t)$, sum of a trend function $\mu(t)$ and a noise process $V(t)$. Suppose that we sample $X(t)$ more and more with a fixed sampling interval $\Delta = t_i - t_{i-1}$, then it is clear that we cannot expect a classical filter or a robust smoother to yield a consistent estimate of the trend at a given time point t_0 , unless the trend function is constant, since values of $X(t)$ at t far from t_0 would provide no information on $\mu(t_0)$. On the other hand, if we sample in a compact interval finer and finer with a sampling interval $\Delta = t_i - t_{i-1} \rightarrow 0$, then we would not achieve consistency if the noise process has a fixed dependency structure since $X(t_{i-1})$ would be highly dependent of $X(t_i)$ and provide no more information on $\mu(t_i)$ as $\Delta \rightarrow 0$. To overcome this difficulty, we consider the following model. We assume that our time series is a member of a sequence of time series of the form

$$Y_i^{(n)} = X^{(n)}(t_i) = m(t_0 + ic_n) + Z_i, \quad -N < i < N \tag{2.1}$$

where m is a smooth function and $c_n \rightarrow 0$ as $n \rightarrow \infty$. The sample size $2N + 1$ is also assumed to depend on n and tends to infinity with n . The model can represent both situations described above. In the first situation when we have a very large sample with a fixed sampling interval, we would assume that the trend function varies so slowly that we may represent it by $\mu(t_i) = m(t_0 + ic_n)$ where n is a large number (c_n small) and m is a smooth function. In the second situation when we have sampled very finely in a compact interval, we would assume that the noise process $V(t)$ is so weakly dependent that we may represent it by $V(t_i) = U(t_0 + (t_i - t_0)/c_n)$ where n is a large number. If we assume further that the sampling interval is proportional to c_n then of course the trend function would be of the form $m(t_0 + ic_n)$ and hence by putting $Z_i = U(t_0 + i)$, the time series may be considered as a member of the sequence $\{X^{(n)}(t_i) = m(t_0 + ic_n) + Z_i, -N < i < N\}$.

In the following we shall consider the model (2.1) and we shall investigate asymptotic properties of the M -smoother as $n \rightarrow \infty$. Note that the sample length increases to infinity with n and since this sample length does not play any apparent role, the index N will be dropped. The M -smoother will be denoted by $S_i^{(n)}$ and is defined as the solution of the equation

$$\sum_i \alpha_j^{(n)} \psi(Y_{i-j}^{(n)} - S_i^{(n)}) = 0 \tag{2.2}$$

where $\{\alpha_j^{(n)}\}$ is a sequence of weight functions and ψ is the function introduced before.

3. SOME ASYMPTOTIC RESULTS

Here we consider the consistency and the asymptotic distribution of the M -smoother $S_0^{(n)}$ defined by (2.2). We restrict ourselves without loss of generality

to the estimation of $\mu_0 = m(t_0)$. The time series is assumed to obey the model (2.1), moreover, we introduce the following assumptions.

ASSUMPTION 1. The trend function $m(t)$ is twice continuously differentiable.

ASSUMPTION 2. The noise process $Z(i)$ is a linear process, that is

$$Z(i) = \sum_{j=-\infty}^{\infty} a_j \varepsilon_{i-j}$$

with

$$\sum_{j=-\infty}^{\infty} |j| |a_j| < \infty$$

and $\{\varepsilon_j\}$ are independent identically distributed random variables with a symmetric distribution having a finite (absolute) first moment.

ASSUMPTION 3. There exists a sequence $\{b_n\}$ of positive real numbers tending to zero as $n \rightarrow \infty$ such that with $M_n = b_n / c_n$

$$\lim_{n \rightarrow \infty} M_n = \infty,$$

and the filter coefficients can be represented as

$$\alpha_j^{(n)} = M_n^{-1} K(j/M_n)$$

where K is a positive, piecewise continuous function satisfying

$$\int K(t) dt = 1$$

$$\text{supp } \{K\} \subset [-A, A], \quad A > 0.$$

ASSUMPTION 4. The ψ -function, defining the M -smoother through (2.7) is nondecreasing, bounded, continuously differentiable and antisymmetric.

Assumption 4 on the ψ -function establishes the robustness of $S_0^{(n)}$. Similar assumptions on ψ are made in the theory of robust estimation of location (Huber, 1981, chapter 4). Assumption 3 on the representation of the filter coefficients ensures that the sequence of piecewise constant function taking value $\alpha_j^{(n)}$ in the interval $[(j-.5)/M, (j+.5)/M]$, up to a homothetical transformation, behaves asymptotically, as $n \rightarrow \infty$, like the piecewise continuous function K . The scaling parameter M_n here plays the role of the *span* $sp(S_0^{(n)})$ of the smoother $S_0^{(n)}$; as defined in Mallows (1980, p. 701). In contrast to Mallows' assumptions, our embedding procedure makes $sp(S_0^{(n)})$ dependent on the model index n .

The following main results will be provided by a couple of lemmata. The cumulative distribution function of $\{Z(i)\}$ will be denoted by F .

THEOREM 3.1. Assume that assumptions 1 to 4 hold, then, as $n \rightarrow \infty$,

$$S_0^{(n)} \xrightarrow{P} \mu_0.$$

The next theorem proves the asymptotic normality of $S_0^{(n)}$.

THEOREM 3.2. Under assumptions 1 to 4 and the additional requirement that

ψ' is Lipschitz continuous,

then as $n \rightarrow \infty$,

$$\sqrt{M_n} \left[S_0^{(n)} - \mu_0 + B_0^{(n)} / \left(\int \psi'(z) dF(z) \right) \right],$$

$$B_0^{(n)} = M_n^{-1} \sum_j K(j/M_n) E_F \psi(Y_j - \mu_0),$$

converges in distribution to a normal variate with mean zero and variance

$$\int K^2(t) dt \sum_K R_\psi(k) / \left[\int \psi'(z) dF(z) \right]^2,$$

where

$$R_\psi(k) = \text{cov} \{ \psi(Z(0)), \psi(Z(k)) \}.$$

LEMMA 3.1. Let

$$H_n(t_0, s) = \sum_j M_n^{-1} K(j/M_n) \psi(Y_j - s)$$

$$H(t_0, s) = \int \psi[m(t_0) + z - s] dF(z),$$
(3.1)

where $F(z)$ is the cumulative distribution function of $\{Z(j)\}$. Then under assumptions 1 to 4, as $n \rightarrow \infty$, for all $s \in \mathbb{R}$

$$\lim_{n \rightarrow \infty} EH_n(t_0, s) = H(t_0, s).$$

PROOF. It is clear that

$$EH_n(t_0, s) = M_n^{-1} \sum_j K(j/M_n) H(t_0 + jc_m, s).$$

Now, by Lebesgue's dominated convergence theorem, the boundedness of ψ and the continuity of m and ψ , it is easy to see that $H(t, s) \rightarrow H(t_0, s)$ as $t \rightarrow t_0$. Hence, as $n \rightarrow \infty$,

$$M_n^{-1} \sum_j K(j/M_n) [H(t_0 + jc_m, s) - H(t_0, s)] \rightarrow 0.$$

Since $\lim_{n \rightarrow \infty} M_n^{-1} \sum_j K(j/M_n) = \int K(t) dt = 1$ by assumption 3 the assertion (3.1) follows.

LEMMA 3.2. *Let*

$$R(k, s) = \text{cov} \{ \psi(Z(j) + s), \psi(Z(j+k) + s) \}, \tag{3.2}$$

then under assumptions 1 to 4 as $n \rightarrow \infty$, for all $s \in \mathbb{R}$

$$M_n \text{ var} \{ H_n(t_0, s) \} \rightarrow \int K^2(t) dt \sum_{k=-\infty}^{\infty} R(k, \mu_0 - s). \tag{3.3}$$

PROOF. The proof of this lemma is very technical since covariance functions do not commute with nonlinear functions such as ψ . The argument is based on the mean value theorem, together with a split-up technique applied to $\{Z(i)\}$ and a careful analysis of remainder terms.

