The Annals of Statistics
1990, Vol. 18, No. 2, 873–890

DATA-DRIVEN BANDWIDTH CHOICE FOR DENSITY ESTIMATION BASED ON DEPENDENT DATA[1]

BY JEFFREY D. HART AND PHILIPPE VIEU

Texas A & M University and Université Paul Sabatier

The bandwidth selection problem in kernel density estimation is investigated in situations where the observed data are dependent. The classical leave-out technique is extended, and thereby a class of cross-validated bandwidths is defined. These bandwidths are shown to be asymptotically optimal under a strong mixing condition. The leave-one out, or ordinary, form of cross-validation remains asymptotically optimal under the dependence model considered. However, a simulation study shows that when the data are strongly enough correlated, the ordinary version of cross-validation can be improved upon in finite-sized samples.

1. Introduction. Let $X_1, \ldots, X_n$ be identically distributed real random variables, and let f denote their common density function. One of the more popular nonparametric estimators of f is the Parzen–Rosenblatt kernel estimator. This estimator uses a kernel function K and a smoothing parameter h (depending on n) and is defined by

$$\hat{f}_n(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i),$$

where

$$K_h(u) = h^{-1}K(u/h).$$

Much of the enormous literature concerned with these estimators [see, e.g., the surveys by Devroye and Györfi (1984) and Silverman (1986)] points out the importance of making a reasonable choice of the bandwidth h of $\hat{f}_n$. In this paper we shall consider data-driven bandwidths that estimate the minimizer of the integrated squared error (ISE),

$$\text{ISE} = \text{ISE}(h) = \int \left(\hat{f}_n(x) - f(x)\right)^2 w(x)\, dx,$$

where w is some nonnegative weight function. The most often used data-driven criteria are based on cross-validation ideas [Rudemo (1982) and Bowman (1984)]. Our aim is to investigate the behavior of cross-validation when the observed data are not independent. In other contexts [see, e.g., Hart and Wehrly (1986)], it has been pointed out that cross-validation can produce much

Received April 1988; revised June 1989.

[1]Research supported in part by ONR Contract N00014-85-K-0723 while the second author was visiting the Department of Statistics at Texas A & M University.

AMS 1980 *subject classifications*. Primary 65G05, 62G20, 62M99; secondary 62M10, 60G10, 60G35.

Key words and phrases. Nonparametric density estimation, kernel estimate, bandwidth selection, α-mixing processes, cross-validation.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics*. www.jstor.org

under-smoothed estimates when the data are positively correlated. By modifying the leave-out technique involved in the cross-validation method, we construct a class of bandwidths that are asymptotically optimal when the data are α-mixing. An obvious application of our results is to the analysis of time series data. Estimating the marginal density f of a stationary time series is of interest since it is reasonable to use f as the predictive density for long-term forecasting.

The bandwidth of $\hat{f}_n$ will be selected to minimize the cross-validation criterion,

$$CV_{l_n}(h) = \int \hat{f}_n^2(x)w(x)\,dx - 2n^{-1}\sum_{i=1}^{n} \hat{f}_n^{(i)}(X_i)w(X_i). \tag{1.1}$$

The quantity $\hat{f}_n^{(i)}$ is the kernel estimator of f based on the observations X_j whose indices j are not too close to i, i.e.,

$$\hat{f}_n^{(i)}(x) = n_{l_n}^{-1} \sum_{|j-i|>l_n} K_h(x - X_j),$$

where $\{l_n\}$ is a sequence of positive integers, called the leave-out sequence, and n_{l_n} is such that

$$nn_{l_n} = \#\{(i,j): |i-j| > l_n\}.$$

Criterion (1.1) is designed for cases where the dependence between X_i and X_j decreases as $|i-j|$ increases. This is often the case, for example, when the X_i's are indexed by time. If X_i and X_j are actually independent for $|i-j| > l_n$, then criterion (1.1) is an unbiased estimator of the risk function

$$R(h) = E[\mathrm{ISE}(h)] - \int f^2(x)w(x)\,dx.$$

More generally, the difference between $R(h)$ and $E[CV_{l_n}(h)]$ will be a function of how highly dependent X_i and X_j are for $|i-j| > l_n$. The way in which we shall quantify strength of dependence is through an α-mixing condition (see Section 2.1).

Our main result is as follows: If the leave-out sequence l_n does not increase too fast, the bandwidth $\hat{h}(l_n)$ that minimizes $CV_{l_n}(h)$ is asymptotically optimal (see Section 2.2). Of particular interest is the fact that, as in the independence case [see Hall (1983), Stone (1984), Hall and Marron (1987) or Marron (1987)], cross-validation is still asymptotically optimal when we leave out only one point (i.e., $l_n = 0$). Proofs of these results will be given in Section 4. To obtain further insight as to the behavior of cross-validation under dependence, we performed a Monte Carlo experiment in which the dependence was first-order autoregressive. Our findings in the simulation study (see Section 3) support the contention that ordinary cross-validation ($l_n = 0$) is robust to moderate amounts of dependence in the data. However, we also show that some improvement in average ISE can be obtained by taking $l_n > 0$ when the data are sufficiently highly correlated.

2. Asymptotic optimality under dependence

2.1. *Dependence structures.* Let N denote the set of positive integers, and for any i and j in $N \cup \{\infty\}$ $(i \leq j)$, define M_i^j to be the σ-algebra spanned by the variables $X_i, X_{i+1}, \ldots, X_j$. The dependence among the observations is usually modelled by some mixing condition. The most often studied of these is the so-called φ-mixing condition. The sequence $\{X_i\}$ is said to be φ-mixing if there exist coefficients $\varphi(m)$ such that

$$\lim_{m \to \infty} \varphi(m) = 0,$$

and for positive integers k and m and for any sets A and B that are, respectively, M_1^k-measurable and M_{k+m}^{∞}-measurable,

$$|P(A \cap B) - P(A)P(B)| \leq \varphi(m)P(A).$$

A less restrictive dependence condition is the α-mixing condition due to Rosenblatt (1956). The sequence $\{X_i\}$ is said to be α-mixing if there exist mixing coefficients $\alpha(m)$ such that

$$\lim_{m \to \infty} \alpha(m) = 0,$$

and for positive integers k and m and for any sets A and B that are, respectively, M_1^k-measurable and M_{k+m}^{∞}-measurable,

$$|P(A \cap B) - P(A)P(B)| \leq \alpha(m).$$

For future reference we introduce the following notation:

$$\tilde{\alpha}(m) = \sup_{j \geq m} \alpha(j).$$

The φ-mixing condition is quite satisfactory for modelling most Markovian phenomena. [See, e.g., Rosenblatt (1971a) for sufficient conditions under which a Markov process is φ-mixing.] However, φ-mixing is too restrictive to include many interesting Gaussian processes [Ibragimov and Linnik (1971)]. For these processes the α-mixing condition is much more appropriate. The reader will find in the survey by Bradley (1985) more complete discussions of the various dependence structures. Chapter 5 in Hall and Heyde (1980) is also very useful for this topic. Consistency results for nonparametric density and regression estimators are found, e.g., in Collomb and Härdle (1986) and Sarda and Vieu (1988) for the φ-mixing case, and in Robinson (1983) or Roussas (1988) for the α-mixing case. Castellana and Leadbetter (1986) obtain consistency properties under another kind of dependence based on joint densities.

