Biometrika (1992), **79**, 4, *pp.* 783-95
Printed in Great Britain

Choice of bandwidth for kernel regression when residuals are correlated

BY EVA HERRMANN

Institut für Angewandte Mathematik, Universität Heidelberg, Im Neuenheimer Feld 294, D-6900 Heidelberg, Germany

THEO GASSER AND ALOIS KNEIP

Zentralinstitut für Seelische Gesundheit, J5, P.O.B. 5970, D-6800 Mannheim 1, Germany

SUMMARY

Standard techniques for selecting the bandwidth of a kernel estimator from the data in a nonparametric regression model perform badly when the errors are correlated. In this paper we propose a modified version of an existing plug-in bandwidth selector. The method is generalized to stationary error variables by estimating a functional of the residual covariance function. The proposed bandwidth selector shows good properties in asymptotic theory and in simulations without assuming a parametric model for the error process.

Some key words: Bandwidth selection; Correlated residuals; Kernel estimator; Nonparametric regression.

1. INTRODUCTION

Assume that data $Y_1, \ldots, Y_n$ follow a regression model with a smooth function g and a fixed design $0 \leqslant t_1 < \ldots < t_n \leqslant 1$:

$$Y_i = Y(t_i) = g(t_i) + \varepsilon_i \quad (i = 1, \ldots, n) \tag{1·1}$$

where the ε_i are zero-mean residuals. When a priori knowledge is vague, it is advisable to avoid the arbitrary specification of a parametric model for g and to apply a nonparametric estimator of g instead. We will use the following kernel estimator (Gasser & Müller, 1979; Müller, 1988, p. 16; Härdle, 1990, pp. 28-32)

$$\hat{g}(t, b) = \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} \frac{1}{b} W\left(\frac{t-u}{b}\right) du \, Y_i, \tag{1·2}$$

where $s_0 = 0$, $s_i = \frac{1}{2}t_i + \frac{1}{2}t_{i+1}$ for $i = 1, \ldots, n-1$ and $s_n = 1$ with a kernel function W and bandwidth b.

Most of the literature on nonparametric regression deals with independent residuals ε_i; see, however, Altman (1990) and Hart (1991). The choice of the bandwidth b is crucial for the performance of the estimates. A data-adaptive optimization of b makes the method also more objective and more economical. Two different approaches have been proposed to estimate the optimal bandwidth, namely (i) minimizing the sum of squares of some type of estimated residuals as in cross-validation, generalized cross-validation, and minimum unbiased risk estimation, and (ii) estimating the asymptotically optimal bandwidth leading to the so called plug-in estimator.

Most of the literature deals with variants of the first approach; see, e.g., Craven & Wahba (1979) and Rice (1984). Härdle, Hall & Marron (1988) showed that these variants are asymptotically equivalent with respect to their integrated squared error. A plug-in estimator was proposed by Gasser, Kneip & Köhler (1991) for independent residuals. It has superior rates of convergence in mean squared error compared to a cross-validation type estimator and also behaves much better in simulations. It is based on the asymptotically optimal bandwidth and involves estimating the residual variance nonparametrically (Gasser, Sroka & Jennen-Steinmetz, 1986) and estimating a functional of the second derivative in an iterative scheme.

Unfortunately all these data-adaptive methods developed for independent residuals break down in the case of correlated residuals. Intuitively, this is plausible, since highly correlated residuals may look smooth and can be mistaken for signal. Figure 1 illustrates this effect. It shows a simulated example with 75 data points where the residuals follow an AR(10) process; see § 4 for further details. The method introduced in § 2 chooses a bandwidth which is close to the optimal one with respect to mean integrated squared error. The naive plug-in method which assumes independent residuals underestimates the appropriate bandwidth and the resulting curve estimator almost interpolates the data. Simulations show that naive cross-validation procedures have the same drawback. This lack of robustness to dependence is disturbing because correlated residuals are quite frequent in regression problems; in our biomedical work we have encountered correlations when analyzing intragastric pH (Bauerfeind et al., 1987) and when analyzing brain potentials (Gasser et al., 1988), and correlations are expected to be the rule for physiological curves.

One solution is to modify cross-validation by leaving out k points with $k \geq 1$ or by down-weighting neighbours (Györfi et al., 1989, p. 131). Unfortunately there is no effective method to choose the modification so far. Further, variability is a problem even in the classical case with $k = 1$ (Gasser et al., 1991). Alternatively an unbiased risk approach

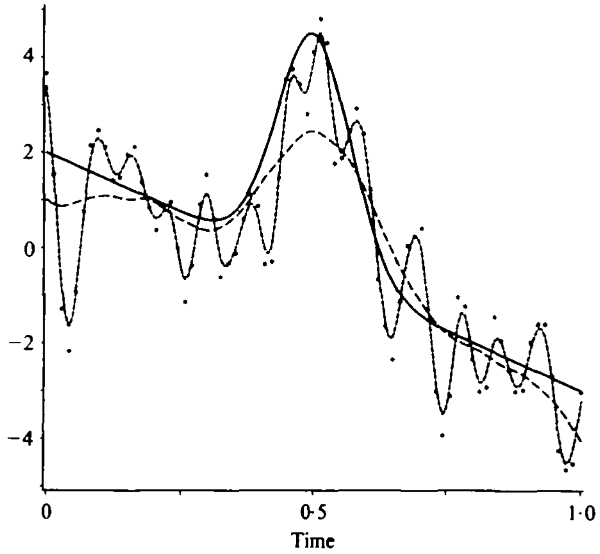

Fig. 1. Simulated example with true regression curve, shown by solid line, data, naive plug-in estimator, dotted line, and a modified version of the plug-in estimator, dashed line.

can be used. A parametric model is fitted to estimated residuals while excluding the low frequency part (Hart, 1991; Chiu, 1989; Altman, 1990). This approach makes unduly strong assumptions in a nonparametric context. Our aim was to develop an estimator for the optimal bandwidth in the case of correlated residuals, satisfying the following requirements: it should be robust towards a broad variety of correlation structures, no parametric model for the time series ε_i should need to be specified, and the loss in efficiency should be small for independent residuals.