We fix s and define $T(j) = \psi(Y_j - s)$. We then have

$$\begin{aligned} M_n \text{ var} \{ H_n(t_0, s) \} &= M_n^{-1} \sum_j \sum_k K(j/M_n) K((j+k)/M_n) \text{cov} \{ T(j), T(j+k) \} \\ &= \sum_k c_{n,k}, \quad \text{say.} \end{aligned} \tag{3.4}$$

Define now

$$Z_p(j) = \sum_{|u| \leq p} a_u \varepsilon_{j-u}, \quad r_p(j) = Z(j) - Z_p(j), \tag{3.5}$$

and apply the mean value theorem to the function ψ to obtain with (3.5)

$$\begin{aligned} T(j) &= \psi[m(t_0 + jc_n) + Z_p(j) - s] + r_p(j) \cdot \psi'(\xi_j) \\ T(j+k) &= \psi[m(t_0 + (j+k)c_n) + Z_p(j+k) - s] + r_p(j+k) \cdot \psi'(\xi_{j+k}) \end{aligned} \tag{3.6}$$

where $\psi'(\xi_j)$ is for all i a bounded random variable.

Now by assumption 2 on the noise process, the first summand on the right-hand sides of (3.6) are independent for $p \leq k/2, k > 1$. Hence

$$\begin{aligned} \text{cov} \{ T(j), T(j+k) \} &= E[r_p(j) \psi'(\xi_j) T(j+k)] \\ &\quad + E\{ \psi[m(t_0 + jc_n) + Z_p(j) - s] r_p(j+k) \psi'(\xi_{j+k}) \}. \end{aligned}$$

Since

$$E|r_p| \leq \left[\sum_{|u| \geq p} |a_u| \right] E|\varepsilon_0|,$$

and since the function ψ and the random variable $\psi'(\xi_i), T(j+k), \psi'(\xi_{j+k})$ are bounded, by taking $p = k/2$ if k is even, $p = (k+1)/2$ otherwise, we get

$$|\text{cov} \{ T(j), T(j+k) \}| \leq C_1 \sum_{|u| \geq k/2} |a_u|$$

for some constant C_1 . Since the function K is bounded and has compact support, we deduce from (3.4) that for some constant C_2

$$|c_{n,k}| \leq C_2 \sum_{|u| > k/2} |a_u| = c_k, \quad \text{say.}$$

Since

$$\sum c_k \leq C_2 \sum_u [4|u| + 1] |a_u| < \infty$$

by assumption 2, by Lebesgue dominated convergence theorem it remains only to compute for fixed k the limit of the $c_{n,k}$, as $n \rightarrow \infty$. This follows from the continuity of m and ψ (assumptions 1 and 4) which ensures that

$$\lim_{n \rightarrow \infty} \text{cov} \{T(j), T(j+k)\} = R(k, \mu_0 - s).$$

So finally by the continuity of K we have that, as $n \rightarrow \infty$,

$$\begin{aligned} c_{k,n} &= M_n^{-1} \sum_j K(j/M_n) K((j+k)/M_n) \text{cov} \{T(j), T(j+k)\} \\ &\rightarrow \int K^2(t) dt R(k, \mu_0 - s), \end{aligned}$$

which shows (3.3) completing the proof of the lemma. \square

The proof of theorem 3.1 follows now immediately from lemma 3.1 and lemma 3.2. Both lemmas together state that, as $n \rightarrow \infty$, $H_n(t_0, s) \xrightarrow{P} H(t_0, s)$, theorem 1 thus follows by monotony of ψ and symmetry of F .

For the proof of theorem 3.2 we need the following lemma.

LEMMA 3.3. *Suppose that assumptions 1 to 4 hold. Then, as $n \rightarrow \infty$, for all s*

$$\sqrt{M_n} [H_n(t_0, s) - EH_n(t_0, s)]$$

converges in distribution to a normal variate with zero mean and variance

$$\int K^2(t) dt \sum_k R(k, \mu_0 - s).$$

PROOF. Recall the definition of $Z_p(j)$ from (3.5) and (3.6) respectively and set for fixed s

$$T_p(j) = \psi(m(t_0 + jc_n) + Z_p(j) - s)$$

and

$$U_p(j) = T(j) - T_p(j).$$

By the same argument used in the proof of lemma 3.2 we obtain that with a constant C_4

$$EU_p^2(j) \leq C_4 \left[\sum_{|u| \geq p} a_u^2 \right].$$

Let p be an even integer and let us split up

$$M_n^{1/2} H_n(t_0, s) = M_n^{-1/2} \sum_j K(j/M_n) T_{p/2}(j) + M_n^{-1/2} \sum_j K(j/M_n) U_{p/2}(j). \quad (3.7)$$

We will show now that the variance of the last term of the right-hand side of (3.7) converges uniformly in n to zero, as $p \rightarrow \infty$, and then we claim that for fixed p , as $n \rightarrow \infty$,

$$M_n^{-1/2} \sum_j K(j/M_n)[T_{p/2}(j) - ET_{p/2}(j)] \tag{3.8}$$

converges in distribution to a normal random variable with mean zero and variance σ_p^2 , where

$$\lim_{p \rightarrow \infty} \sigma_p^2 = \int K^2(t) dt \sum_k R(k, \mu_0 - s).$$

Once these two claims are proved, the result follows from Bernstein's lemma, as stated in Hannan (1970, p. 242). The variance of the second term on the right-hand side of (3.7) is equal to

$$\sum_k \left[M_n^{-1} \sum_j K(j/M_n)K((j+k)/M_n) \text{cov} \{U_{p/2}(j), U_{p/2}(j+k)\} \right].$$

Note that $U_{p/2} \leq 2 \sup |x\psi(x)|$ for all p . Hence, by the same argument as in the proof of lemma 3.2, there exists a constant C_3 such that

$$|\text{cov} \{U_{p/2}(j), U_{p/2}(j+k)\}| \leq C_3 \left[\sum_{|u| \geq p/2} |a_u| \right]. \tag{3.9}$$

For the range $k \geq p$ we will use another bound. Let $p/2 \leq q \leq k/2$, we have:

$$\begin{aligned} U_{p/2}(j) &= U_q(j) + [T_q(j) - T_{p/2}(j)] \\ U_{p/2}(j+k) &= U_q(j+k) + [T_q(j+k) - T_{p/2}(j+k)] \end{aligned}$$

by definition of $U_q(j)$ and $T_q(j)$. Since the terms inside the square brackets are independent by construction and since the random variables $T_k(j)$ are bounded,

$$|\text{cov} \{U_{p/2}(j), U_{p/2}(j+k)\}| \leq C_4 E[|U_q(j)| + |U_q(j+k)|] \leq C_5 \sum_{u=q} |a_u|$$

where C_4, C_5 are constant. Since this inequality holds for all q in the range $p/2 \leq q \leq k/2$ we have that

$$|\text{cov} \{U_{p/2}(j), U_{p/2}(j+k)\}| \leq C_6 \left[\sum_{|u| \leq k/2} |a_u| \right]. \tag{3.10}$$

Now from (3.10) and (3.9) we may conclude

$$\begin{aligned} \text{var} \left\{ M_n^{-1/2} \sum_j K(j/M_n) U_{p/2}(j) \right\} &\leq C_7 \sum_{|k| \geq p} \sum_{u > p/2} |a_u| + C_8 \sum_{|k| < pu \leq k/2} |a_u| \\ &\leq C_9 \sum_{|u| \geq p/2} [4|u| + 1] |a_u|. \end{aligned}$$

for some constants C_7, C_8, C_9 . This tends, as $p \rightarrow \infty$, to zero by assumption 2.

We now show that (3.8) has a normal limit. For this let $r \geq p$ a given integer and group the terms in (3.8) into blocks

$$M_n^{-1/2} \sum_{k \neq 0} B_k + M_n^{-1/2} \sum_k B'_k$$

where B'_0 in the sum of terms in (3.8) with subscripts ranging from $-p+1$ to $p-1$. Similar B_k , $k \geq 1$ denotes the sum of terms with indices from $k(r+p)-r$ to $k(r+p)-1$, whereas B'_k , $k \geq 1$ are the blocks ranging from $k(r+p)$ to $k(r+p)+p-1$. B_{-k} and B'_{-k} are defined in the same way with $-j$ playing the role of j . By construction of the blocks B'_k it is evident that the different B'_k are independent and thus

$$\text{var} \left\{ M_n^{-1/2} \sum_k B'_k \right\} = M_n^{-1} \sum_k \text{var} \{ B'_k \}. \tag{3.11}$$

Now since K has compact support by assumption 3, $B'_k = 0$ for $|k| > AM_n/(r+p)$. The boundedness of K and $T_{p/2}(j)$ yields now

$$|B'_k| \leq C_{10}p, \quad |B'_0| \leq 2C_{10}p,$$

since B'_0 is a block of length $2p-1$. Hence the right-hand side of (3.11) is bounded by

$$M_n^{-1} 2C_{10}^2 p^2 A(M_n+2)/(r+p)$$

which tends to zero, uniformly in n , as $r \rightarrow \infty$.