The present paper will deal with the α-mixing condition, which is less restrictive than the ones usually assumed in density estimation.

2.2. *Asymptotic optimality.* In order to get an asymptotic optimality property we make the following assumptions. The kernel function is assumed to be

such that

(K.1) $\exists\, \overline{K}$, $0 < \overline{K} < \infty$, such that $\forall\, x \in \mathscr{R}$, $|K(x)| \leq \overline{K}$;
(K.2) K is Lipschitz continuous, i.e., $\exists\, C_K$, $0 < C_K < \infty$, such that

$$|K(x) - K(y)| \leq C_K|x - y| \qquad \forall\, x, y \in \mathscr{R};$$

(K.3) K is compactly supported;
(K.4) K is symmetric;
(K.5) $\int K(x)\,dx = 1$, $\quad 0 < \int x^{\nu}K(x)\,dx < \infty$,

and $\int x^{k}K(x)\,dx = 0$, $\quad k = 1, \ldots, \nu - 1$;

(K.6) The Fourier transform of K is absolutely integrable.

The selected bandwidth is

$$\hat{h}(l_n) = \arg\min_{h \in H_n} CV_{l_n}(h), \tag{2.1}$$

where

(H.1) $H_n = [An^{-a}, Bn^{-b}]$, $\quad 0 < b \leq 1/(2\nu + 1) \leq a < 2/(1 + 4\nu)$,

and A and B are finite positive constants. The weight function w is such that

(W.1) w is bounded by 1 and $S = \text{support}(w)$ is compact.

The leave-out sequence $(l_n)_N$ and the mixing coefficients satisfy

(L.1) $l_n \leq l_n^* = n^{\tau_1}$ for some $0 < \tau_1 < (2 - a(1 + 4\nu))/2$,

and

(L.2) $\tilde{\alpha}(l_n^*) = o(n^{-\tau_2})$ for $\tau_2 = U + V + (2a + 4\nu a)(2 + U/V)$,

where

$$U = 1 + 2a + 2\nu a - b \quad \text{and} \quad V = 2 - a(1 + 4\nu) - 2\tau_1. \tag{2.2}$$

The nonparametric model is defined by the following assumptions on the density function:

(F.1) f has ν continuous derivatives, $\nu \in N$, $\nu \geq 1$;
(F.2) $\max(f(x), f(-x)) \to 0$ as $x \to \infty$;
(F.3) $\exists\, M_1$, $0 < M_1 < \infty$, such that $\forall\, x \in \mathscr{R}$, $f(x) \leq M_1$;
(F.4) For any j, (X_j, X_{j+1}) has a density f_j with respect to Lebesgue measure.

THEOREM 1. *Assume that the sequence $(X_i)_N$ is α-mixing and that conditions* (H.1), (W.1), (L.1), (L.2), (K.1)–(K.6) *and* (F.1)–(F.4) *hold. Then the cross-validated bandwidth defined by* (1.1) *and* (2.1) *is asymptotically optimal in the sense that, as $n \to \infty$,*

$$\text{ISE}\big(\hat{h}(l_n)\big) \Big/ \Big(\inf_{h \in H_n} \text{ISE}(h)\Big) \to 1 \quad a.s.$$

REMARK 2.1. We first note that condition (H.1) is reasonable since A and B can be chosen to insure that the set H_n contains the optimal bandwidth (i.e., the global minimizer of ISE), which is of order $n^{-1/(2\nu+1)}$. To clarify the

conditions (H.1), (L.1) and (L.2), consider the particular case where f has two continuous derivatives (i.e., when $\nu = 2$). In this case it is reasonable to choose $a = b = \frac{1}{5}$, and τ_1 can be taken, for example, to be $\frac{1}{20}$. Conditions (L.1) and (L.2) become, respectively,

$$l_n \leq l_n^* = n^{1/20} \quad \text{and} \quad \tilde{\alpha}(l_n^*) = o(n^{-46.1}).$$

Although these conditions may appear somewhat restrictive, they do allow the process to have algebraically decreasing mixing coefficients of the form

$$\alpha(m) = sm^{-t} \quad \text{for } t > 922.$$

This is at least a small improvement upon an assumption of geometrically decaying $\alpha(m)$.

REMARK 2.2. Theorem 1 could be stated similarly for φ-mixing data just by changing $\alpha(n)$ to $\varphi(n)$ in condition (L.2).

REMARK 2.3 (Choice of the leave-out sequence). Theorem 1 states that there exists a class of asymptotically optimal data-driven bandwidths indexed by the leave-out sequence $(l_n)_N$. Although it seems clear that the problem of choosing l_n is less important than the initial bandwidth selection problem, the simulation study in Section 3 shows that the number of data points left out does have some influence on the behavior of the cross-validated kernel estimate. An interesting problem would be to try to find a method of determining a good choice for l_n. While this problem is not theoretically addressed here, the results of our simulation yield some insight as to the effect of l_n.

REMARK 2.4 (Extensions). In order to make the presentation clearer, this paper deals only with univariate X_i. It is easy to see, by following the proofs, that all the results stated herein apply also in the setting of $\mathscr{R}^q$-valued X_i. For this it suffices to change a and b to qa and qb in conditions (L.1) and (L.2) and in formulas (C.1) (of Section 4) and (2.2). Another possible extension would be to use location-adaptive bandwidth selectors by letting w depend on n and on the location x, exactly as described for independent data in Mielniczuk, Sarda and Vieu (1989).

3. A simulation study. To investigate the finite sample behavior of cross-validation with dependent data, we performed a small-scale Monte Carlo study. The process $\{X_j\}$ we considered was first-order autoregressive and Gaussian, i.e.,

$$X_j = \rho X_{j-1} + Z_j, \qquad j \in N,$$

where $|\rho| < 1$ and the Z_j's are independent random variables all having the same normal distribution. As shown by Godoretskii (1977), this process is α-mixing with coefficients

$$\alpha(m) = C \sum_{j=m}^{\infty} j|\rho|^j = O(m|\rho|^m),$$

and thus satisfies the conditions of Theorem 1.

