Probab. Th. Rel. Fields 74, 567–581 (1987)

Probability Theory and Related Fields

Extent to which Least-Squares Cross-Validation Minimises Integrated Square Error in Nonparametric Density Estimation

Peter Hall [1] * and James Stephen Marron [2] **

[1] Australian National University, Dept. of Statistics, Faculty of Economics and Commerce, GPO Box 4, Canberra ACT 2601, Australia

[2] University of North Carolina, Chapel Hill

Summary. Let h_0, $\hat{h}_0$ and $\hat{h}_c$ be the windows which minimise mean integrated square error, integrated square error and the least-squares cross-validatory criterion, respectively, for kernel density estimates. It is argued that $\hat{h}_0$, not h_0, should be the benchmark for comparing different data-driven approaches to the determination of window size. Asymptotic properties of $h_0-\hat{h}_0$ and $\hat{h}_c-\hat{h}_0$, and of differences between integrated square errors evaluated at these windows, are derived. It is shown that in comparison to the benchmark $\hat{h}_0$, the observable window $\hat{h}_c$ performs as well as the so-called "optimal" but unattainable window h_0, to both first and second order.

1. Introduction

Let $X_1, \ldots, X_n$ be a random sample from a distribution with unknown density f on $\mathbb{R}$, and let

$$f_n(x|h) \equiv (nh)^{-1} \sum_{i=1}^{n} K\{(x-X_i)/h\}$$

be a nonparametric estimator of f based on kernel K and window h. The problem of choosing h so as to "minimise error", in some sense, is legion in the theory and practice of nonparametric density estimation. Commonly, the criterion used to measure loss is mean integrated square error (MISE),

$$M(h) \equiv \int E\{f_n(x|h)-f(x)\}^2\, dx.$$

See for example Rosenblatt [17]. This approach has its roots in classical theory of nonparametric density estimation, where the window h is taken to be non-random. Of course, the value h_0 which minimises $M(h)$ depends on the unknown

* On leave from Australian National University. – Work of first author supported by U.S.A.F. Grant No. F 49620 82 C 009.

** Research of second author partially supported by NSF Grant DMS-8400602.

density f. Any attempt to estimate this "optimal" h must result in a window which is a function of the sample values. That is, the value of h must in practice be a random variable. Bearing this in mind, it seems to us that one should try from the outset to minimise integrated square error (ISE),

$$\Delta(h) \equiv \int \{f_n(x|h) - f(x)\}^2 \, dx,$$

instead of MSE. If $\hat{h}_0$ (a random variable) minimizes Δ, and h_0 (non-random) minimizes M, then $E\{\Delta(h_0)\} \geq E\{\Delta(\hat{h}_0)\}$. In this sense, $\hat{h}_0$ improves on h_0.

Let $\hat{h}$ be a "data-driven" bandwidth, estimated from the sample in some way. Our aim in this paper is to examine the distance between $\hat{h}$ and $\hat{h}_0$, and the distance between $\Delta(\hat{h})$ and $\Delta(\hat{h}_0)$. Of course, $\Delta(\hat{h}) \geq \Delta(\hat{h}_0)$. We ask: how much greater than the minimum, $\Delta(\hat{h}_0)$, is $\Delta(\hat{h})$?

There are at least two approaches to constructing $\hat{h}$: the classical argument, which essentially tries to estimate h_0; and least-squares cross-validation (Bowman [2, 3]; Rudemo [19]). The cross-validatory window is that value $\hat{h}_c$ which minimizes

$$CV(h) \equiv \int f_n^2(x|h) \, dx - 2n^{-1} \sum_{i=1}^{n} f_{ni}(X_i|h),$$

where $f_{ni}(x|h) \equiv \{(n-1)h\}^{-1} \sum_{j \neq i} K\{(x - X_j)/h\}$ is the kernel density estimate obtained by leaving out sample value X_i. The intuitive appeal of cross-validation is that it sidesteps secondary issues such as theoretical properties of MISE, and goes straight to the heart of the problem, by minimizing an estimate of $\Delta(h) - \int f^2$. (Notice that $CV(h)$ is unbiased for $M(h) - \int f^2$.) We shall show that this directness pays dividends. In a range of situations, including the multivariate case, the difference between $\Delta(\hat{h}_c)$ and $\Delta(\hat{h}_0)$ is of the same order of magnitude as the difference between $\Delta(h_0)$ and $\Delta(\hat{h}_0)$, under minimal smoothness conditions on f. (The common order is n^{-1}.) In this sense, the classical "best but unachievable strategy" of using h_0 is no better than the achievable strategy of least-squares cross-validation. Furthermore, neither h_0 not $\hat{h}_c$ consistently outperforms the other, since probabilities

$$P\{\Delta(\hat{h}_c) > \Delta(h_0)\}, \quad \mathrm{P}\{\Delta(\hat{h}_c) < \Delta(h_0)\}$$

both converge to strictly positive limits.