To prove that $M_n^{-1/2} \sum_{k \neq 0} B_k$ is asymptotically normal, we employ the Lindeberg condition

$$\lim_{n \rightarrow \infty} M_n^{-1} \sum_{k \neq 0} EB_k^2 I(B_k^2 > \varepsilon/M_n) = 0. \tag{3.12}$$

By the same compactness argument as above

$$B_k = 0 \quad \text{for } |k| > AM_n/(r+p)$$

and

$$|B_k| \leq C_{10}r, \quad k \neq 0,$$

which ensures that (3.12) holds. On the other hand

$$\begin{aligned} \text{var} \left\{ M_n^{-1/2} \sum_{k \neq 0} B_k \right\} &= \sum_{|i| \leq p} M_n^{-1} \left\{ \sum_{j \in I_i} K(j/M_n) K((j+i)/M_n) \right. \\ &\quad \left. \times \text{cov} \{ T_{p/2}(j), T_{p/2}(j+i) \} \right\} \end{aligned} \tag{3.13}$$

where I_i is the set of subscripts j such that both i and j belong to the set

$$\bigcup_{k=1}^{\infty} [\{kr+kp-r, \dots, kr+kp-1\} \cup \{1-kr-kp, \dots, r-kr-kp\}].$$

By the same continuity argument as in the proof of lemma 3.2, we see that (3.13), as $n \rightarrow \infty$, converges to

$$r/(r+p) \int K^2(t) dt \sum_{|k| \leq p} R_{p/2}(k, \mu_0 - s).$$

This term itself has $\sigma_p^2 = \int K^2(t) dt \sum_{|k| \leq p} R_{p/2}(k, \mu_0 - s)$ as a limit, as $r \rightarrow \infty$.

Therefore from Bernstein's lemma, the quantity in (3.8) is asymptotically normally distributed with mean zero and variance σ_p^2 . Clearly $R_{p/2}(k, \mu_0 - s) \rightarrow R(k, \mu_0 - s)$, as $p \rightarrow \infty$. An application of Lebesgue's dominated convergence theorem thus completes the proof of lemma 3.3. \square

The proof of the next lemma is straightforward

LEMMA 3.4. Under the conditions of theorem 3.1, as $n \rightarrow \infty$,

$$\frac{\partial}{\partial s} H_n(t_0, s) \xrightarrow{p} - \int \psi'(\mu_0 + z - s) dF(z).$$

PROOF OF THEOREM 3.2. The proof of theorem 3.2 follows now directly from

$$\begin{aligned} 0 &= H_n(t_0, S_0^{(n)}) \\ &= H_n(t_0, \mu_0) + (S_0^{(n)} - \mu_0) \cdot \frac{\partial}{\partial s} H_n(t_0, \mu_0 - s)|_{s=\xi^{(n)}} \end{aligned}$$

where $|\xi^{(n)} - \mu_0| < |S_0^{(n)} - \mu_0|$. Now by Lipschitz continuity of ψ' , theorem 3.1 and lemma 3.4, as $n \rightarrow \infty$,

$$\frac{\partial}{\partial s} H_n(t_0, \mu_0 - s)|_{s=\xi^{(n)}} \xrightarrow{p} - \int \psi'(z) dF(z).$$

Hence

$$S_0^{(n)} - \mu_0 = H_n(t_0, \mu_0) \int \left[\int \psi'(z) dF(z) \right] + o_p(1)$$

and theorem 3.2 follows from lemma 3.3.

REMARK. If K is symmetric and ψ'' exists then it is easily seen that

$$B_0^{(k)} = b_n^2 \int K(t) t^2 dt \int [\psi''(z) m'(t_0) \psi'(z) m''(t_0)] dF(z) + o(b_n^2).$$

The asymptotic bias is thus of the order $O(b_n^2)$.

4. THE ROBUSTNESS OF M-SMOOTHERS

In the previous sections we have introduced and discussed the M -smoother $S_0^{(n)}$ with asymptotic variance

$$\beta_k [V_H(\psi, F) + V_c(\psi, F)] \tag{4.1}$$

where

$$\beta_k = \int K^2(t) dt$$

$$V_H(\psi, F) = E_F \psi^2(z) / [E_F \psi'(z)]^2$$

$$V_c(\psi, F) = \sum_{k \neq 0} E_F \{ \psi(Z(0)) \psi(Z(k)) \} / E_F \psi'(z)^2.$$

The factor β_k occurs also in the asymptotic variance of linear smoothers, the summand $V_H(\psi, F)$ is exactly the variance of Huber's M -estimates for location

(Huber, 1981) and $V_c(\psi, F)$ is the summand involving the correlation structure of the noise process $Z(i)$. Suppose now that $Z(i)$ is white noise. Then $V_c(\psi, F) = 0$ and therefore the theory of robust estimation of location applies since β_k is independent of ψ and F . More precisely, once we decide to take the asymptotic variance of $S_0^{(n)}$ to measure the performance of M -smoothers, we have the same minimax results as in the robust estimation of location (Huber, 1981, chapter 4), provided $\{Z(i)\}$ is white noise. In this sample case, optimal M -smoothers can be designed by the following device. First decide on the degree of smoothness of the trend, i.e., choose such a linear filter $\{\alpha_i^{(n)}\}$ which reproduces a polynomial of a certain order and minimizes β_k . Secondly decide on the class of contaminations $\mathcal{G}_\gamma = \{F: F - (1 - \gamma)\Phi + \gamma H\}$ and determine an optimal ψ -function as in Huber (1981). Suppose for instance we deduced that $\{\alpha_i^{(n)}\}$ should pass a linear trend. We thus have to solve the following optimization problem

$$\int K^2(t) dt = \min!$$

subject to

$$\int K(t)t dt = 0$$

$$\int K(t)t^2 dt = 1.$$

The last condition was introduced to keep the bias constant (see the remark after the proof of theorem 3.2). This problem was solved by Rosenblatt (1971) giving the following function K

$$K(t) = \begin{cases} (3/4)(1 - t^2), & |t| \leq 1 \\ 0, & \text{elsewhere.} \end{cases}$$

This gives us the 'linear part' of the asymptotic variance of the M -smoother. The 'nonlinear part' of the asymptotic variance is in the case simply $V_H(\psi, F)$, which gives us for the contamination model (with *replacing* outliers) Huber's ψ -function, as defined in (1.3),

$$\psi(y) = \max \{-\kappa, \min \{y, \kappa\}\},$$

where κ is related to the contamination rate γ by

$$(1 - \gamma)^{-1} = 2\kappa^{-1}\Phi(\kappa) + 2\Phi(\kappa) - 1.$$

The case where $\{Z(i)\}$ is *not* white noise or equivalently, $V_c(\psi, F) \neq 0$, is more delicate since it is not apparent how optimal M -smoothers can be found. First of all, it is not clear how an outlier model for $\{Z(i)\}$ can be formulated. One approach could be taken in analogy to Steve Portnoy's work (Portnoy, 1977) and to describe the presence of outliers through a contamination model \mathcal{G}_γ for the *innovation process*. Outliers of this kind may occur in practice, the graph of the United Kingdom exports (Brillinger, 1975, fig. 1.1.4) at the end of the year 1967

seems to indicate an outlier of the kind described above. However, the application of Portnoy's work to real problems is difficult, since he considered a noise process $Z(i) = a_{-1}\varepsilon_{i-1} + a_0\varepsilon_i + a_1\varepsilon_{i+1}$, and derived results under the assumption that $|a_{-1}| = |a_1|$ is very small which may not be true in practice.