Each set of data in our simulation was obtained by generating a random sample $Y_1, \ldots, Y_n$ from the $N(0, 1)$ distribution and then taking $X_1 = Y_1$ and

$$X_j = \rho X_{j-1} + (1 - \rho^2)^{1/2} Y_j \quad \text{for } j \geq 2.$$

This produces X_j's that each have a $N(0, 1)$ distribution. We considered two sample sizes, $n = 50$ and 200, and four values for ρ (0, 0.3, 0.6 and 0.8). One hundred independent replications were performed for each of the eight combinations of n and ρ. Taking K to be a standard normal density, we computed the cross-validation curves $CV_{l_n}(h)$, $l_n = 0, 1, \ldots, 10$, for each replication. [Our choice of a normal kernel is concordant with assumption (K.3) since for computational reasons $K(x)$ must be set equal to 0 for $|x|$ sufficiently large.] For each l_n, the minimizer $\hat{h}(l_n)$ of $CV_{l_n}(h)$ for $h \in H_n$ was determined, where $H_{50} = \{0.05 + 1.95(i - 1)/49\colon\ i = 1, \ldots, 50\}$ and $H_{200} = \{0.02 + 1.48(i - 1)/49\colon\ i = 1, \ldots, 50\}$. Finally, for each set of data, we obtained $\tilde{h}$, the minimizer of ISE(h) over H_n, and we recorded ISE($\tilde{h}$) and ISE($\hat{h}(l_n)$), $l_n = 0, 1, \ldots, 10$. For convenience, we took the weight function w in the ISE to be identical to 1.

The results of the simulation are summarized in Figures 1–3 and Tables 1 and 2. Figures 1 and 2 provide evidence that ordinary cross-validation is reasonably robust to moderate amounts of correlation. Apparently, when $0 \leq \rho \leq 0.6$, one obtains at best a very small improvement in average ISE by choosing $l_n > 0$. However, for $\rho = 0.8$, a statistically significant improvement results from using $l_n > 0$ rather than $l_n = 0$. We reached this conclusion by performing Friedman's rank test, where each of the 100 replications was regarded as a block, and the responses within a block were ISE($\hat{h}(j)$), $j = 0, 1, \ldots, 10$. For $n = 50$, the P-value for the test of no differences among the 11 methods was 0.0001, while the P-value for a direct comparison of $l_n = 0$ and $l_n = 2$ was $(4.7)10^{-5}$. At $n = 200$ and $\rho = 0.8$, the P-values for both Friedman's test comparing the 11 methods and a direct comparison of $l_n = 0$ with $l_n = 4$ were 0.0002.

Some information concerning the behavior of the cross-validated bandwidths is given in Tables 1 and 2. A notable aspect of these tables is that when $\rho = 0.8$, $\hat{h}(0)$'s distribution is shifted to the left of $\tilde{h}$'s. This is an indication of the undersmoothing phenomenon we mentioned in the introduction. Taking $l_n > 0$ (for $\rho = 0.8$) makes the difference between $E(\hat{h}(l_n))$ and $E(\tilde{h})$ smaller. Also of interest is the strong positive correlation between $\hat{h}(0)$ and the other $\hat{h}(j)$, and the negative correlation between $\tilde{h}$ and each $\hat{h}(j)$. The discouraging negative correlation has been pointed out by Scott and Terrell (1987) in the setting of independent data.

Figure 3 reinforces the last few comments and shows why taking $l_n > 0$ does not yield any more of an improvement in ISE than it does. One can see that the regression of $\hat{h}(4)$ on $\tilde{h}$ is shifted above that of $\hat{h}(0)$ on $\tilde{h}$. This indicates that $\hat{h}(4)$ tends to be the better estimate of $\tilde{h}$ for large $\tilde{h}$, while $\hat{h}(0)$ tends to be better for small $\tilde{h}$. The fact that taking $l_n > 0$ yields some improvement in ISE is undoubtedly due to the fact that undersmoothing usually results in a larger ISE than does oversmoothing.

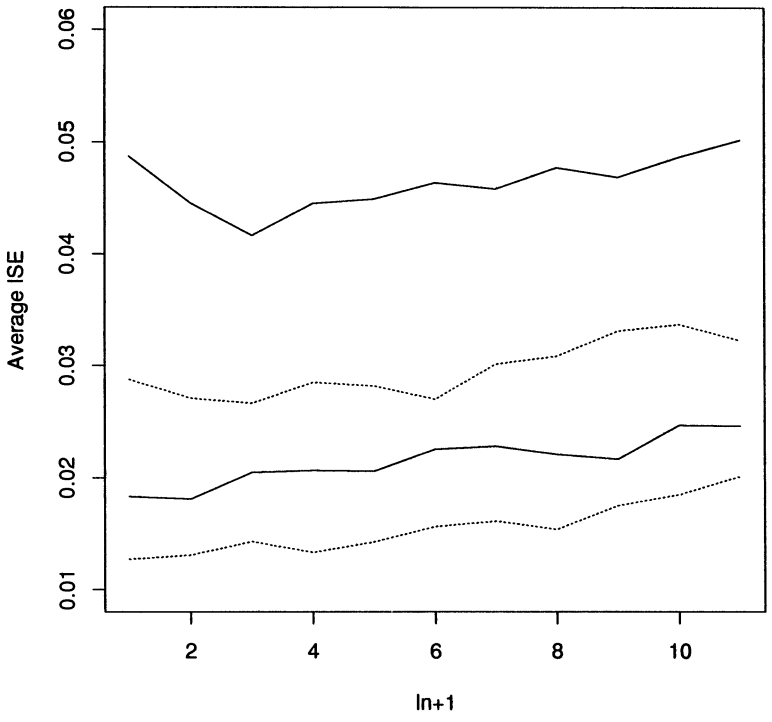


FIG. 1. *Average ISE as a function of* $l_n + 1$ ($n = 50$). *From top to bottom, the curves correspond to* $\rho = 0.8, 0.6, 0.3$ *and* 0. *Each curve is an average over* 100 *independent sets of data.*

Before concluding this section, some comments are in order concerning the choice of l_n. At this stage of our investigation we cannot give a definitive rule for choosing l_n. However, our simulation provides evidence that if the dependence among the data is not substantial, taking $l_n = 1$ or 2 will, on the average, not increase ISE by any important amount and may decrease it. On the other hand, if there is strong positive dependence among the data we definitely advise taking $l_n > 0$. A reasonable data analytic approach would be to examine the density estimates corresponding to $\hat{h}(0)$ and to the first few $\hat{h}(j)$, $j > 0$. One way of inferring the strength of dependence among the data would be to estimate the autocorrelation function of the observed process.