One class of competitors to $\hat{h}_c$ consists of two-stage ("plug-in") procedures, which aim to estimate the constant c_0 in the asymptotic formula $h_0 \sim c_0 n^{-1/5}$ (valid in one dimension). They cannot be expected to perform better than if the precise value of h_0 had been available. They can produce windows $\hat{h}$ for which $\Delta(\hat{h}) - \Delta(\hat{h}_0)$ is of larger order of magnitude than $\Delta(h_0) - \Delta(\hat{h}_0)$, depending on their construction and the extent of additional smoothness assumptions.

We shall close this section by relating our contributions to recent work in the area. Theorem 2.3 of Rice [16] is close to our Theorem 2.1, but in the context of nonparametric regression. Asymptotic first-order optimality of least-squares cross-validation in density estimation has been established by Hall [11, 13] and Stone [21]; Stone's work assumes minimal conditions on f. Other forms of cross-validation in nonparametric density estimation have been considered by Habbema, Hermans and van den Broek [9], Duin [8], Chow, Geman and Wu [5], Bowman, Hall and Titterington [4] and Marron [14, 15]. The last three papers take quite a general view of the principle of cross-validation. A recent survey by Titterington [22] sets cross-validation into context as a smoothing technique. First- and second-order properties of the difference between ISE and MISE have been examined by Bickel and Rosenblatt [1], Rosenblatt [18], Csörgő and Révész [6] (pp. 228–229) and Hall [10, 12]. Finally, we should point out that although L^2 measures of error, such as MISE, are very widely accepted, there do exist alternatives – examples include supremum measures (Silverman [20]) and L^1 measures (Devroye and Györfi [7]).

2. Results

For the sake of clarity and brevity we shall state and prove our main results for the case of one-dimensional data, in the context of a positive kernel. Towards the end of this section we shall show that the theorems are readily extendible to any finite number of dimensions, and to more general kernels which may become negative in order to reduce bias.

We impose the following conditions on K and f:

(2.1) K is a compactly supported, symmetric function on $\mathbb{R}$ with Hölder-continuous derivative K', and satisfies

$$\int K=1, \qquad \int z^2 K(z)\,dz \equiv 2k \neq 0.$$

(A function g is Hölder continuous if $|g(x)-g(y)| \leqq c|x-y|^{\varepsilon}$ for some $c, \varepsilon>0$ and all x, y.)

(2.2) f is bounded and twice differentiable, f' and f'' are bounded and integrable, and f'' is uniformly continuous.

Define integrated square error Δ, mean integrated square error $M \equiv E(\Delta)$, and the cross-validatory criterion CV as in Sect. 1. Set $D \equiv \Delta - M$, and notice that $CV = \Delta + \delta - \int f^2$, where

$$\tfrac{1}{2}\delta \equiv \int f f_n - n^{-1} \sum_{i=1}^{n} f_{ni}(X_i).$$

Recall that $\hat{h}_0$, $\hat{h}_c$ and h_0 minimize Δ, CV and M, respectively. Observe that

$$M(h)=(nh)^{-1}\int K^2+(1-n^{-1})\int\{\int K(z)f(x-hz)\,dz\}^2\,dx \\ -2\int f(x)\,dx\int K(z)f(x-hz)\,dz+\int f^2.$$