Because of its good performance in the independent case, we started with the plug-in estimator of Gasser et al. (1991). The correlation structure is taken into account by a second parameter which is a functional of the residual process. Thereby, again, we introduce additional parameters for which appropriate fixed values are proposed.

Section 2 contains a description of the method. Asymptotic theory is given in § 3 and simulation results are summarized in § 4. Theoretical details are deferred to an appendix. Equally spaced data are assumed so as to simplify the analysis.

2. Definition of the method

Assume that the residuals in model (1·1) follow a strictly stationary time series with $E(\varepsilon_i)=0$, $\mathrm{var}(\varepsilon_i)=\sigma^2$, $E(\varepsilon_i\varepsilon_{i+\nu})=\gamma_\nu$ and that the regression function g is smooth. Our mean integrated squared error criterion for a kernel estimator $\hat{g}$ with bandwidth b is given by

$$\textsc{mise}\,(b)=E\left[\int_0^1 v(t)\{g(t)-\hat{g}(t,b)\}^2\,dt\right],$$

where v is a twice continuously differentiable function with support $[\delta, 1-\delta]$ and $v(t)>0$ for all $t\in(\delta, 1-\delta)$ and some $\delta>0$. The function v is used at the bandwidth selection stage to avoid boundary problems; when estimating g itself appropriate kernels are used for estimating near the boundary (Gasser, Müller & Mammitzsch, 1985). Assume also that g'' does not vanish identically on $[\delta, 1-\delta]$. The optimal bandwidth b_M minimizes the mean integrated squared error. In a recent paper Hall & Hart (1990) obtained optimal rates for mean integrated squared error. Our goal is to estimate b_M in order to optimize the nonparametric regression fit. For this we will use the bandwidth b_A minimizing the asymptotic mean integrated squared error. For a kernel of order 2 it is given by

$$b_A=\left(C_1S\Big/\left[nC_2^2\int_0^1 v(t)\{g''(t)\}^2\,dt\right]\right)^{1/5}, \qquad (2\cdot1)$$

with

$$C_1=\int_0^1 v(t)\,dt\int_{-1}^1\{W(x)\}^2\,dx,\quad C_2=\int_{-1}^1 x^2W(x)\,dx,\quad S=\sigma^2+2\sum_{\nu=1}^{\infty}\gamma_\nu.$$

Gasser et al. (1991) introduced a method to estimate the optimal bandwidth b_M by estimating the unknown terms in b_A. However, this relies essentially on independence and breaks down in case of correlated residuals. Here, a generalization to stationary time series errors is proposed. Equation (2·1), already given by Hart (1991), shows well the impact of assuming a correlated or an uncorrelated model. When wrongly assuming uncorrelated errors, the bandwidth becomes too small when $S>\sigma^2$. This is frequent in practice, where low frequency noise dominates high frequency noise. In the simulated example of Fig. 1 we have $S\simeq 6\cdot1\sigma^2$. The other way round it would become too large which is a rarer case.

In the case of uncorrelated residuals S is simply their variance σ^2. Plug-in estimators for b_M use this formula with the known constants C_1 and C_2 while plugging in estimators for σ^2 and for $\int v(t)\{g''(t)\}^2\,dt$. In order to obtain a good estimator for the latter functional, an iterative procedure was suggested by Gasser et al. (1991). When replacing the estimator for the variance by an estimator $\hat{S}$ for S, this plug-in estimator can be generalized, leading to the following algorithm:

(i) set the starting value $b_0 = 1/n$,

(ii) set

$$b_i = \left[\frac{C_1\hat{S}}{nC_2^2\int_0^1 v(t)\{\hat{g}_2(t, b_{i-1}n^{1/10})\}^2\,dt}\right]^{1/5} \quad (i = 1, \ldots, i^*), \tag{2·2}$$

(iii) set $\hat{b} = b_{i^*}$.

According to theoretical considerations the number of iterations i^* is fixed at 11. Generally, further iterations do not improve the estimator and less iterations increase the variance in theory and simulations.

The estimator $\hat{g}_2$ of g'' is given by a kernel estimator with bandwidth b, defined by

$$\hat{g}_2(t, b) = \sum_{i=1}^{n}\int_{s_{i-1}}^{s_i}\frac{1}{b^3}W_2\left(\frac{t-u}{b}\right)du\,Y_i$$

as in the uncorrelated case (Müller, 1988, p. 27). In the iterations the bandwidth b_{i-1} is inflated in an asymptotically appropriate proportion with respect to the bandwidth b_M to be estimated.

While in the independent case the estimation of the functional of g'' is the crucial point, the estimation of S becomes now an important tuning parameter. Obviously, the functional S summarizes information about the covariance function γ_ν in an asymptotically appropriate way. For independent residuals Gasser et al. (1986) proposed a simple method for estimating σ^2. For equally spaced design, it depends on pseudo-residuals $\hat{\varepsilon}_i = Y_i - \frac{1}{2}Y_{i+1} - \frac{1}{2}Y_{i-1}$. However, it does not carry over to the independent case in a straightforward way. Intuitively, the problem when estimating S can be described as follows: S is equivalent to the value of the true residual spectrum at zero frequency. However, at zero frequency the spectrum of residuals and the spectrum of the regression functions usually both have large power. They are thus difficult to disentangle without relying on restrictive assumptions. Even the assumption of parametric models for the residual process may necessitate further precautions, such as the exclusion of a frequency interval $[0, \nu_0)$ when estimating the parameters (Chiu, 1989). For m-dependent residuals Müller & Stadtmüller (1988) suggested an estimator for S based on first order differences of residuals with lag ν $(\nu \geq 1)$, that is $\hat{\varepsilon}_{i,\nu} = Y_i - Y_{i-\nu}$. Using second order differences leads us to consider lagged residuals defined as

$$\hat{\varepsilon}_{i,\nu,\mu} = Y_i - \frac{\nu}{\nu+\mu}\,Y_{i+\mu} - \frac{\mu}{\nu+\mu}\,Y_{i-\nu}.$$