4. Proofs

4.1. *Proof of Theorem* 1. The asymptotic optimality property will follow [see Marron (1987)] from

$$\sup_{h \in H_n} \left| CT_{l_n}(h) \right| \Big/ \mathrm{ISE}(h) \to 0 \quad \text{a.s.} \tag{4.1}$$

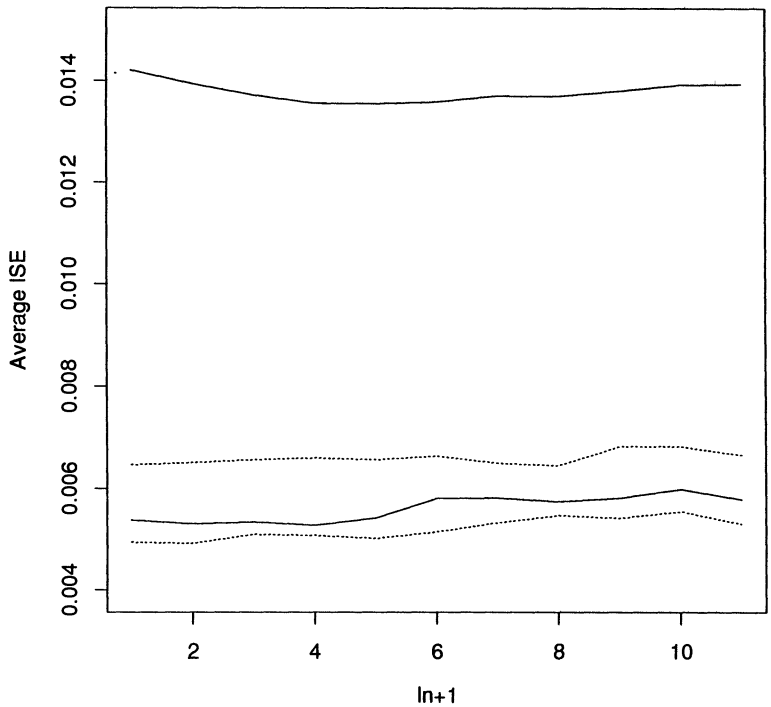


FIG. 2. *Average ISE as a function of* $l_n + 1$ ($n = 200$). *From top to bottom, the curves correspond to* $\rho = 0.8, 0.6, 0.3$ *and* 0. *Each curve is an average over* 100 *independent sets of data.*

where

$$CT_{l_n}(h) = n^{-1} \sum_{i=1}^{n} \hat{f}_n^{(i)}(X_i) w(X_i) - \int f(x) \hat{f}_n(x) w(x)\, dx$$
$$- n^{-1} \sum_{j=1}^{n} f(X_j) w(X_j) + \int f^2(x) w(x)\, dx.$$

Because of the following lemma it is enough to prove (4.1) for $l_n = l_n^*$.

LEMMA 1. *Under the conditions of Theorem* 1 *we have*

$$\sup_{h \in H_n} \left| CT_{l_n}(h) - CT_{l_n^*}(h) \right| \Big/ \mathrm{ISE}(h) \to 0 \quad a.s.$$

Now denote by H_n' a finite subset of H_n composed of equally spaced elements, and such that

(C.1) $$\#H_n' = n^{2a+2va-b+\xi} \quad \text{for some } 0 < \xi < V.$$

The following lemmas complete the proof.

TABLE 1
Summary statistics for cross-validated bandwidths ($n = 50$).

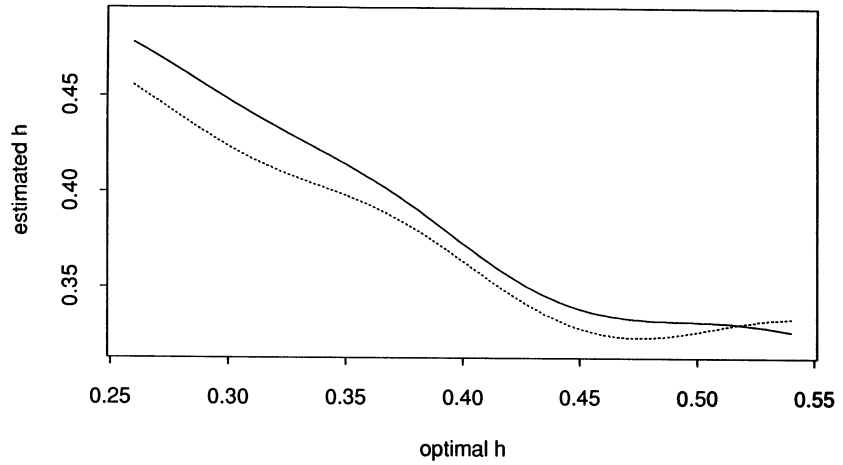
l_n	Mean of $\hat{h}(l_n)$	SD of $\hat{h}(l_n)$	$\hat{\rho}(\hat{h}, \tilde{h})$	$\hat{\rho}(\hat{h}, \hat{h}(0))$
	$\rho = 0.3$, mean($\tilde{h}$) = 0.512, SD($\tilde{h}$) = 0.101			
0	0.516	0.168	−0.544	1
1	0.529	0.176	−0.482	0.912
2	0.513	0.196	−0.450	0.804
3	0.516	0.197	−0.429	0.788
4	0.518	0.197	−0.449	0.845
5	0.508	0.212	−0.444	0.770
	$\rho = 0.8$, mean($\tilde{h}$) = 0.701, SD($\tilde{h}$) = 0.209			
0	0.411	0.186	−0.157	1
1	0.532	0.250	−0.199	0.782
2	0.625	0.289	−0.135	0.641
3	0.656	0.334	−0.116	0.578
4	0.683	0.363	−0.115	0.510
5	0.703	0.387	−0.093	0.484

Note: $\hat{\rho}$ denotes Pearson's correlation coefficient calculated from the 100 independent replications of an experiment, and SD means standard deviation.

TABLE 2
Summary statistics for cross-validated bandwidths ($n = 200$).

l_n	Mean of $\hat{h}(l_n)$	SD of $\hat{h}(l_n)$	$\hat{\rho}(\hat{h}, \tilde{h})$	$\hat{\rho}(\hat{h}, \hat{h}(0))$
	$\rho = 0.3$, mean($\tilde{h}$) = 0.371, SD($\tilde{h}$) = 0.070			
0	0.386	0.097	−0.517	1
1	0.400	0.090	−0.608	0.854
2	0.400	0.090	−0.670	0.813
3	0.404	0.089	−0.620	0.809
4	0.400	0.094	−0.616	0.849
5	0.394	0.104	−0.553	0.872
	$\rho = 0.8$, mean($\tilde{h}$) = 0.469, SD($\tilde{h}$) = 0.122			
0	0.341	0.102	−0.461	1
1	0.387	0.120	−0.522	0.931
2	0.412	0.127	−0.582	0.843
3	0.433	0.130	−0.597	0.773
4	0.444	0.133	−0.595	0.734
5	0.453	0.136	−0.602	0.711

Note: $\hat{\rho}$ denotes Pearson's correlation coefficient calculated from the 100 independent replications of an experiment, and SD means standard deviation.