We may derive expressions for $M'(h)$ and $M''(h)$ by differentiating under the integral signs in this formula. In that way we may deduce that, with $c_1 \equiv \int K^2$ and $c_2 \equiv k^2 \int (f'')^2$ we have

$$M(h)=c_1(nh)^{-1}+c_2h^4+o\{(nh)^{-1}+h^4\},$$
$$M''(h)=2c_1(nh^3)^{-1}+12c_2h^2+o\{(nh^3)^{-1}+h^2\}$$

as $h \to 0$ and $n \to \infty$. Consequently, $h_0 \sim c_0 n^{-1/5}$ where $c_0 = (c_1/4c_2)^{1/5}$, and $M''(h_0) \sim c_3 n^{-2/5}$ where $c_3 = 2c_1 c_0^{-3} + 12 c_2 c_0^2$. Set

$$L(z) \equiv -zK'(z),$$
$$\sigma_0^2 \equiv (2/c_0)^3 (\int f^2) \int [\int K(y+z)\{K(z) - L(z)\}\, dz]^2\, dy + (4kc_0)^2 \{\int (f'')^2 f - (\int f'' f)^2\},$$
$$\sigma_c^2 \equiv (2/c_0)^3 (\int f^2)(\int L^2) + (4kc_0)^2 \{\int (f'')^2 f - (\int f'' f)^2\}.$$

The structure of our arguments is very simple, and so we shall prove our main results here. The lemmas in Sect. 3 supply all the rigour needed, and we shall refer to them as required.

First, we prove a limit theorem for $\hat{h}_0 - h_0$. Observe that

$$0 = \Delta'(\hat{h}_0) = M'(\hat{h}_0) + D'(\hat{h}_0) = (\hat{h}_0 - h_0)\, M''(h^*) + D'(\hat{h}_0), \tag{2.3}$$

where h^* lies inbetween h_0 and $\hat{h}_0$. By Lemma 3.3, $\hat{h}_0 = h_0 + O_p(n^{-1/5-\varepsilon})$ for some $\varepsilon > 0$, and so by Lemma 3.2 (with $h_1 = h_0$), $D'(\hat{h}_0) = D'(h_0) + o_p(n^{-7/10})$. But Lemma 3.4 declares that $n^{7/10} D'(h_0) \xrightarrow{\mathscr{D}} N(0, \sigma_0^2)$, and so $n^{7/10} D'(\hat{h}_0)$ must have the same weak limit. Since $h^*/h_0 \xrightarrow{p} 1$, it is easily shown that $M''(h^*) = c_3 n^{-2/5} + o_p(n^{-2/5})$. Combining the estimates from (2.3) down, we conclude that

$$n^{3/10}(\hat{h}_0 - h_0) \xrightarrow{\mathscr{D}} N(0, \sigma_0^2 c_3^{-2}). \tag{2.4}$$

Next we prove a limit theorem for $\hat{h}_c - \hat{h}_0$. Notice that

$$\begin{aligned} 0 = CV'(\hat{h}_c) &= M'(\hat{h}_c) + D'(\hat{h}_c) + \delta'(\hat{h}_c) \\ &= (\hat{h}_c - h_0)\, M''(h^*) + D'(\hat{h}_c) + \delta'(\hat{h}_c), \end{aligned} \tag{2.5}$$

where on this occasion h^* lies inbetween h_0 and $\hat{h}_c$.

Using Lemmas 3.2 and 3.3 in the same manner as before, we find that $D'(\hat{h}_c) + \delta'(\hat{h}_c) = D'(h_0) + \delta'(h_0) + o_p(n^{-7/10})$. Lemmas 3.4 and 3.5 imply that $D'(h_0) + \delta'(h_0) = O_p(n^{-7/10})$. Since $h^*/\hat{h}_0 \xrightarrow{p} 1$, it is easily shown that $M''(h^*) = c_3 n^{-2/5} + o_p(n^{-2/5})$. Using these results in (2.5), we find that

$$0 = (\hat{h}_c - h_0)\, c_3 n^{-2/5} \{1 + o_p(1)\} + O_p(n^{-7/10}),$$

and so $\hat{h}_c - h_0 = O_p(n^{-3/10})$. This means that

$$(\hat{h}_c - h_0)\, M''(h^*) = (\hat{h}_c - h_0)\, c_3 n^{-2/5} + o_p(n^{-7/10}),$$

and so we may refine (2.5) as follows:

$$0 = (\hat{h}_c - h_0)\, c_3 n^{-2/5} + D'(h_0) + \delta'(h_0) + o_p(n^{-7/10}).$$

We already know from the previous paragraph that

$$0 = (\hat{h}_0 - h_0)\, c_3 n^{-2/5} + D'(h_0) + o_p(n^{-7/10}).$$

Subtracting:

$$0 = (\hat{h}_c - \hat{h}_0)\, c_3 n^{-2/5} + \delta'(h_0) + o_p(n^{-7/10}).$$

This result and Lemma 3.5 entail

(2.6) $$n^{3/10}(\hat{h}_c - \hat{h}_0) \xrightarrow{\mathscr{D}} N(0, \sigma_c^2 c_3^{-2}).$$

We pause to combine (2.4) and (2.6) into a theorem.