Our estimator $\hat{S}_m$ for S is defined as follows:

$$\hat{S}_m = \sum_{\nu=-m}^{m}\hat{\gamma}_\nu, \tag{2·3}$$

with

$$\hat{\gamma}_0=\frac{1}{n-2m-2}\sum_{i=m+2}^{n-m-1}\tfrac{2}{3}\hat{\varepsilon}^2_{i,m+1,m+1},\quad \hat{\gamma}_\nu=\frac{1}{n-m-1-\nu}\sum_{i=\nu+1}^{n-m-1}\alpha_\nu\hat{\varepsilon}^2_{i,\nu,m+1}+\beta_\nu\hat{\gamma}_0,$$

$$\hat{\gamma}_{-\nu}=\frac{1}{n-m-1-\nu}\sum_{i=m+2}^{n-\nu}\alpha_\nu\hat{\varepsilon}^2_{i,m+1,\nu}+\beta_\nu\hat{\gamma}_0,$$

$$\alpha_\nu=\frac{m+1+\nu}{2m+2},\quad \beta_\nu=-\alpha_\nu\left\{\left(\frac{\nu}{m+1+\nu}\right)^2+\left(\frac{m+1}{m+1+\nu}\right)^2+1\right\},$$

for $\nu=1,\ldots,m$.

The estimator $\hat{S}_m$ is unbiased in the case of m-dependent residuals ε_i and linear regression functions g; $\hat{S}_0$ reduces to the residual variance estimator proposed by Gasser et al. (1986). Asymptotic results show that $\hat{S}_m$ is consistent when $m\to\infty$ and $m^2/n\to 0$ as $n\to\infty$ for residuals that are not necessarily m-dependent but satisfy some mixing conditions, and if the regression function is smooth but not necessarily linear. The mixing conditions are satisfied, e.g. by an autoregressive moving average, ARMA, process. Note that S is an auxiliary parameter so that a simple estimator may be appropriate. Simulations with a suitable choice of m revealed good properties for this estimator.

The parameter m might thus be used as a tuning parameter when estimating the optimal bandwidth in order to account for the correlation present in the residuals. Subsequently, we sketch two approaches for a data-dependent choice of the tuning parameter m, which plays a crucial role for the suitability of the estimator $\hat{S}_m$. Let $\hat{b}_m$ denote the estimator for the optimal bandwidth given by algorithm (2·2) with $\hat{S}=\hat{S}_m$.

Method (i). A simple method which works quite well, especially when the covariances γ_ν are nonnegative and rapidly decreasing with ν, is to choose m as the largest integer such that $\hat{b}_m\geqslant\alpha_1\hat{b}_{m-1}$ and $m\leqslant\alpha_2 n^{\frac{1}{4}}$, with constants $\alpha_1>1$ and $\alpha_2>0$. These rules are based on the following heuristic reasoning: assume the important case of a correlation structure which is dominated by positive autocorrelations, i.e. at least $S>\sigma^2$ has to hold. Then the sequence $\hat{b}_m$ tends to have a substantial increase as long as m is too small. For an appropriate m an asymptote is reached sometimes with some superposed wiggles. It is thus reasonable to stop when the increase becomes small. On the other hand m should increase slowly with n to guarantee a reasonable variance; see Remark 1 in § 4 which motivates this rule. We tested different values for α_1 and α_2 in simulations. The choice $\alpha_1=\frac{6}{5}$ and $\alpha_2=\frac{1}{5}$ leads to good results and has been used in the final simulations.

Method (ii). After smoothing one can use the differences $Y_i-\hat{g}(t_i)$ as estimated residuals and compute the covariance function estimator which is based on these residuals. More explicitly, derive for all $m\leqslant M$ the estimators

$$\tilde{S}_m=\frac{C_3}{n}\sum_{\nu=-k}^{k}\sum_{i=\max(1,\nu+1)}^{\min(n,n-\nu)}\{Y_i-\hat{g}(t_i,\hat{b}_m)\}\{Y_{i+\nu}-\hat{g}(t_{i+\nu},\hat{b}_m)\}v(\tfrac{1}{2}t_i+\tfrac{1}{2}t_{i+\nu}),\qquad(2\cdot4)$$

with $C_3=1/\int v(t)\,dt$. All these $\tilde{S}_m$ are consistent estimators of S, as shown by Remark 2 in § 3, but in our simulations they differ considerably. Like the bandwidths in (i), they first increase with increasing m, then stabilize for moderate values of m and for large values of m they become very variable. Therefore we select the median value, say $\tilde{S}$, from all values $\tilde{S}_m$ $(m=1,\ldots,M)$. Then we choose the m with minimal distance $(\hat{S}_m-\tilde{S})^2$. Simulations have shown that the additional parameters k and M do not affect this selection rule for m very much. In our final simulations we used for k and M the nearest integers to $n^{1/4}$ and $n^{\frac{1}{2}}$ respectively which have given good results. This method is preferable to (i) if there is a more complex dependence structure for the residuals, but differences between the two approaches turn out to be small; see § 3.

3. ASYMPTOTIC RESULTS

We will show that the good asymptotic properties of the plug-in estimator $\hat{b}_{\text{opt}}$ derived by Gasser et al. (1991) for independent residuals remain valid in the case of correlated residuals, with some modifications. The regression model (1·1) is considered with the following assumptions.

Assumption 1. The design is equally spaced, $t_i = i/n - 1/(2n)$ for $i = 1, \ldots, n$.

Assumption 2. The residuals follow a strictly stationary time series satisfying $E(\varepsilon_i) = 0$, $E(|\varepsilon_i|^k) < \infty$ for all $k \in \mathbb{N}$. The cumulants are summable, that is $\Sigma\,|\text{cum}\,\{\varepsilon_1, \varepsilon_{i_1}, \ldots, \varepsilon_{i_{k-1}}\}| < \infty$ for all $k \in \mathbb{N}$ and additionally $S > 0$ and $\Sigma\, \nu|\gamma_\nu| < \infty$. The spectral density is bounded away from zero.