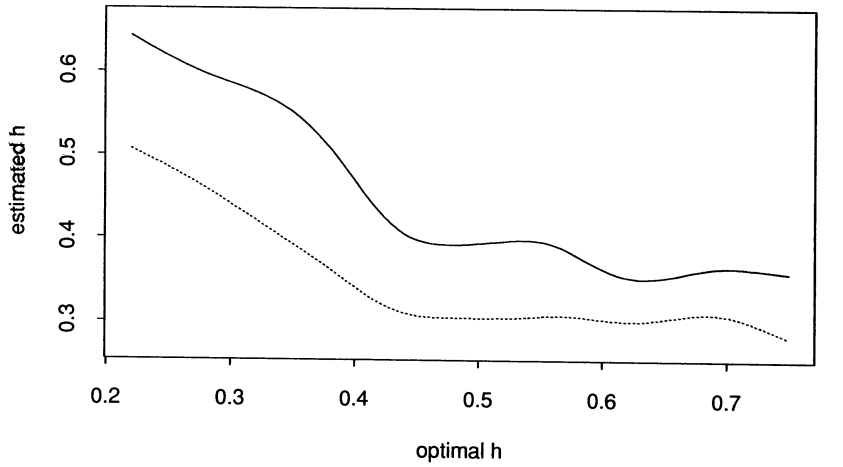


FIG. 3. *Estimated regression of* $\hat{h}(j)$, $j = 0$ *and* 4, *on the optimal bandwidth* $\tilde{h}$ ($n = 200$, $\rho = 0.3$ *and* 0.8). *Each of the four curves is a smoothed scatter plot of* $\hat{h}(j)$ *vs.* $\tilde{h}$. *In each plot, the top (bottom) curve is the estimated regression of* $\hat{h}(4)$ ($\hat{h}(0)$) *on* $\tilde{h}$. *The upper (lower) plot is based on* $\rho = 0.3$ (0.8).

LEMMA 2. *Under the conditions of Theorem* 1 *we have*

$$\sup_{h \in H_n'} \left| CT_{l_n^*}(h) \right| \Big/ \mathrm{ISE}(h) \to 0 \quad a.s.$$

LEMMA 3. *Under the conditions of Theorem* 1 *we have*

$$\sup_{h \in H_n} \left| CT_{l_n^*}(h) - CT_{l_n^*}(h^*) \right| \Big/ \mathrm{ISE}(h) \to 0 \quad a.s.,$$

where, for each h *in* H_n, h^* *denotes the element of* H_n' *that is closest to* h.

An important tool in proving these results is inequality (4.2) below. The ISE is essentially the sum of two positive components [see, e.g., Rosenblatt (1971b)], a variance component and a squared bias component. The squared bias component,

$$b(h) = \int \left(E\hat{f}_n(x) - f(x)\right)^2 w(x)\, dx,$$

does not depend on the multidimensional distribution of $X_1, \ldots, X_n$, and is therefore the same as for independent samples. Following the techniques previously described by Parzen (1962) in the case $\nu = 2$, one can show by Taylor expansion of f that under the conditions (F.1), (W.1) and (K.5), $b(h)$ is asymptotically of the order $h^{2\nu}$. It now follows from Theorem 2.1 in Vieu (1989) that there exists a finite positive constant C_b such that for n large enough we have for any h in H_n,

$$\text{ISE}(h) \geq C_b h^{2\nu}. \tag{4.2}$$

4.2. *Main mathematical tool.* Asymptotic optimality properties are usually obtained in the case of independent data [see, e.g., Hall (1983), Stone (1984) or Marron (1987)] by using certain bounds on moments of sums of i.i.d. variables. To deal with the dependence structure described above we need to have some "equivalent" tool, namely the following proposition.

PROPOSITION 1. *Let K be a kernel function satisfying the conditions* (K.1)–(K.6) *and assume that the sequence $(X_i)_N$ is α-mixing. Let $j(1), \ldots, j(p)$ be p distinct positive integers, and define*

$$g(X_{j(1)}, \ldots, X_{j(p)}) = \prod_{r=1}^{q} \prod_{\substack{s=1 \\ s \neq r}}^{q} K_h(X_{j(r)} - X_{j(s)})^{\beta_{r,s}} \prod_{i=1}^{p} g_{j(i)}(X_{j(i)}), \tag{4.3}$$

where the $g_{j(i)}$ are real-valued functions such that $|g_{j(i)}(\cdot)| \leq M_i < \infty$, the $\beta_{r,s}$ are nonnegative integers and $q \leq p$. Let $A_1, \ldots, A_v$ be a partition of $\{j(1), \ldots, j(p)\}$. Then, there exists a finite positive constant G such that

$$\left| \int g\, dP_{(X_{j(1)}, \ldots, X_{j(p)})} - \int g\, dP_{(X_t, t \in A_i)} \cdots dP_{(X_t, t \in A_v)} \right| \leq G \prod_{i=1}^{p} M_i h^{-\beta} \tilde{\alpha}(d),$$

where

$$d = \inf\{d(A_i, A_j) : i, j = 1, \ldots, v, i \leq j\},$$

$$d(A_i, A_j) = \inf\{|u - u'|, u \in A_i, u' \in A_j\},$$

and

$$\beta = \sum_{r=1}^{q} \sum_{\substack{s=1 \\ s \neq r}}^{q} \beta_{r,s}.$$

PROOF. Such a result already exists for functions g of the form $g(u_1, \ldots, u_p) = \prod_{i=1}^{p} g_i(u_i)$ [Hall and Heyde (1980), Theorem A5, page 277].