Huber (1964) showed that $V_H(\psi, F)$ is convex in ψ and will thus have a unique minimum ψ_0 over the class of ψ -functions for which $V_H(\psi, F)$ is finite. Hence, if $V_c(\psi, F)$ is small, we could expect a similar result as for the robust M -estimates of location, but we were unable to compute such an optimal ψ_0 or a 'least favorable' F_0 such that $V = V_H + V_c$ enjoys the well-known saddlepoint property

$$\min_{\psi} V(\psi, F_0) = V(\psi_0, F_0) = \max_F V(\psi_0, F).$$

5. SOME REMARKS AND OPEN QUESTIONS

From the calculations of section 4, considering the robustness properties of M -smoothers, it is clear that further research will be necessary to completely understand the different phenomena affecting a nonlinear smoother. We have presented in the previous section some investigations when the noise process is white. However, there are several questions of practical and theoretical nature, on which some progress should be made.

QUESTION 1. Throughout this paper we assumed that the process $\{Y_i\}$ is unstationary because $\{Y_i\}$ was defined to be a sum of a deterministic nonlinear trend function and a stationary noise process $\{Z(i)\}$. This addresses discussion point (i) of Mallows (1980, p. 710) but still leaves the following question open. How do M -smoothers react against *unstationary* noise?

QUESTION 2. Is it possible to show in the class of M -smoothers an asymptotic minimax result for $V(\psi, F) = V_H(\psi, F) + V_c(\psi, F)$?

QUESTION 3. Is there a selection rule which allows to choose the span M_n of the M -smoother in some (asymptotically) optimal way?

QUESTION 4. Suppose we pre-smooth a time-series $\{Y_i\}$ to arrive at an estimated trend function $S_i^{(n)}$ and construct the estimated residual series $\hat{R}_i = Y_i - S_i^{(n)}$. Then perform a whitening of \hat{R}_i yielding \hat{W}_i and apply now a robust smoother to

$$\hat{Y}_i = S_i^{(n)} + \hat{W}_i.$$

Does this post-whitening have any merits?

ACKNOWLEDGEMENT

This work has been supported in part by the Deutsche Forschungsgemeinschaft Sonderforschungsbereich 123 "Stochastische Mathematische Modelle."

REFERENCES

- BRILLINGER, D. R. (1975) *Time Series Data Analysis and Theory*. Holt, Rinehart and Winston, Inc.
- HANNAN, E. J. (1970) *Multiple Time Series*. Wiley: New York.
- HOROWITZ, J. (1980) Extreme Values from a Nonstationary Stochastic Process: In Application to Air Quality Analysis. *Technometrics* 22, 469-482.
- HUBER, P. J. (1964). Robust Estimation of a Location Parameter. *Ann. Math. Stat.* 35, 73-101.
- HUBER, P. J. (1981) *Robust Statistics*. J. Wiley & Sons: New York.
- MALLOWS, C. L. (1980) Some Theory of Nonlinear Smoothers. *Ann. Stat.* 8, 695-715.
- PORTNOY, S. L. (1977) Robust Estimation in Dependent Situations. *Ann. Stat.* 5, 22-43.
- ROSENBLATT, M. (1971) Curve Estimates. *Ann. Math. Stat.* 42, 1815-1842.
- TUKEY, T. W. (1977) *Exploratory Data Analysis*. Addison-Wesley: Reading, Mass.
- VELLEMAN, P. F. (1980) Definition and Comparison of Robust Nonlinear Data Smoothing Algorithms. *J. Amer. Stat. Ass.* 75, 609-615.

Fetal cerebral function and intrauterine hypoxia in sheep fetuses

J. F. H. GAUWERKY¹, K. WERNICKE¹, T. HÖLTING¹, P. MATTHIS², W. HÄRDLE³ and F. KUBLI¹

¹Department of Obstetrics and Gynecology, ²Department of Neuropediatrics and ³Special Research Department 123, University of Heidelberg, *Vofßstraße 9, D-6900 Heidelberg, FRG*

Abstract. The fetal EEG was studied in seven chronically prepared sheep fetuses (gestational age 115–120 days) under different degrees of hypoxia. The EEG was evaluated by spectrum analysis. Hypoxia was induced by clamping the common hypogastric artery. During normoxia a cyclic high voltage – low voltage (HV-LV) pattern with typical frequency shifts occurred. First sign of mild hypoxia was a shortening of the HV-LV cycle. Further increasing hypoxia caused a reduction of mean power, especially in the HV-phase. Reduction of pO_2 below 16 mmHg resulted in a loss of the cyclic changes of the fetal EEG. An increasing pO_2 caused a recovery of the endogeneous HV-LV dynamics with its typical frequency shifts. Only in one case during persistent hypoxia and increasing acidosis a slowing of the fetal EEG pattern was observed. In all other cases frequency pattern during hypoxia was comparable to a normal HV- or LV-phase.

Key words: intrauterine hypoxia – fetal cerebral function – sheep fetuses

Introduction

At about 110 days of gestation the EEG of the sheep fetus begins to differentiate and an episodic pattern of high voltage slow activity (HV) and low voltage fast activity (LV) is established [Ruckebusch 1972, Dawes et al. 1972]. The differentiated and highly organized pattern of activity reflects the maturation of higher brain centers associated with the production of rapid and non-rapid eyemovement sleep. Fetal behavior, i.e., fetal muscle activity and breathing movements, is associated with the different sleep stages [Dawes et al. 1980] at about 120 days of gestation. Isocapnic hypoxia causes a reduction of fetal breathing [Clewlow et al. 1983, Gauwerky et al. 1982] and muscle activity [Gauwerky et al. 1982].

The central control mechanism of fetal behavior during hypoxia is still unknown. A recent publication has suggested that the decrease of breathing movements during hypoxia is caused by an active process arising from supracollicular structures [Dawes et al. 1983]. Up to now, no studies have been reported on the EEG-related changes of fetal behavior during hypoxia on a computerized basis. There is also minimal knowledge about the influence of different degrees of hypoxia on the activity of the cerebral cortex in the full term fetus. We have therefore

examined the sequential changes of fetal EEG activity during different states of hypoxia.

Methods

The studies were carried out on seven chronically prepared sheep fetuses (crossbred Merino) with a gestational age of 115–120 days (term 145 days). Under halothane anesthesia, a catheter was implanted into a fetal carotid artery and two pairs of fetal EC electrodes were implanted bilaterally on the parietal dura. The electrodes were implanted following the midline incision through the scalp. To create controlled hypoxia an inflatable vascular occluder was applied to the common hypogastric artery. On the 4th postoperative day the experiments were started to measure the fetal response to hypoxia. At this time all animals were in a steady state based on respiratory and cardiovascular parameters. Mean values of fetal arterial blood pH and pO_2 during normoxia were 7.37 ± 0.03 and 25.1 ± 27 mmHg, respectively. The data were recorded on FM-tape (Hewlett Packard HP 3698 A, 8 channel) and digitized. For further analysis, data were summarized on 20 second records and evaluated by spectrum analysis. The mean power, as well as the relative band energy, was calculated in the ranges of 1–25 Hz, 25–5 Hz, 5–10 Hz, 10–15 Hz and 20–30 Hz. With the help of robust regression methods [Härdle and Gassner 1984] the expected values were determined. In the figures the mean values (and the approx. 95% confidence limit) of the estimated values are shown. Fig. 1 shows data and mean values in one case as an example. The blood samples for blood gas analysis were taken at 5 minute intervals during hypoxia.

Correspondence to Dr. J. F. H. Gauwerky:

Results

Normoxia: $pO_2 > 20 \text{ mmHg}$

The EEG of the full term fetus is subject to variations with the cyclic increase and decrease of the mean power (Figure 1). The different phases do not correspond with any specific condition, but are an expression of a continually changing process. Similarly to the maximum amplitude, the duration of the cycle length is subject to an individual variability (average cycle duration: $20 \text{ min} \pm 3.6 \text{ min SE}$). Parallel to the fluctuation of the mean power, clear shifts of frequency occur. High mean power is correlated with a high proportion of low frequencies, while low mean power occurs with a high proportion of higher frequencies (Figures 2 and 3).