Assumption 3. The regression function g is l_p times differentiable and

$$|g^{(l_p)}(x) - g^{(l_p)}(y)| \leq L|x - y|^{p - l_p}$$

for all x, $y \in [0, 1]$ and some $p > 2$, $l_p \in \mathbb{N}$ with $l_p < p \leq l_p + 1$ and $L < \infty$. For at least one $l \in \{0, \ldots, l_p - 1\}$ the derivative $g^{(l+1)}$ does not vanish in $[\delta, 1-\delta]$ and $g^{(l)}$ achieves an absolute maximum or absolute minimum in $[\delta, 1-\delta]$.

Assumption 4. The kernel W is a Lipschitz continuous symmetric probability density with support $[-1, 1]$. The kernel function W_2 used for estimating g'' is assumed to be Lipschitz continuous, symmetric at zero with support $[-1, 1]$ and satisfies $\int W_2(x)\,dx = 0$ and $\int W_2(x)x^2\,dx = 2$.

The first assumption simplifies the arguments. Assumption 2 gives the mixing condition. A broad class of stationary time series satisfies these assumptions, especially the popular autoregressive and moving average processes. The first part of Assumption 3 gives smoothness conditions for the regression function. Gasser et al. (1991) also give convergence rates when these smoothness conditions are not satisfied. Their Theorem 3 is supposed to hold also for the modified plug-in estimator with straightforward modifications. The second part of Assumption 3 ensures only that the asymptotic formula (2·1) holds and excludes, in particular, linear functions g. Assumption 4 is a common one and is satisfied for optimal kernels (Gasser et al., 1985).

In the following let $q = \min\,(p-2, 2)$. We first prove consistency of the estimator $\hat{S}_m$ defined by (2·3).

PROPOSITION 1. *Suppose that Assumptions* 2 *and* 3 *are satisfied.*

(*a*) *If the residuals are m_0-dependent, then $\hat{S}_m$ is a $n^{\frac{1}{2}}$-consistent estimator for S for every fixed $m \geq m_0$.*

(*b*) *Let $m \to \infty$, $m^2 n^{-1} \to 0$ for $n \to \infty$. Then $\hat{S}_m$ is a consistent estimator for S with*

$$\text{var}\,(\hat{S}_m) = O(m^2 n^{-1}),$$

$$E(\hat{S}_m) = S - \sum_{\nu=m+1}^{\infty} a_\nu \gamma_\nu + O\{(m+1)^3 n^{-2}\} = S + o(1),$$

where $3+2m \leqslant a_{m+1} \leqslant 4+3m$, $1 < a_{\nu+m+1} < 2$ *for* $\nu = 1, \ldots, m$, $|a_{2m+2}| \leqslant 2+m$ *and* $a_{\nu+2m+2} = 2$ *for all* $\nu \geqslant 1$.

The proof is given in the Appendix. The bounds for the constants a_ν show that they are positive if $\nu \neq 2m+2$ and that a_{m+1} dominates each of the others. Rates of convergence can be obtained by requiring a further assumption on the decrease of the covariances. For example if $\Sigma\, \nu^{\alpha}|\gamma_\nu| < \infty$ holds for some $\alpha \geqslant 1$, then $\hat{S}_m = S + O_p(mn^{-\frac{1}{2}}) + o(m^{-\alpha+1})$.

Now we establish the rate of convergence for b_A to the mean integrated squared error optimal bandwidth b_M.

PROPOSITION 2. *Suppose that Assumptions* 1-4 *hold. Then* b_A *defined by* (2·1) *satisfies*

$$|b_A - b_M| = O(n^{-(q+2)/10}).$$

The proof is straightforward, following Gasser et al. (1991) with the modifications sketched in the Appendix.

In the following theorem we assume $\hat{S}$ to be a consistent estimator for S. Besides $\hat{S}_m$, other consistent estimators could be used, for example, if a parametric model for the error series is available.

THEOREM. *If Assumptions* 1-4 *are satisfied and if* $\hat{S}$ *is a consistent estimator for* S, *then the estimator* $\hat{b}_{\text{opt}}$ *for the optimal bandwidth given by* (2·2) *satisfies*

$$\hat{b}_{\text{opt}} = b_M[1 + O\{n^{-q/10} + |E(\hat{S}) - S|\} + O_p\{n^{-(q+3)/10} + |\hat{S} - E(\hat{S})|\}].$$

Note that our basic assumptions lead to $q \leqslant 2$. This theorem is proved in the Appendix. The first term is related to the bias of the estimator and the second term to the variability. The general form can be specified if rates for the bias and variance of $\hat{S}$ can be obtained following Proposition 1.

The following remarks help in understanding when the two criteria for choosing m will work. Their proofs are similar to that of the Theorem.

Remark 1. Suppose Assumptions 1-4 hold. Then the estimator $\hat{b}_m$ satisfies

$$\hat{b}_m = b_M\left[\left\{\frac{E(\hat{S}_m)}{S}\right\}^{1/5} + O(n^{-q/10}) + O_p(n^{-(3+q)/10}) + O_p(mn^{-\frac{1}{2}})\right]. \qquad (3\cdot1)$$

Note that the variability of $\hat{b}_m$ depends strongly on the variability of $\hat{S}_m$ which is related to the last term in (3·1) and therefore increases with m. Suppose that the residuals have nonnegative covariances γ_ν which are rapidly decreasing with lag ν. Then the mean $E(\hat{S}_m)$ will increase against S when $m \rightarrow \infty$; compare Proposition 1. The first criterion will choose m where the increase of $E(\hat{S}_m)$ is mainly finished and where the variance of $\hat{S}_m$ is still low.

Remark 2. Suppose Assumptions 1-4 hold. If $\hat{b}_m > n^{-1+\eta}$ for some $\eta > 0$, then the estimators $\tilde{S}_m$ in (2·4) can be written as

$$\tilde{S}_m = S - \frac{2S}{n\hat{b}_m}(2k+1)\,W(0) + \text{MISE}\,(\hat{b}_m)(2k+1)\{1 + o_p(1)\} + O_p\left(\frac{1}{k} + kn^{-\frac{1}{2}}\right).$$

Remark 1 shows that for every $m \leqslant M$ with $M = o(n^{\frac{1}{2}})$ the estimated bandwidth $\hat{b}_m$ has a rate of decrease of about $n^{-1/5}$ and then Remark 2 establishes that for such a bandwidth at least $\tilde{S}_m = S + O_p(n^{-1/4})$ holds. This gives some justification for the second criterion. Simulations show that $\tilde{S}_m$ varies considerably with m. But choosing the median yields a robust estimator $\tilde{S}$ leading to a suitable choice of m for the simulations done.