Here, the main idea behind the proof is to express our g as

$$g(X_{j(1)},\dots,X_{j(p)}) = \prod_{r=1}^{q}\prod_{\substack{s=1\\ s\neq r}}^{q}\left[\frac{1}{2\pi}\int \xi(hy)\exp\{-iy(X_{j(r)}-X_{j(s)})\}\,dy\right]^{\beta_{r,s}} \times \prod_{i=1}^{p} g_{j(i)}(X_{j(i)}),$$

where ξ is the Fourier transform of K. It is now clear that g may be written as a multiple integral whose integrand (as a function of $X_{j(1)},\dots,X_{j(p)}$) is of the form $\prod_{r=1}^{p} h_r(X_{j(r)})$. An induction argument and Theorem A5 of Hall and Heyde (1980) can now be used to obtain the result. For the sake of brevity the details are omitted. □

4.3. *Proof of Lemma* 1. Defining $K_h(X_r - X_s)$ to be 0 when r or s is not in $\{1,\dots,n\}$, we can write

$$\left|CT_{l_n}(h) - CT_{l_n^*}(h)\right| \le |R_1(h)| + |R_2(h)|,$$

where

$$R_1(h) = (nn_{l_n})^{-1}\sum\sum_{l_n<|i-j|\le l_n^*} K_h(X_i - X_j)w(X_i),$$

and

$$R_2(h) = \left((nn_{l_n})^{-1} - (nn_{l_n^*})^{-1}\right)\sum\sum_{|i-j|>l_n^*} K_h(X_i - X_j)w(X_i).$$

Writing

$$R_1(h) = (nn_{l_n})^{-1}\sum_{|k|=l_n+1}^{l_n^*}\sum_{i=1}^{n} K_h(X_i - X_{i-k})w(X_{i-k}),$$

we get by using (W.1), (K.1) and (K.3)

$$|R_1(h)| \le 2(l_n^* - l_n)(n_{l_n})^{-1}\int |K(u)|\,du \sup_{u\in S}\bar{f}_n(u),$$

where $\bar{f}$ is the kernel estimate of f constructed with the kernel function $|K|/\int|K(u)|\,du$. Proposition 4.1 in Roussas (1988) gives, together with (F.3), an almost sure finite bound for $\sup_{u\in S}\bar{f}_n(u)$ which is uniform over $h\in H_n$. Therefore, using the fact that $n = O(n_{l_n})$,

$$\sup_{h\in H_n}|R_1(h)| = O\left(n^{-1}l_n^*\right) \quad \text{a.s.}$$

Similarly, we have

$$|R_2(h)| \le \left((n_{l_n^*})^{-1} - (n_{l_n})^{-1}\right)(n - 2l_n^*)\int |K(u)|\,du \sup_{u\in S}\bar{f}_n(u),$$

and, as before, this leads to

$$\sup_{h\in H_n} |R_2(h)| = O(n^{-1}l_n^*) \quad \text{a.s.}$$

Finally, Lemma 1 follows from these two results together with (L.1) and (4.2).

4.4. *Proof of Lemma* 2. We first write

$$CT_{l_n^*}(h) = (nn_{l_n^*})^{-1} \sum\sum_{|i-j|>l_n^*} \Gamma(i,j),$$

where

$$\Gamma(i,j) = K_h(X_i - X_j)w(X_j) - \int f(x)K_h(x - X_i)w(x)\,dx$$

$$- f(X_j)w(X_j) + \int f^2(x)w(x)\,dx.$$

Now define

$$\Gamma^*(j) = \int K_h(u - X_j)w(X_j)f(u)\,du - \iint K_h(x-u)f(x)w(x)f(u)\,du\,dx$$

$$- f(X_j)w(X_j) + \int f^2(x)w(x)\,dx,$$

$$\psi(i,j) = \Gamma(i,j) - \Gamma^*(j) \quad \text{and} \quad T(h) = \sum\sum_{|i-j|>l_n^*} \psi(i,j).$$

Noting that $n_{l_n^*} \sim n$ and using (4.2), all we have to prove is that

$$\sup_{h\in H_n'} n^{-2}h^{-2\nu}|T(h)| \to 0 \quad \text{a.s.} \tag{4.4}$$

and

$$\sup_{h\in H_n'} n^{-1}h^{-2\nu}\left|\sum_{j=1}^{n}\Gamma^*(j)\right| \to 0 \quad \text{a.s.} \tag{4.5}$$

The proof of (4.5) is given in Marron [(1987), Formula 7.3] when the data are independent. The proof in our case proceeds over the same steps by using the exponential inequality for α-mixing variables given by Roussas [(1988), Theorem A.2] in place of Bernstein's classical one.

All that remains is to prove (4.4). For this, decompose the term $T(h)$ in the following way:

$$T(h) = T^+(h) + T^-(h), \tag{4.6}$$

where

$$T^+(h) = \sum\sum_{i+l_n^*<j\le n} \psi(i,j) \quad \text{and} \quad T^-(h) = \sum\sum_{1\le j<i-l_n^*} \psi(i,j).$$

We just give the details concerning T^+; the same argument can be applied to

T^-. Define $\psi(i, j) = 0$ when (i, j) is not in the set $\{(i, j): i + l_n^* < j \le n\}$. The first step is to rewrite T^+ in the following way:

$$T^+(h) = \sum_{s=0}^{1} \sum_{t=0}^{1} \sum_{j=1}^{l_n^*} \sum_{k=1}^{l_n^*} \sum_{q=1}^{n_1} \sum_{m=1}^{q} \psi(k + 2(m-1)l_n^* + sl_n^*, j + (2q-1)l_n^* + tl_n^*),$$

where n_1 is the greatest integer less than or equal to $n/(2l_n^*)$. We will now bound for some integer p the moments of order $2p$ of T^+. By Minkowski's inequality,

$$E\left[T^+(h)^{2p}\right] \le (2l_n^*)^{4p} \sup_{j,k,s,t} E\left[\left(\sum_{q=1}^{n_1} \sum_{m=1}^{q} \psi(m', q')\right)^{2p}\right], \tag{4.7}$$

where for j, k, s and t fixed we use the simplified notation,

$$m' = k + 2(m-1)l_n^* + sl_n^* \quad \text{and} \quad q' = j + (2q-1)l_n^* + tl_n^*.$$

REMARK 4.1. The following facts are important in the sequel. If m_1 differs from m_2, then $|m_1' - m_2'| > l_n^*$. Similarly, if q_1 differs from q_2, then $|q_1' - q_2'| > l_n^*$. Finally, whenever m differs from q and from $q+1$, $|m' - q'| > l_n^*$.