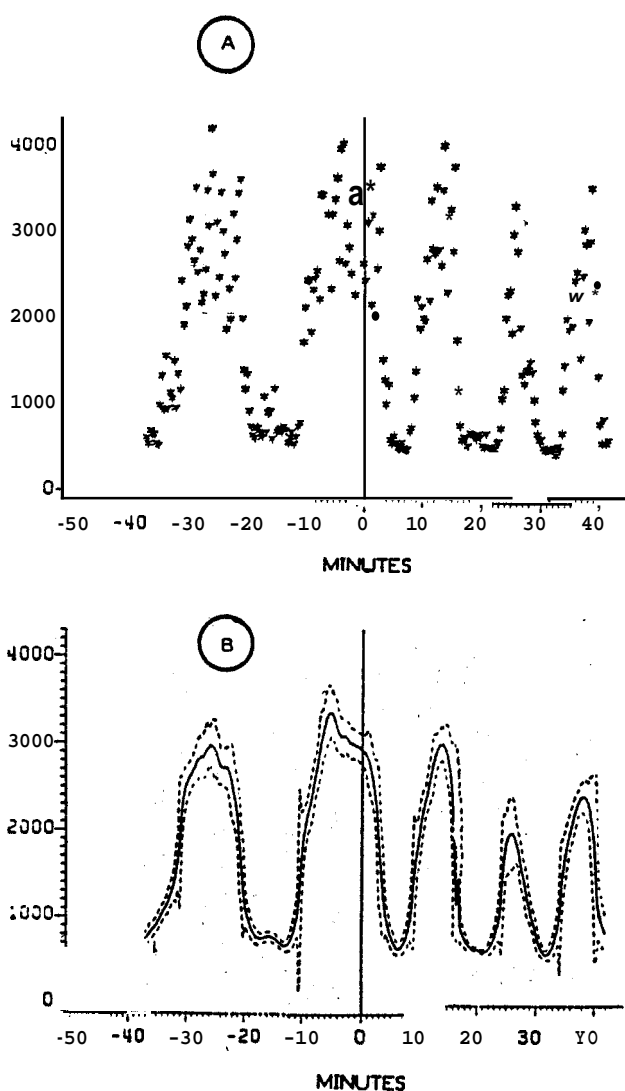


Fig. 1 Mean power in animal 3. Section A: basic data. Section B: estimated values ($\pm 95\%$ confidence limit).

Hypoxia: $pO_2 < 20 \text{ mmHg}$

Under the influence of mild hypoxia ($\Delta pO_2 < 5 \text{ mmHg}$, $pO_2 > 16 \text{ mmHg}$) we observed shortening of the "high voltage-low voltage" cycle (Figure 4). In the case shown in Figure 4, the duration of the cycle was reduced from 21 min during the normoxic phase to 12 min during hypoxia. The change of pO_2 was 3.6 mmHg in this case, the pH remained unal-

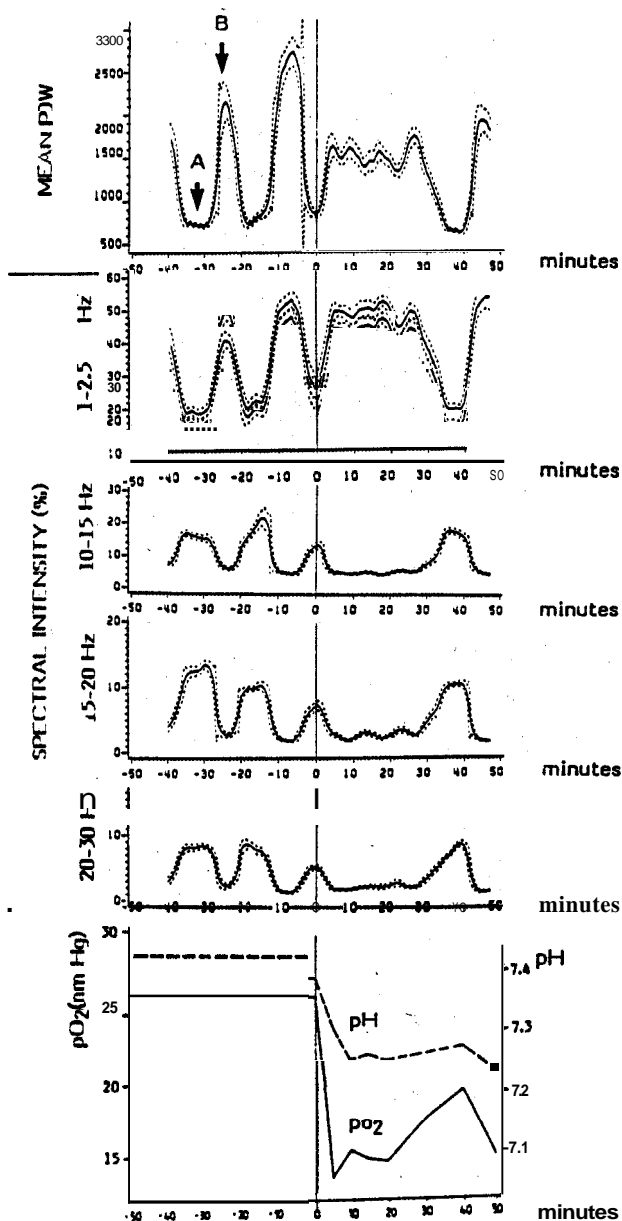


Fig. 2 Mean power and spectral intensity during normoxia and hypoxia in animal 2. Point A corresponds with a low-voltage phase (LV), point B with a high-voltage phase (HV). Distribution of frequencies see also Fig.3. During hypoxia an inhibition of the cyclic HV-LV changes occurs. An increase of the pO_2 leads to reinstatement of the cyclic changes.

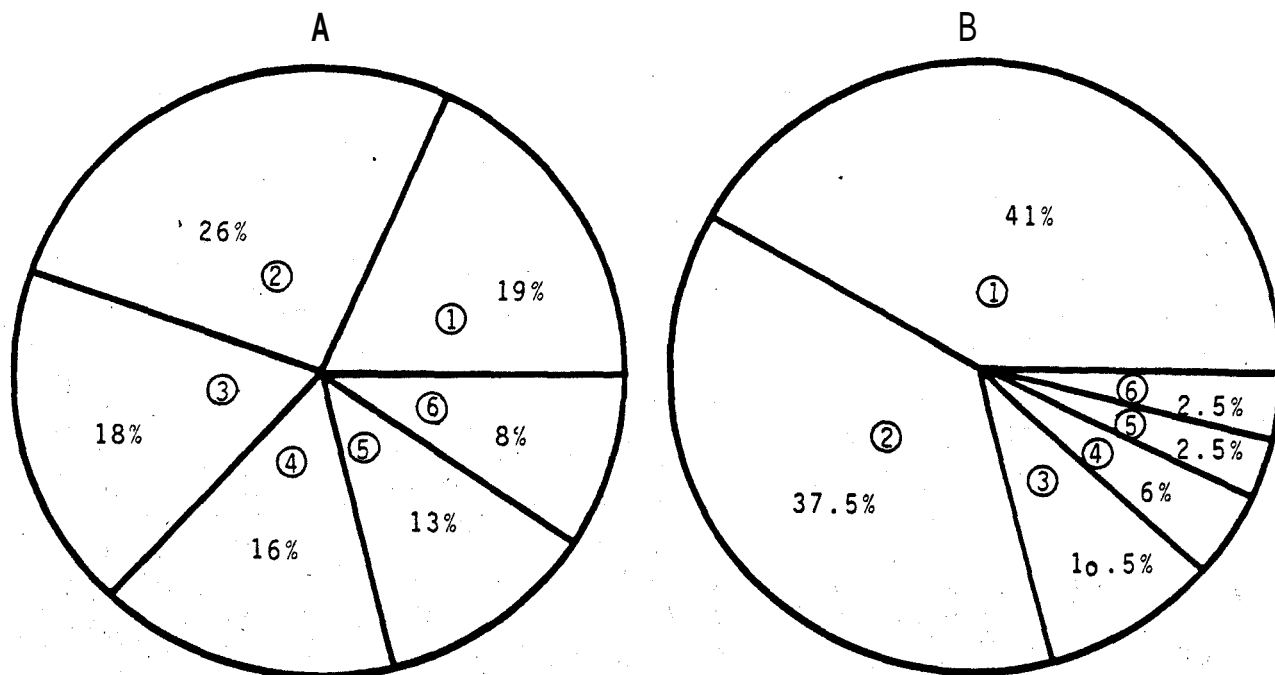


Fig.3 Distribution of frequencies during LV and HV (point A and B, Fig.2). 1: 1-2.5 Hz, 2: 2.5-5 Hz, 3: 5-10 Hz, 4: 10-15 Hz, 5: 15-20 Hz, 6: 20-30 Hz.

tered. During mild hypoxia no frequency shifts were observed. The first sign of a further increasing hypoxia is a reduction of the mean power, especially in the high voltage phase (Figure 4, point A) without any frequency shifts. Further reduction of oxygen tension ($pO_2 < 16$ mmHg) caused a loss of the cyclic changes of the EEG with high voltage and low voltage (Figures 2, 5, 6 and 7).