4. SIMULATIONS AND APPLICATIONS

The finite sample properties of the estimator $\hat{b}$ and the resulting $\hat{g}(t, \hat{b})$ have been studied by simulation. After describing the simulation design we shall illustrate the results for some typical and interesting situations. Noise processes, regression functions and sample sizes varied across simulations. As noise processes we took Gaussian ARMA processes normed to variance one:

(i) white noise,
(ii) AR(1) processes with coefficients from −0·8 to 0·98,
(iii) an AR(3) process from the literature (Hurwich, 1985) which has a spectrum with a broad peak and further power in the higher frequency part and almost no power in the low frequency part,
(iv) autoregressive processes of order between 5 and 10 which were models for EEG data with 0, 1 or 2 spectral peaks and large deviations from white noise,
(v) moving average processes of order 2 and of order 4.

As regression functions we mainly used

$$g_1(t) = p_1 + p_2 t + p_3 \exp\{-(t-p_4)^2/p_5\},$$

with $p_1 = 2$, $p_2 = -5$, $p_3 = 5$, $p_4 = \frac{1}{2}$ and $p_5 = \frac{1}{100}$, and

$$g_2(t) = p_1 \sin(p_2 2\pi t),$$

with $p_1 = p_2 = 2$.

The sample size varied between 25 and 625. The number of replicates was 400. The following bandwidth selectors entered into the evaluation:

(i) $\hat{b}_m$, for fixed $0 \leq m \leq n^{\frac{1}{2}}$, where $\hat{b}_0$ is the naive plug-in estimator to be used when assuming independent residuals;
(ii) $\hat{b}_{(i)} = \hat{b}_{\hat{m}(i)}$, $\hat{b}_{(ii)} = \hat{b}_{\hat{m}(ii)}$, where $\hat{m}(i)$, $\hat{m}(ii)$ are obtained by selecting m with automatic methods (i), (ii) as in § 2;
(iii) $\hat{b}_{m(\mathrm{opt})}$ where $m(\mathrm{opt})$ is the sample optimum for m with respect to integrated squared error;
(iv) $\hat{b}(S)$ with the true S;
(v) b_I which is the sample optimal bandwidth, optimal with respect to integrated squared error.

Additionally we computed one automatic spline curve estimator (Silverman, 1984) which assumes independent residuals. In order to understand the properties of the different methods we evaluated mean, standard deviation and some characteristic quantiles of $\hat{m}(i)$, $\hat{m}(ii)$ and $m(\mathrm{opt})$, of the different bandwidths, and of the integrated squared error for the different curve estimators.

Now we illustrate some of the main results. For white noise the naive plug-in estimator and the spline curve estimator are as good as expected. But only a small price has to be paid when using $\hat{m}(i)$, $\hat{m}(ii)$ instead of $m = 0$. For example, for $n = 100$ and regression function g_1 the mean integrated squared error increased from 0·123 for $\hat{b}_0$ to 0·124 for $\hat{b}_{(i)}$ and 0·137 for $\hat{b}_{(ii)}$. However, substantial gains can usually be achieved by the new methods when correlation is present. For an AR(1) process with positive correlation the integrated squared error of estimates based on $\hat{m}(i)$, $\hat{m}(ii)$ follows closely the true optimum while the naive plug-in estimator and even more the spline estimator became worse, tending towards interpolating the data. This is not surprising (Diggle & Hutchinson, 1989). Table 1 illustrates the behaviour of median integrated squared error for $n = 100$, regression function g_1 and AR(1) processes. Here and in all other examples the estimator $\hat{b}(S)$ behaved by-and-large about as well as $\hat{b}_{(i)}$ and $\hat{b}_{(ii)}$ and consistently worse than $\hat{b}_{m(\text{opt})}$. This indicates that the auxiliary parameter m to account for correlation structure is a useful concept in bandwidth choice. For AR(1) processes with negative correlations the integrated squared error is smaller than in the case of independence and differences between the different methods become small.

Table 1. *Medians of the integrated squared error,* ISE, *in the case of* AR(1) *residuals*

Correl.	ISE($\hat{b}_0$)	ISE($\hat{b}_{(i)}$)	ISE($\hat{b}_{(ii)}$)	ISE(b_I)	ISE(spline)
−0·3	0·082	0·082	0·085	0·068	0·073
−0·1	0·100	0·100	0·108	0·094	0·092
0·0	0·118	0·118	0·121	0·110	0·107
0·1	0·138	0·134	0·140	0·127	0·137
0·3	0·231	0·194	0·187	0·171	0·349
0·5	0·415	0·272	0·268	0·240	0·773
0·7	0·622	0·426	0·404	0·370	0·887
0·9	0·844	0·741	0·704	0·648	0·936

For the AR(3) process with a broad peak and much high frequency power and $S = 0{\cdot}03\sigma^2$ none of the approaches came close to the optimum. They achieved roughly the performance of the naive plug-in rule which in case of predominantly negative correlations is already quite good. In this case a different approach would be needed for optimizing the bandwidth.

In the case of autoregressive processes of higher order, ignoring the correlation structure and using the naive plug-in selector yields consistently much too small a bandwidth close to interpolation; see also Fig. 1. Both modified plug-in rules based on selecting the tuning parameter m from the data behave reasonably well and similar to $\hat{b}_{m(\text{opt})}$ and $\hat{b}(S)$. Table 2 shows results for the bandwidths in a typical example, with $n = 100$, the sine regression function and an AR(10) process with two spectral peaks and $S = 6{\cdot}1$. The behaviour is also reflected in the mean relative inefficiency which is in this case 4·11 for the naive selector, 1·20 for the first rule and 1·13 for the second. The naive plug-in rule becomes worse for increasing n and more so the spline estimator while both modified versions of the plug-in rule behave close to the optimum. This is illustrated in Table 3 for median integrated squared error: the regression function g_1 is superposed by an AR(6) process with one distinct spectral peak and predominantly low frequency power otherwise and $S \simeq 3{\cdot}4\sigma^2$. For the moving average processes the results were again favourable for the proposed methods.