We now consider the moments [which exist by (F.1)–(F.3)]

$$E\left[\left(\sum_{q=1}^{n_1} \sum_{m=1}^{q} \psi(m', q')\right)^{2p}\right] = \sum_{q_1=1}^{n_1} \sum_{m_1=1}^{q_1} \cdots \sum_{q_{2p}=1}^{n_1} \sum_{m_{2p}=1}^{q_{2p}} E\left[\prod_{i=1}^{2p} \psi(m_i', q_i')\right].$$

Now define

$$I = \{J = (m_1, q_1, \ldots, m_{2p}, q_{2p}): 1 \le m_i \le q_i \le n_1, i = 1, \ldots, 2p\},$$

and consider the subset I_1 of I composed of the $4p$-tuples $J = (m_1, q_1, \ldots, m_{2p}, q_{2p})$ for which at least one index among $(m_1, q_1, \ldots, m_{2p}, q_{2p})$ differs from all other indices by at least 2. An important point is that

$$\#I_1 = O(n_1^{4p}) \quad \text{and} \quad \#(I - I_1) = O(n_1^{2p}). \tag{4.8}$$

We now investigate

$$E\left[\left(\sum_{q=1}^{n_1} \sum_{m=1}^{q} \psi(m', q')\right)^{2p}\right] = \sum_{J \in I_1} EZ(J) + \sum_{J \in I - I_1} EZ(J), \tag{4.9}$$

where for $J = (m_1, q_1, \ldots, m_{2p}, q_{2p})$, $Z(J)$ is defined by

$$Z(J) = \prod_{i=1}^{2p} \psi(m_i', q_i').$$

First, consider $J = (m_1, q_1, \ldots, m_{2p}, q_{2p})$ in I_1 and assume that $m_{i(0)}$ is an index which differs from all others in J by at least 2. (The proof would be similar if this index were some q_j.) One can write $Z(J)$ as a finite sum of

terms each of the form (4.3). Proposition 1 can then be applied to each of these terms by taking (using the notation introduced in the proposition)

$$v = 3, \qquad \beta = 2p,$$

$$A_1 = \{i \in \{m'_1, q'_1, \ldots, m'_{2p}, q'_{2p}\}, i < m'_{i(0)}\},$$

$$A_2 = \{m'_{i(0)}\}$$

$$A_3 = \{i \in \{m'_1, q'_1, \ldots, m'_{2p}, q'_{2p}\}, i > m'_{i(0)}\}.$$

Note that by Remark 4.1 we have $d \geq l_n^*$. By Proposition 1 there is a finite positive constant C_1 for which

$$|EZ(J)| \leq C_1 h^{-2p} \tilde{\alpha}(l_n^*) + \left| \int \left[\int \left[\int Z(J)\, dP_{X_{m'_{i(0)}}} \right] dP_{(X_t, t \in A_1)} \right] dP_{(X_t, t \in A_3)} \right|,$$

and, by definition of $\psi(i, j)$, the second term on the right-hand side of this inequality is 0. Along with (4.8) this leads to

$$\left| \sum_{J \in I_1} EZ(J) \right| = O\left(n_1^{4p} h^{-2p} \tilde{\alpha}(l_n^*)\right). \tag{4.10}$$

Now consider indices J in the set $I - I_1$ and define the partition of $I - I_1$,

$$I - I_1 = \bigcup_{\gamma=1}^{4p} I^\gamma,$$

where $I^\gamma = \{J = (m_1, q_1, \ldots, m_{2p}, q_{2p}) \in I - I_1 : \#\{m_1, q_1, \ldots, m_{2p}, q_{2p}\} = \gamma\}$. Let $J = (m_1, q_1, \ldots, m_{2p}, q_{2p})$ be in the set I^γ, and denote by $m_{a(1)}, \ldots, m_{a(\gamma_1)}, m_{a(\gamma_1+1)}, \ldots, m_{a(\gamma_2)}, q_{b(1)}, \ldots, q_{b(\gamma_1)}, q_{b(\gamma_1+1)}, \ldots, q_{b(\gamma_3)}$ (where $\gamma_2 + \gamma_3 = \gamma$) the γ distinct elements of $\{m_1, q_1, \ldots, m_{2p}, q_{2p}\}$, classified in such a way that

$$\text{for } r = 1, \ldots, \gamma_1, \qquad m_{a(r)} = q_{b(r)} + 1,$$

$$\text{for } r = \gamma_1 + 1, \ldots, \gamma_2, \quad m_{a(r)} \neq q_{b(j)} + 1, \qquad j = 1, \ldots, \gamma_3,$$

and

$$\text{for } r = \gamma_1 + 1, \ldots, \gamma_3, \quad q_{b(r)} \neq m_{a(j)} - 1, \qquad j = 1, \ldots, \gamma_2.$$

Now apply Proposition 1 with

$$v = \gamma - \gamma_1, \qquad \beta = 2p,$$

$$A_r = \{m'_{a(r)}, q'_{b(r)}\} \quad \text{for } r = 1, \ldots, \gamma_1,$$

$$A_r = \{m'_{a(r)}\} \quad \text{for } r = \gamma_1 + 1, \ldots, \gamma_2,$$

$$A_r = \{q'_{b(r-\gamma_2+\gamma_1)}\} \quad \text{for } r = \gamma_2 + 1, \ldots, \gamma - \gamma_1,$$

and note that Remark 4.1 implies that $d \geq l_n^*$. Hence, for some finite positive

constant C_2

$$|EZ(J)| \le C_2 h^{-2p}\tilde{\alpha}(l_n^*) + \left|\int Z(J)\prod_{r=1}^{v} dP_{(X_t, t\in A_r)}\right|.$$

By using (F.4) and integration by substitution, one gets

$$\left|\int Z(J)\prod_{r=1}^{v} dP_{(X_t, t\in A_r)}\right| = O(h^{-2p+\gamma/2}).$$

A result much like the last one is given by Marron and Härdle [(1986), Formula 3.4]. [See also Mielniczuk, Sarda and Vieu (1989).] In our case the proof would be the same.

The last two results lead to

$$|EZ(J)| = O\big(h^{-2p}\tilde{\alpha}(l_n^*)\big) + O(h^{-2p+\gamma/2}) \quad \text{for any } J \in I^{\gamma}. \tag{4.11}$$

Noting that $\#I^{\gamma} = O(n_1^{\gamma})$, and, because of (4.8), $\#I^{\gamma} = O(n_1^{2p})$, we have for some finite positive constant C_3

$$\begin{aligned}&\left|\sum_{J\in I-I_1} EZ(J)\right| \\ &\le C_3\left(n_1^{2p}h^{-2p}\tilde{\alpha}(l_n^*) + \sum_{\gamma=1}^{2p} n_1^{\gamma}h^{-2p+\gamma/2} + \sum_{\gamma=2p+1}^{4p} n_1^{2p}h^{-2p+\gamma/2}\right) \\ &= O\big(n_1^{2p}h^{-2p}\tilde{\alpha}(l_n^*)\big) + O\big(n_1^{2p}h^{-p}\big).\end{aligned} \tag{4.12}$$

To get (4.12) we have used the fact that the two sums to the right of the last inequality sign are geometric sums.