In cases in which hypoxia started at the end of a low voltage phase (Figures 2 and 7) a change to a high voltage phase with its typical frequency pattern always occurred. The mean power was reduced to values between high voltage and low voltage levels. An increasing pO_2 caused a prompt recovery of the endogenous high voltage - low voltage dynamics with its typical frequency shifts. The increasing pO_2 caused the onset of a low voltage phase followed by a high voltage phase (Figure 2).

A similar pattern is shown in Figure 5. Hypoxia started at the beginning of a high voltage phase. During hypoxia no significant frequency shifts occurred. The mean power decreased with persisting hypoxia and decreasing pH. A reduction of the frequencies in the range 1-2.5 Hz was observed only at the end of this registration. When hypoxia starts at the beginning of a low voltage phase, the distribution of frequencies showed the typical pattern of a low voltage phase (Figure 6). The mean power was reduced below the level of a normal low voltage phase. At the end of this registration, an increasing acidosis

caused further reduction of the mean power and a further decrease of lower frequencies (1-2.5 Hz).

The data shown in Figure 7 are in agreement with these observations. In this case, hypoxia starts again at the beginning of a high voltage phase. During persistent hypoxia and increasing acidosis the mean power decreased. We could observe a slightly slowing of the fetal EEG pattern only in this experiment. The frequency shifts seem to occur especially from the band 2 (2.5-5 Hz) into the band 1 (1-2.5).

Table 1 Acid-base values before and during hypoxia. Values are given as mean \pm SE. In the first group of four animals during hypoxia no cyclic changes of HV and LV could be observed, whereas in the other animals with only slightly decreased pO_2 the HV-LV pattern during hypoxia was at least partially present.

Ani- mal No.	HV-LV pattern during Normoxia Hypoxia	Normoxia		Hypoxia		
		pH	pO_2 (mmHg)	pH	pO_2 (mmHg)	DpO_2 (mmHg)
1	no	7.41	25.2	7.32 \pm 0.05	15.1 \pm 2.8	10.1
4	no	7.37	28.6	7.14 \pm 0.11	13.8 \pm 2.9	14.8
6	no	7.32	26.5	7.16 \pm 0.07	13.3 \pm 1.6	12.2
7	no	7.34	27.6	7.20 \pm 0.02	12.9 \pm 1.1	14.7
2	partial	7.38	25.7	7.27 \pm 0.02	16.4 \pm 2.0	9.3
5	partial	7.36	21.3	7.13 \pm 0.06	14.3 \pm 3.0	7.0
3	yes	7.42	20.9	7.41 \pm 0.02	17.3 \pm 2.3	3.6

In Table I acid base values of all animals are summarized. As shown in this table the HV-LV pattern was present during hypoxia only in cases with slightly reduced pO_2 .

Discussion

The normal EEG of the full term fetus is subject to cyclic fluctuations of the mean power, as well as of the frequency distribution. This fact was established

by Ruckebusch [1972] and Dawes et al. [1972] in animal experiments. Sokol et al. [1976] observed the same in the human fetus. In further animal experiments [Ruckebusch et al. 1977], a definite pattern of fetal behavior was associated with the EEG pattern. Fetal behavior was compared with the sleep-waking cycle of the newborn and the adult. The shifts of power and frequencies of the EEG during these behavioral states have not been described precisely so far. We could show that the EEG phases described in the literature as "high voltage" or "low voltage"

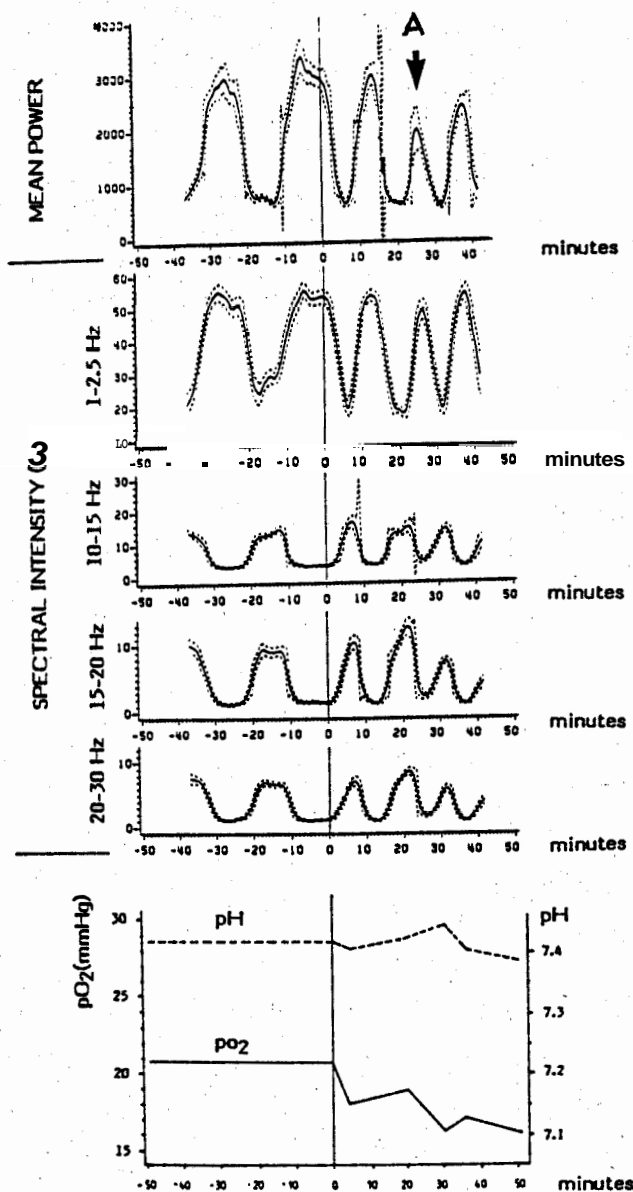


Fig. 4 Mean power and spectral intensity during normoxia and hypoxia in animal 3. Reduction of mean power during HV (point A) caused by further reduction of pO_2 .