In summary, the goal of obtaining an objective regression fit with reasonable statistical performance, allowing for a broad class of regression functions and residual structure, seems realistic.

Table 2. *Characteristics of the different bandwidth estimators*

	$\hat{b}_0$	$\hat{b}_{(i)}$	$\hat{b}_{(ii)}$	b_I
10%	0·014	0·094	0·080	0·070
Mean	0·015	0·109	0·097	0·090
Median	0·015	0·108	0·097	0·087
90%	0·017	0·126	0·115	0·113

Table 3. *Medians of integrated squared error*, ISE, *for different n*; AR(6) *residuals*

n	ISE($\hat{b}_0$)	ISE($\hat{b}_{(i)}$)	ISE($\hat{b}_{(ii)}$)	ISE(b_I)	ISE(spline)
25	1·976	2·043	1·804	1·601	2·547
50	1·917	1·317	1·379	1·020	2·593
100	1·928	0·789	0·766	0·672	2·719
225	1·856	0·455	0·448	0·411	2·729
400	1·785	0·300	0·301	0·283	2·718
625	1·691	0·234	0·232	0·220	2·726

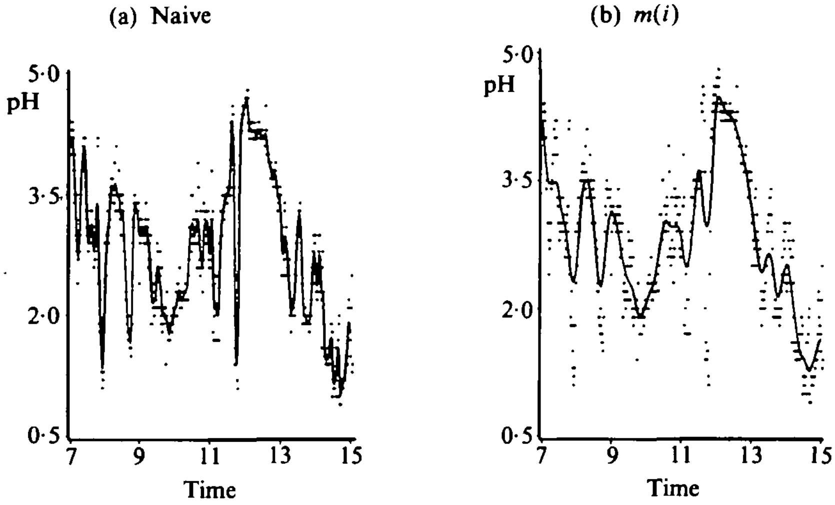

Fig. 2. Intragastric pH recordings from 7 a.m. to 3 p.m. (a) data and kernel estimate using naive bandwidth selection with $m = 0$; (b) data and kernel estimate using the modified plug-in bandwidth selection with method (i) and $\hat{m}(i) = 3$.

Figure 2 illustrates one application of the method to real data, i.e. intragastric pH measurements. The data are obtained by the following technique. An electrode is placed through the nose into the stomach and acidity is measured every 5 seconds continuously, usually for 24 hours, and the data stored on a small device carried in the pocket (Bauerfeind et al., 1987). Noise is usually correlated for physiological reasons and here probably also due to oversampling. Figure 2 shows a scatterplot on the left with a kernel fit obtained by naive bandwidth selection. Only every 10th measurement is depicted for graphical reasons. Evidently, the fit follows not only the signal but also the noise. The bandwidth selection based on the two approaches presented in this paper leads to a more meaningful pattern with little difference between the two methods: method (i) is illustrated here. The strong peak at about 13 hours is due to lunch: the meal reduces the acidity of the stomach significantly, an effect much appreciated by ulcer patients. The quantification of the effect of meals or antacid drugs on pH is facilitated by appropriate smoothers of the pH recordings.

ACKNOWLEDGEMENTS

This work has been performed as part of the research programme of the Sonderforschungsbereich 123 at the University of Heidelberg, and was made possible by financial support from the Deutsche Forschungsgemeinschaft.

Appendix

Proofs

Proof of Proposition 1. Note that $\hat{S}_m$ is a symmetric quadratic functional of the observations Y_i and can be written as $\hat{S}_m = \Sigma\, s_{i,j} Y_i Y_j$. Obviously there exists a constant $c_1 < \infty$ with $|s_{i,i+\nu}| \leqslant c_1 m/n$ and $|s_{i,i+\nu}| \leqslant c_1/n$, for $i = 1, \ldots, n-\nu$, $\nu = 0$, $\nu = m+1$, $\nu = 2m+2$ and $\nu = 1, \ldots, m$, $m+2, \ldots, 2m+1$ respectively. Hence $\operatorname{var}(\hat{S}_m) = O(m^2/n)$ follows from Assumption 2. The expression for the bias can be shown easily using the mean value theorem. □

Proof of Proposition 2. With lower and upper bounds for variance and bias the rest of the proof can be obtained easily. We restrict our attention to lower and upper bounds of the variance. Assumption 2 yields

$$\sum_{i \geqslant 1} \sum_{j \leqslant 0, j \geqslant n+1} |\gamma_{i-j}| = 2\left(\sum_{\nu=1}^{n-1} \nu|\gamma_\nu| + \sum_{\nu=n}^{\infty} n|\gamma_\nu|\right) < \infty.$$