It follows from (4.7), (4.9), (4.10), (4.12) and the fact that n_1 is of the same order as n/l_n^*, that

$$E\big[T^+(h)^{2p}\big] = O\big(n^{4p}h^{-2p}\tilde{\alpha}(l_n^*)\big) + O\big(n^{2p}(l_n^*)^{2p}h^{-p}\big).$$

Now, using Chebyshev's and Boole's inequalities, we have

$$P\left[\sup_{h\in H_n'} n^{-2}h^{-2\nu}|T^+(h)| > \varepsilon\right] \le \#H_n'\varepsilon^{-2p}\sup_{h\in H_n'} E\big[\big(n^{-2}h^{-2\nu}T^+(h)\big)^{2p}\big].$$

By taking p to be the integer such that $1 + U/V < p \le 2 + U/V$, and using (L.1), (L.2), (H.1) and (C.1) together with the last two expressions, the Borel–Cantelli lemma implies that

$$\sup_{h\in H_n'} n^{-2}h^{-2\nu}|T^+(h)| \to 0 \quad \text{a.s.}$$

Using (4.6) and exactly the same proof as above for $T^-(h)$ yields (4.4).

4.5. *Proof of Lemma* 3. We can write

$$CT_{l_n^*}(h) - CT_{l_n^*}(h^*) = \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV}, \tag{4.13}$$

where

$$\mathrm{I} = (nn_{l_n^*})^{-1} \sum\sum_{|i-j|>l_n^*} (1/h - 1/h^*)K\big((X_i - X_j)/h\big)w(X_i),$$

$$\mathrm{II} = (nn_{l_n^*})^{-1} \sum\sum_{|i-j|>l_n^*} (1/h^*)\big[K\big((X_i - X_j)/h\big) - K\big((X_i - X_j)/h^*\big)\big]w(X_i),$$

$$\mathrm{III} = n^{-1} \sum_i \int (1/h^* - 1/h)K((x - X_i)/h)f(x)w(x)\,dx$$

and

$$\mathrm{IV} = n^{-1} \sum_i \int (1/h^*)\big[K((x - X_i)/h^*) - K((x - X_i)/h)\big]f(x)w(x)\,dx.$$

Because the points in H_n' are equally spaced, we have by (H.1),

$$1/h - 1/h^* = O\big((\#H_n')^{-1}n^{2a-b}\big). \tag{4.14}$$

Therefore, because K and w are bounded,

$$|\mathrm{I}| + |\mathrm{III}| = O\big((\#H_n')^{-1}n^{2a-b}\big). \tag{4.15}$$

Since K is Lipschitz continuous and compactly supported,

$$\big|K\big((X_i - X_j)/h\big) - K\big((X_i - X_j)/h^*\big)\big| = O(h^*|1/h - 1/h^*|). \tag{4.16}$$

Obviously, (4.16) remains valid with $x - X_i$ in place of $X_i - X_j$, and so by (4.14), (4.16) and (H.1),

$$|\mathrm{II}| + |\mathrm{IV}| = O\big((\#H_n')^{-1}n^{2a-b}\big). \tag{4.17}$$

It follows from (4.2), (H.1), (C.1), (4.13), (4.15) and (4.17) that

$$\sup_{h\in H_n} \big|CT_{l_n^*}(h) - CT_{l_n^*}(h^*)\big| \Big/ \mathrm{ISE}(h) = O(n^{-\xi}).$$

Acknowledgments. The authors express their gratitude to two referees and an Associate Editor whose careful reading pointed out a number of errors in the first draft of this paper.

REFERENCES

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

BOWMAN, A. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71** 353–360.

BRADLEY, R. (1985). Some remarks on strong mixing conditions. In *Proc. 7th Conf. in Probab. Theory, Brasov, Romania, 1982* (M. Losifescu, ed.) 65–72. Published jointly by Editura Academici Republicii, Romania, and VNU Science Press, Utrecht.

CASTELLANA, J. V. and LEADBETTER, M. R. (1986). On smoothed probability density estimation for stationary processes. *Stochastic Process. Appl.* **21** 179–193.

COLLOMB, G. and HÄRDLE, W. (1986). Strong uniform convergence rates in robust nonparametric time series analysis. *Stochastic Process. Appl.* **23** 77–89.

DEVROYE, L. and GYŐRFI L. (1984). *Nonparametric Density Estimation: The L_1 View*. Wiley, New York.

GORODETSKII, V. V. (1977). On the strong mixing condition for linear sequences. *Theory Probab. Appl.* **22** 411–413.

HALL, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11** 1156–1174.

HALL, P. and HEYDE, C. (1980). *Martingale Limit Theory and Its Applications*. Academic, New York.

HALL. P. and MARRON, J. S. (1987). On the amount of noise inherent in bandwidth selection for a kernel density estimator. *Ann. Statist.* **15** 163–181.

HART, J. and WEHRLY, T. (1986). Kernel regression estimation using repeated measurements data. *J. Amer. Statist. Assoc.* **81** 1080–1088.

IBRAGIMOV, I. A. and LINNIK, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff, Groningen.

MARRON, J. S. (1987). A comparison of cross-validation techniques in density estimation. *Ann. Statist.* **15** 152–162.

MARRON, J. S. and HÄRDLE, W. (1986). Random approximation to some measures of accuracy in nonparametric estimation. *J. Multivariate Anal.* **20** 91–113.

MIELNICZUK, J., SARDA, P. and VIEU, P. (1989). Local data-driven bandwidth choice for density estimation. *J. Statist. Plann. Inf.* **23** 53–69.

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065–1079.

ROBINSON, P. M. (1983). Nonparametric estimates for time series. *J. Time Series Anal.* **4** 185–201.

ROSENBLATT, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci.* **42** 43–47.

ROSENBLATT, M. (1971a). *Markov Processes: Structure and Asymptotic Behavior*. Springer, Berlin.

ROSENBLATT, M. (1971b). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.

ROUSSAS, G. (1988). Nonparametric regression estimation under mixing conditions. Preprint.

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimates. *Scand. J. Statist.* **9** 65–78.

SARDA, P. and VIEU, P. (1988). Empirical distribution function for mixing random variables. Application in nonparametric hazard estimation. *Statistics.* **20** 559–571.

SCOTT, D. and TERRELL, G. (1987). Biased and unbiased cross-validation in density estimation. *J. Amer. Statist. Assoc.* **82** 1131–1146.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimation. *Ann. Statist.* **12** 1285–1297.

VIEU, P. (1989). Quadratic errors for nonparametric estimates under dependence. Unpublished manuscript.

DEPARTMENT OF STATISTICS
TEXAS A & M UNIVERSITY
COLLEGE STATION, TEXAS 77843-3143

LABORATOIRE DE STATISTIQUE ET PROBABILITÉS
UNIVERSITÉ PAUL SABATIER, UA CNRS 745
31062 TOULOUSE
FRANCE