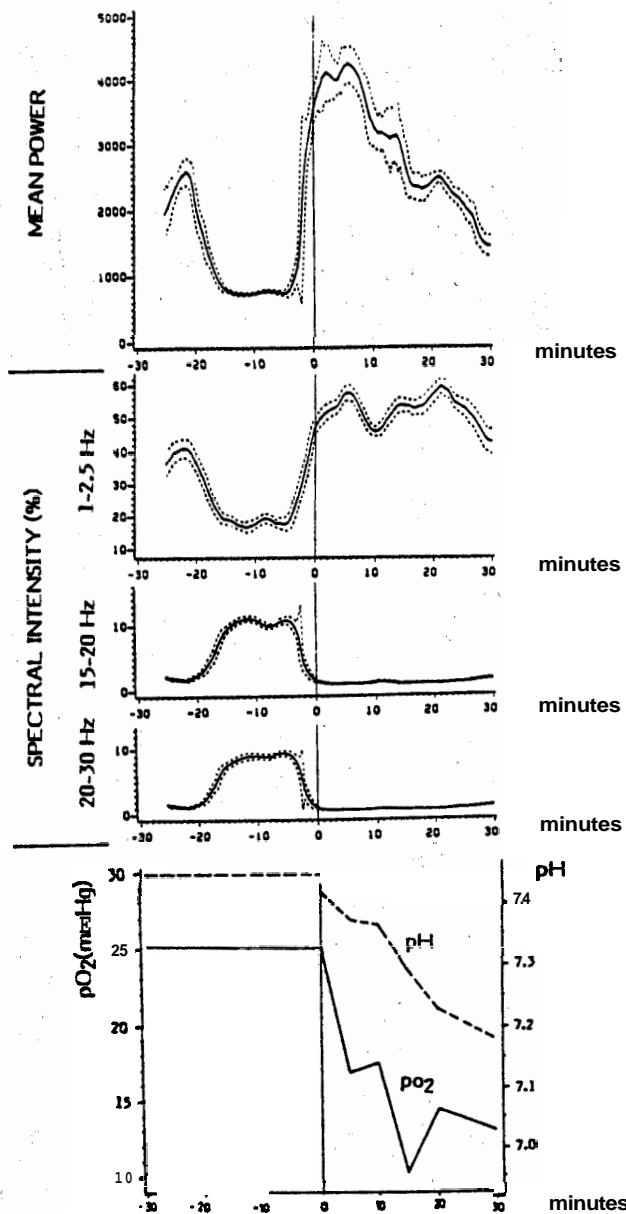


Fig. 5 Mean power and spectral intensity during normoxia and hypoxia in animal 1. Hypoxia starts at the beginning of an HV phase. Spectral distribution during hypoxia is that of an HV phase during normoxia. Mean power is continuously reduced.

[Ruckebusch et al. 1972, 1977 and Dawes et al. 1972] do not correspond to any specific condition. Moreover, a permanent and constant change of the EEG pattern occurs. It might be consistent with a cybernetical model of a feed-back mechanism [Wiener 1963].

In our experiments, the duration of one cycle was 20 min in average. The data was obtained from animals of the same gestational age. The influence of the gestational age on the duration of one sleep-waking cycle is still unknown. Investigations on the

sleep-waking cycles of premature newborns [Parmelee 1974] show that no changes of the total cycle length occur in the human fetus from the 36th-40th week of gestation. However, prior to this time it seems (measured in resting and activity cycles) that the total cycle length is clearly shorter.

A differentiation of fetal EEG between high voltage and low voltage phases occurs in the sheep in the last trimester (0.8 . term) [Ruckebusch 1972]. Investigations on the activity of the human fetus [Dreyfuß-Brisac 1975], as well as the EEG of the

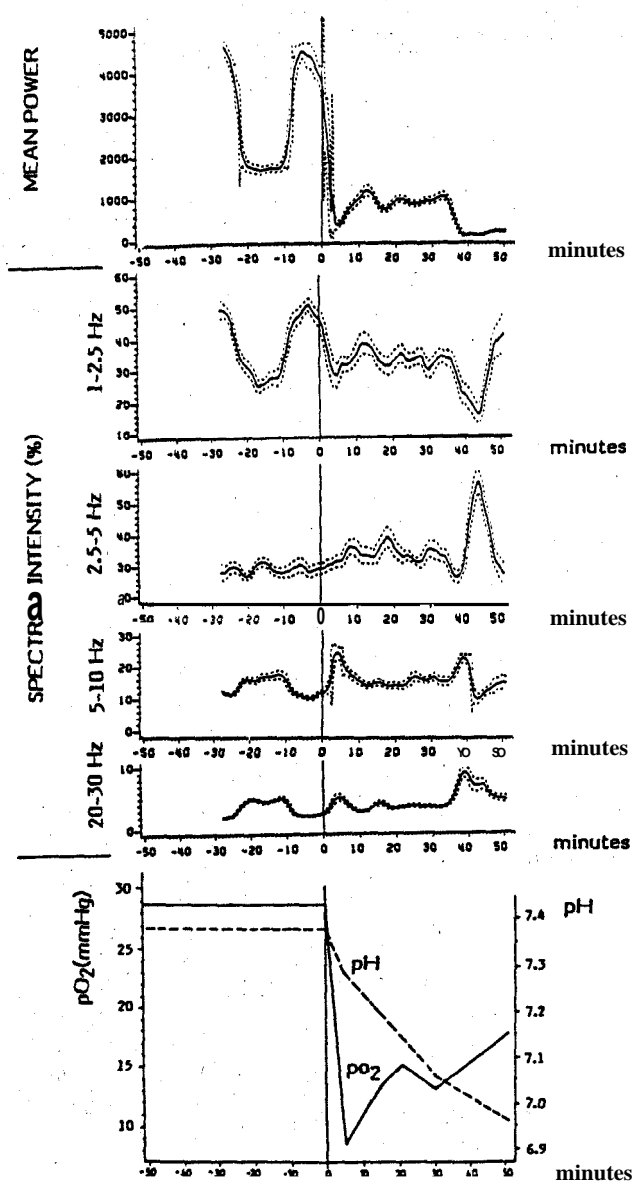


Fig.6 Mean power and spectral intensity during normoxia and hypoxia in animal 4. Hypoxia starts at the beginning of an LV-phase. Mean power is reduced below the values of LV. Distribution of frequencies corresponds to an LV-phase in the first 10 minutes of hypoxia.

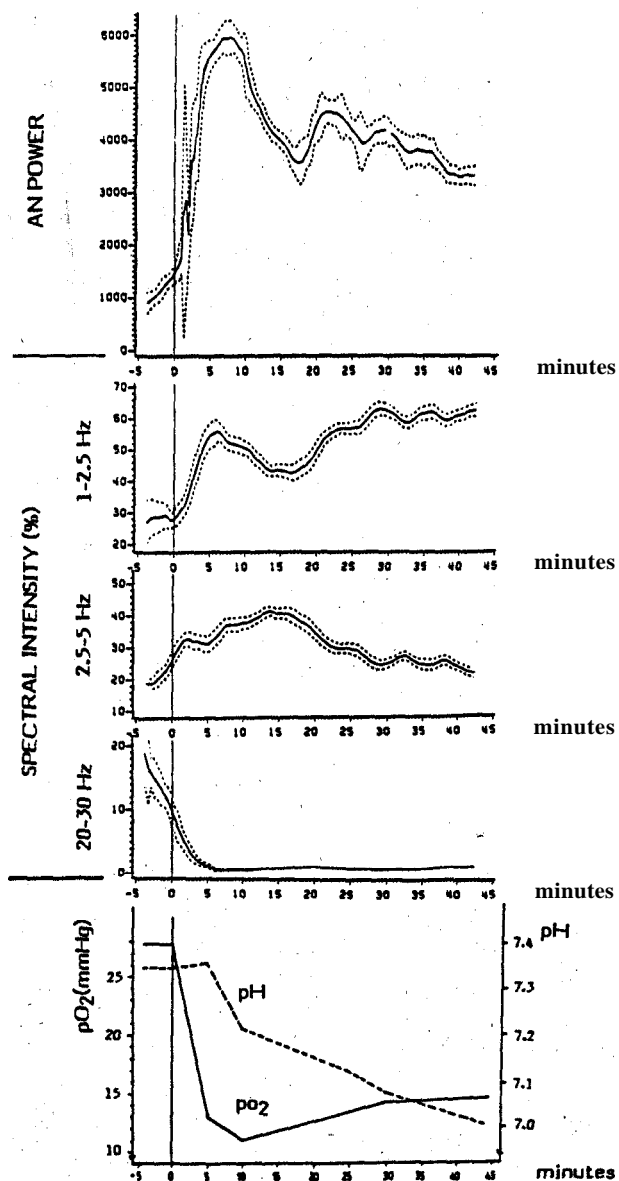


Fig.7 Mean power and spectral intensity during normoxia and hypoxia in animal 7. Hypoxia starts at the beginning of an HV-phase. With prolonged asphyxia, a slowing of frequencies and reduction of mean power occurs.

premature newborn [Dreyfuß-Brisac 1962], allows us to postulate that in human beings a corresponding differentiation takes place after the 32nd week of gestation.

Although our experiments do not allow definite conclusions on the basis of a relatively small number of investigations, we could show that changes of the EEG phase dynamics occur depending on the degree of hypoxia as well as the hypoxic gradient. First sign of mild hypoxia (ΔpO_2 5–6 mmHg, $pO_2 > 16$ mmHg) is the shortening of the high voltage – low voltage phases. This could be interpreted as a change towards a more premature EEG. Investigations carried out by Karch et al. [1977] on newborns with perinatal hypoxia confirmed these results. In contrast to these findings, Amiel-Tison [1980] demonstrated an acceleration of the cerebral maturity in children born after an intrauterine high risk incidence.