Because of the Lipschitz continuity of W there exists a constant $c_1 < \infty$ with

$$\begin{aligned} \left|\operatorname{var}\{\hat{g}(t,b)\} - \frac{S}{nb}\int_{-1}^{1} W^2(x)\,dx\right| &= \left|\sum_{i,j=1}^{n} \gamma_{i-j} \int_{s_{i-1}}^{s_i} \int_{s_{j-1}}^{s_j} \frac{1}{b} W\left(\frac{t-u}{b}\right)\left\{W\left(\frac{t-w}{b}\right) - W\left(\frac{t-u}{b}\right)\right\} dw\,du \right.\\ &\quad \left. + \frac{1}{nb}\sum_{i=1}^{n}\left\{\int_{s_{i-1}}^{s_i} \frac{1}{b} W^2\left(\frac{t-u}{b}\right) du\right\} \sum_{j \leqslant 0, j \geqslant n+1} \gamma_{i-j}\right| \\ &\leqslant \frac{c_1}{n^2 b^2} \end{aligned}$$

for all $t \in [\delta, 1-\delta]$, $b \in [n^{-1}, \delta]$. The spectral density f of the error process is at least continuous because of Assumption 2. If it is positive we get for all $t \in [\delta, 1-\delta]$, $b \in [n^{-1}, \delta]$ and a suitable constant $c_2 > 0$

$$\begin{aligned} \operatorname{var}\{\hat{g}(t,b)\} &= \sum_{i,j=1}^{n} \left\{\int_{s_{i-1}}^{s_i} \frac{1}{b} W\left(\frac{t-u}{b}\right) du\right\}\left\{\int_{s_{j-1}}^{s_j} \frac{1}{b} W\left(\frac{t-w}{b}\right) dw\right\} \gamma_{i-j} \\ &\geqslant \left\{\min_{\lambda \in [0,\pi]} 2\pi f(\lambda)\right\} \sum_{i=1}^{n} \left\{\int_{s_{i-1}}^{s_i} \frac{1}{b} W\left(\frac{t-u}{b}\right) du\right\}^2 \\ &\geqslant \frac{c_2}{nb}. \end{aligned}$$

□

Lemma 1. *Suppose Assumption* 2 *holds. Let* $c(i)$, $a(i,j) \in \mathbf{R}$ *for all* $i, j = 1, \ldots, n$ *with* $a(i,j) = 0$ *if* $|i-j| \geqslant k/2$ *for some even integer* k. *Then*

(*i*) $$E\left\{\sum_{i,j} a(i,j)(\varepsilon_i \varepsilon_j - \gamma_{i-j})\right\}^{2s} \leqslant C_s n^s k^s \sup_{i,j} |a(i,j)|^{2s},$$

(*ii*) $$E\left\{\sum_i \varepsilon_i c(i)\right\}^{2s} \leqslant C_s n^2 \sup_i |c(i)|^{2s}$$

for all n, $s \in \mathbf{N}$ *with a constant* C_s *that depends only on* s *and the moments up to order* s *of the error series.*

This lemma can be proved by using the theory of cumulants. It is also given in an unpublished manuscript of Pham Din Tuan.

Lemma 2. *Suppose Assumptions* 1-4 *hold and let*

$$\tilde{C} = \int_0^1 v(t)\,dt \int_{-1}^{1} \{W_2(x)\}^2\,dx.$$

(*i*) *Suppose that* $\tilde{b} = cn^{-\alpha}\{1 + o_p(n^{-\beta})\}$ *with* $\frac{3}{10} \leqslant \alpha \leqslant \frac{9}{10}$ *and* $c > 0$, $0 \leqslant \beta < \frac{1}{10}$. *Then*

$$\int_0^1 v(t)\{\hat{g}_2(t, \tilde{b})\}^2\, dt = \frac{S\tilde{C}}{c^5} n^{-1+5\alpha}\{1 + o_p(n^{-\beta})\}.$$

(*ii*) *Suppose that* $b = cn^{-\alpha}\{1 + o_p(n^{-\beta})\}$ *with* $\alpha = \frac{1}{5}$ *and* $c > 0$, $0 \leqslant \beta < \frac{1}{10}$. *Then*

$$\int_0^1 v(t)\{\hat{g}_2(t, \tilde{b})\}^2\, dt = \int_0^1 v(t)\{g''(t)\}^2\, dt + \frac{S\tilde{C}}{c^5} + o(1) + o_p(n^{-\beta}).$$

(*iii*) *Suppose that* $\tilde{b} = cn^{-\alpha}\{1 + o(1) + o_p(n^{-\beta})\}$ *for* $\alpha = \frac{1}{10}$ *and some* $c > 0$, $\beta \geqslant 0$ *then*

$$\int_0^1 v(t)\{\hat{g}_2(t, \tilde{b})\}^2\, dt = \int_0^1 v(t)\{g''(t)\}^2\, dt + O(n^{-q/10}) + O_p(n^{-(3+q)/10} + n^{-\beta-q/10}).$$

Proof of Lemma 2. The estimated integral can be expressed as a sum of several variance and bias components,

$$\int v(t)\{\hat{g}_2(t, \tilde{b})\}^2\, dt = B(\tilde{b}) + M(\tilde{b}) + V_1(\tilde{b}) + V_2(\tilde{b}),$$

where

$$B(b) = \int_0^1 v(t)\left\{\sum_{i=1}^n \int_{s_{i-1}}^{s_i} \frac{1}{b^3} W_2\left(\frac{t-u}{b}\right) du\, g(t_i)\right\}^2,$$

$$M(b) = 2\int_0^1 v(t)\left\{\sum_{i=1}^n \varepsilon_i \sum_{j=1}^n \int_{s_{i-1}}^{s_i} \int_{s_{j-1}}^{s_j} \frac{1}{b^6} W_2\left(\frac{t-u}{b}\right) W_2\left(\frac{t-w}{b}\right) du\, dw\, g(t_j)\right\} dt,$$

$$V_1(b) = \int_0^1 v(t)\left\{\sum_{i=1}^n \sum_{j=1}^n \int_{s_{i-1}}^{s_i} \int_{s_{j-1}}^{s_j} \frac{1}{b^6} W_2\left(\frac{t-u}{b}\right) W_2\left(\frac{t-w}{b}\right) du\, dw\, \gamma_{i-j}\right\} dt,$$

$$V_2(b) = \int_0^1 v(t)\left\{\sum_{i=1}^n \sum_{j=1}^n \int_{s_{i-1}}^{s_i} \int_{s_{j-1}}^{s_j} \frac{1}{b^6} W_2\left(\frac{t-u}{b}\right) W_2\left(\frac{t-w}{b}\right) du\, dw(\varepsilon_i\varepsilon_j - \gamma_{i-j})\right\} dt.$$