It may well be that during acute hypoxia, just as in our experiments and as reported by Karch et al. [1977], a form of reaction occurs other than in a prolonged state of intrauterine stress as reported by Amiel-Tison [1980]. Our experiments confirm the statement by Challamel et al. [1974] that existence of fetal sleep-wake cycles during labor can be interpreted as a sign of fetal well-being, even though our experiments show that changes of the phase dynamics already occur during mild hypoxia. In the neonate, sleep cycles are present only under conditions of normal oxygenation and normal acid-base balance [Radvanyi et al. 1973].

As shown in Figure 4, further increased hypoxia ($pO_2 < 16$ mmHg) leads to a decrease of the mean power, especially in the high voltage phase. During this period no frequency shifts were observed. Further increase of the hypoxia with pO_2 values under 16 mmHg causes an inhibition of the cyclic changes of high voltage and low voltage phases. When hypoxia starts at the beginning of a high voltage phase, this phase remains unchanged regardless of the mean power. The mean power is depressed, but the frequency distribution is the same as in a high voltage phase during normoxia. With the beginning of hypoxia at the transition to a low voltage phase, similar results were obtained. The frequency distribution is in accordance with that of a low voltage phase during normoxia but the mean power is reduced as compared to levels during normoxia.

Only in one case (Figure 7) did we see a minor reduction of frequency during prolonged acidosis (reduction of the spectral intensity in the band 2.5–5 Hz, increase of spectral intensity in the band 1–2.5 Hz). These findings are not in agreement with the acute experiments of Rosen et al. [1973, 1967] and Symmes et al. [1970], who found that the slowing of frequency is one of the first signs during hypoxia. On

Table 2 Fetal electroencephalography and FHK monitoring compared [Viniker 1979].

Feature	Continuous FHR monitoring	Fetal electroencephalography
Preceding experience	Fetal heart auscultation	New technique
Signal	"Regular" Easily obtainable Easily checked by auscult.	Random Technically difficult to obtain No simple check
Trace	Convenient Parameters of fetal distress well defined Simple to read	Voluminous Parameters of fetal distress not established Expertise required
Esperience	Large	Limited to a few centres
Cerebral function	Not directly related	Directly related
Current status	Clinically accepted	Research technique only

the basis of our experiments, the following model for the regulation of the fetal EEG is possible: higher brain centers control the activity of the cerebral cortex. The result is a cyclic variation of high voltage and low voltage. Under the influence of hypoxia the activity of these structures is depressed causing first a shortening of the high voltage – low voltage cycles and after that a total inhibition of the cyclic changes. Parallel to that, hypoxia causes a diminution of the power of the cortical neurons. Possibly, the frequency distribution may be influenced by acidosis.

Corresponding with these results, investigations on the regulation of fetal behavior during hypoxia should be done with recourse to the fetal EEG.

The extent to which the fetal EEG is useful for the diagnosis of antepartum or intrapartum stress situation is still unknown. In Table 2 the fetal EEG and the registration of the fetal heart rate are compared. In contrast to the fetal EEG, the signal collection and the date interpretation of the FHR seems to be relatively easy. The fetal EEG, however, is directly related to the cerebral function. Symmes et al. [1970] concluded: "The times at which statistically significant changes occurred were not earlier during the progressive fall of arterial pO_2 than visual estimates of abnormality made on line from the paper record and were in all cases later than significant cardiovascular changes." He presented the fetal EEG as a purely competitive method to the fetal CTG with regard to the early diagnosis of an intrauterine high risk situation. Our study shows that the fetal EEG is subject to physiological long-term dynamics, and thus changes due to hypoxia cannot always reliably be distinguished from physiological variants observing only

the momentary state. On the other hand, the registration of the fetal EEG is the only method for understanding the fetal brain function available at present. Considering the increased interest in newborn morbidity, it could be of great importance for the recognition and prevention of brain damage. Sureau [1977] summarized that the fetal EEG might improve our understanding of physiological changes in the fetus, but technical difficulties have impaired its practicability in the clinical day-to-day management. However, the rapid development of microprocessor techniques in recent years could lead to clinically useful methods for the fetal EEG registration and interpretation.

REFERENCES

- Amiel-Tison C* 1980 Possible acceleration of neurological maturation following high risk pregnancy. *Am. J. Obstet. Gynecol.* 138: 303
- Challamel MJ, Revol M, Bremond A, Fargier P* 1974 EEG foetal au cours du travail. *Rev. Fr. Gynec. Obstet.* 70: 235
- Clewlow F, Dawes GS, Johnston BM, Walker MW* 1983 Changes in breathing, electrocortical and muscle activity in unanaesthetized fetal lambs with age. *J. Physiol.* 341: 463
- Dawes GS, Fox HE, Leduc BM, Liggins GC, Richards RT* 1972 Respiratory movements and rapid eye movement sleep in the foetal lamb. *J. Physiol.* 220: 119
- Dawes GS, Gardner WN, Johnston BM, Walker DW* 1980 Activity of intercostal muscles in relation to breathing movements, electrocortical activity and gestational age in fetal lambs. *J. Physiol.* 307: 47
- Dawes GS, Gardner WN, Johnston BM, Walker DW* 1983 Breathing in fetal lambs: the effects of brain stem section. *J. Physiol.* 331: 535
- Dreyfuß-Brisac C* 1962 The electroencephalogram of the premature infant. *World Neurology* 3: 5
- Dreyfuß-Brisac C* 1975 Neurophysiological studies in human premature and full-term newborns. *Biol. Psychiatry* 10: 485
- Gauwerky J, Wernicke K, Boos R, Kubli F* 1982 Heart rate variability, breathing and body movements in hypoxic fetal lambs. *J. Perinat. Med.* 10 (Suppl 2): 113
- Härdle W, Gassner P* 1984 Robust non-parametric function fitting. *J. R. Statist. Soc. B.* 46: 42
- Karch D, Kastl E, Sproch I, Bernuth H* 1977 Perinatal hypoxia and bioelectric brain maturation of the newborn infant. *Neuropädiatrie* 8: 253
- Parmelee Jr AH* 1974 Ontogeny of sleep patterns and associated periodicities in infants. Pre- and postnatal development of the human brain. *Mod. Probl. Paediat.* 13: 298
- Radvanyi MF, Monod N, Dreyfuß-Brisac C* 1973 Electroencéphalogramme et sommeil chez le nouveau-né en détresse respiratoire. Etude de l'influence des variations de la PaO₂ et de l'équilibre acidobasique. *Bull. Physiopathol. Resp.* 91: 1569
- Rosen MG* 1967 Effects of asphyxia on the fetal brain. *Obstet. Gynecol.* 29: 687
- Rosen MG, Scibetta J, Chik L, Borgstedt AD* 1973 An approach to the study of brain damage. *Am. J. Obstet. Gynecol.* 115: 37
- Ruckebusch Y* 1972 Development of sleep and wakefulness in the foetal lamb. *Electroenceph. Clin. Neurophysiol.* 32: 119
- Ruckebusch Y, Gaujoux M, Eghbali B* 1977 Sleep cycles and kinesis in the foetal lamb. *Electroenceph. Clin. Neurophysiol.* 42: 226
- Sokol RJ, Rosen MG, Chik L* 1976 Fetal electroencephalography. In *Beard RW, Nathanielz PW (eds)*. Fetal Physiology and Medicine. W. B. Saunders, Co., London, p 476
- Sureau C* 1977: In: *Philipp EE, Barnes J, Neuton M (eds)* Scientific Foundations of Obstetrics and Gynecology. 2nd edn, Heinemann, London, p 882
- Symmes D, Prichard JW, Mann LI* 1970 Spectral analysis of fetal sheep EEG during hypoxia. *Electroenceph. Clin. Neurophysiol.* 29: 511
- Viniker A* 1979 The fetal EEG (detections of oxygen deprivation). *Br. J. Hosp. Med.* 6: 504
- Wiener N* 1963 *Kybernetik*. Econ-Verlag GmbH, Düsseldorf, p 145