Bounds for these terms can be obtained by analogue with the proofs of Lemmas 1 to 4 of Gasser et al. (1991), using our Lemma 1 instead of the Whittle inequality. □

Proof of Theorem 1. Firstly we compare the estimator $\hat{b}$ with the asymptotic optimal bandwidth b_A. If we start the iteration with $b_0 = n^{-1}$ the first eight iteration steps reduce the order of b_k. Additional iteration steps up to $k > 9 + 2/q$ reduce the order of the variability of b_k. This can be shown by induction. Note that $b_0 = 1/n$. Suppose first that

$$b_{k-1} = c_{k-1} n^{-(11-k)/10}\{1 + o_p(n^{-\beta_0})\}$$

for some $1 \leqslant k \leqslant 7$, $c_{k-1} > 0$ and some $0 \leqslant \beta_0 < \frac{1}{10}$ with $(S - \hat{S}) = o_p(n^{-\beta_0})$. Lemma 2(i) gives that

$$\int_0^1 v(t)\{\hat{g}_2(t, b_{k-1} n^{1/10})\}^2\, dt = \frac{S\tilde{C}}{c_{k-1}^5} n^{-1+(10-k)/2}\{1 + o_p(n^{-\beta_0})\}.$$

The whole iteration step implies that

$$b_k = \left[\frac{\hat{S}C_1}{nC_2^2 S\tilde{C} c_{k-1}^{-5} n^{-1+(10-k)/2}\{1 + o_p(n^{-\beta_0})\}}\right]$$

and from an appropriate Taylor expansion and the consistency of $\hat{S}$ follows

$$b_k = \left(\frac{C_1}{C_2\tilde{C}}\right)^{1/5} c_{k-1} n^{-(10-k)/10}\{1 + o_p(n^{-\beta_0})\}$$

for $k \leqslant 7$ which proves the induction step. The same argument using Lemma 2(ii) yields

$$\int_0^1 v(t)\{\hat{g}_2(t, b_7 n^{1/10})\}^2\,dt = \int_0^1 v(t)\{g''(t)\}^2\,dt + \frac{S\tilde{C}}{c_7^5} + o(1) + o_p(n^{-\beta}),$$

$$b_8 = C_1 S^{1/5}\left[nC_2^2 \int_0^1 v(t)\{g''(t)\}^2\,dt + nC_2^2 S\tilde{C}c_7^{-5}\right]^{-1/5}\{1 + o(1) + o_p(n^{-\beta_0})\}$$

$$= n^{-1/5} c_8\{1 + o(1) + o_p(n^{-\beta_0})\}.$$

Application of Lemma 2(iii) establishes the reduction of the rate of the variability for estimating the integral. Let

$$b_{k-1} = c_{k-1} n^{-1/5}\{1 + o(1) + O_p(n^{-q(k-1)/10-\beta_1}) + O_p(n^{-(3+q)/10}) + O_p(S - \hat{S})\}$$

for every $0 \leqslant \beta_1 < \frac{1}{10}$ and for $k \geqslant 9$, $c_{k-1} > 0$. The iteration step and application of Lemma 2(iii) gives

$$b_k = b_A + O(n^{-q/10}) + O_p(n^{-(3+q)/10}) + O_p(n^{-q(k-8)/10-\beta_1}) + O_p(S - \hat{S}),$$

$$b_k = b_A + O(n^{-q/10}) + O_p(n^{-(3+q)/10}) + O_p(S - \hat{S}),$$

for all $k > 9 + 2/q$. Applying Proposition 2 completes the proof. □

References

Altman, N. S. (1990). Kernel smoothing with correlated errors. *J. Am. Statist. Assoc.* **85**, 749-59.

Bauerfeind, P., Cilluffo, T., Fimmel, C., Emde, C., von Ritter, C., Köhler, W., Gugler, R., Gasser, T. & Blum, A. L. (1987). Does smoking interfere with the effect of histamine H$_2$ receptor antagonists on intragastric acidity in man? *Gut* **28**, 549-56.

Chiu, S. T. (1989). Bandwidth selection for kernel estimate with correlated noise. *Statist. Prob. Letters* **8**, 347-54.

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31**, 377-403.

Diggle, P. J. & Hutchinson, M. F. (1989). On spline smoothing with autocorrelated errors. *Aust. J. Statist.* **31**, 166-82.

Gasser, T., Kneip, A. & Köhler, W. (1991). A flexible and fast method for automatic smoothing. *J. Am. Statist. Assoc.* **86**, 643-52.

Gasser, T. & Müller, H. G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*, Ed. T. Gasser and M. Rosenblatt, pp. 23-68. New York: Springer.

Gasser, T., Müller, H. G. & Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *J. R. Statist. Soc.* B **47**, 238-52.

Gasser, T., Pietz, J., Schellberg, D. & Köhler, W. (1988). Visual evoked potentials of mildly mentally retarded and control children. *Develop. Med. Child Neurol.* **30**, 638-45.

Gasser, T., Sroka, L. & Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625-33.

Györfi, L., Härdle, W., Sarda, P. & Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series.* New York: Springer.

Hall, P. & Hart, J. D. (1990). Nonparametric regression with long-range dependence. *Stoch. Processes Applic.* **36**, 339-51.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press.

Härdle, W., Hall, P. & Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Am. Statist. Assoc.* **83**, 86-95.

Hart, J. D. (1991). Kernel regression estimation with time series errors. *J. R. Statist. Soc.* B **53**, 173-88.

Hurwich, C. M. (1985). Data-driven choice of a spectrum estimate: extending the applicability of cross-validation methods. *J. Am. Statist. Assoc.* **80**, 933-40.

Müller, H.-G. (1988). *Nonparametric Regression Analysis of Longitudinal Data.* Berlin: Springer.

Müller, H.-G. & Stadtmüller, U. (1988). Detecting dependencies in smooth regression models. *Biometrika* **75**, 639-50.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215-30.

Silverman, B. W. (1984). A fast and efficient cross-validation method for smoothing parameter choice in spline regression. *J. Am. Statist. Assoc.* **79**, 584-9.

[*Received November* 1990. *Revised September* 1991]