Theorem 2.1. *Under conditions* (2.1) *and* (2.2),

$$n^{3/10}(\hat{h}_0 - h_0) \xrightarrow{\mathscr{D}} N(0, \sigma_0^2 c_3^{-2}) \quad \text{and} \quad n^{3/10}(\hat{h}_c - \hat{h}_0) \xrightarrow{\mathscr{D}} N(0, \sigma_c^2 c_3^{-2}).$$

Having derived these formula, it is only a short step to describe the amount by which h_0 and $\hat{h}_c$ fail to minimize integrated square error. For that purpose we impose an additional condition on K:

(2.7) K has a second derivative on $\mathbb{R}$, and K'' is Hölder continuous.

Let h denote either h_0 or $\hat{h}_c$, and notice that

$$\Delta(h) - \Delta(\hat{h}_0) = \tfrac{1}{2}(h - \hat{h}_0)^2 \Delta''(h^*),$$

where h^* lies inbetween h and $\hat{h}_0$. In view of Lemma 3.6 and the fact that $h^*/h_0 \xrightarrow{p} 1$, $\Delta''(h^*) = M''(h^*) + o_p(n^{-2/5})$. But $M''(h^*) = c_3 n^{-2/5} + o_p(n^{-2/5})$, and so, since $h - \hat{h}_0 = O_p(n^{-3/10})$,

$$\Delta(h) - \Delta(\hat{h}_0) = \tfrac{1}{2}(h - \hat{h}_0)^2 c_3 n^{-2/5} + o_p(n^{-1}).$$

Our next result is now immediate from Theorem 2.1.

Theorem 2.2. *Under conditions* (2.1), (2.2) *and* (2.7),

$$n\{\Delta(h_0) - \Delta(\hat{h}_0)\} \xrightarrow{\mathscr{D}} \tfrac{1}{2}\sigma_0^2 c_3^{-1} \chi_1^2 \quad \text{and} \quad n\{\Delta(\hat{h}_c) - \Delta(\hat{h}_0)\} \xrightarrow{\mathscr{D}} \tfrac{1}{2}\sigma_c^2 c_3^{-1} \chi_1^2.$$

Remarks

2.1. Higher Dimensions. The proofs of Theorems 2.1 and 2.2 work for higher dimensional data, although with more elaborate notation. In the case of p dimensions we should define L by

$$L(z) \equiv -p^{-1} \sum_{i=1}^{p} z^{(i)} K_i(z),$$

where $z = (z^{(1)}, \ldots, z^{(p)})$ and $K_i(z) = (\partial/\partial z^{(i)}) K(z)$. We assume p-dimensional versions of (2.1), (2.2) and (2.7), and define $2k \equiv \int z^{(i)2} K(z)\, dz$ (not depending on i), $c_1 \equiv \int K^2$, $c_2 \equiv k^2 \int (\nabla^2 f)^2$, $c_0 \equiv (p c_1 / 4 c_2)^{1/(p+4)}$, $c_3 \equiv p(p+1) c_1 c_0^{-(p+2)} + 12 c_2 c_0^2$,

$$\sigma_0^2 \equiv 8 p^2 c_0^{-p-2} (\textstyle\int f^2) \int [\int K(y+z)\{K(z) - L(z)\}\, dz]^2\, dy$$
$$+ (4 k c_0)^2 \{\textstyle\int f (\nabla^2 f)^2 - (\int f \nabla^2 f)^2\},$$
$$\sigma_c^2 \equiv 8 p^2 c_2^{-p-2} (\textstyle\int f^2)(\int L^2) + (4 k c_0)^2 \{\int f (\nabla^2 f)^2 - (\int f \nabla^2 f)^2\}.$$

Theorem 2.2 holds as before, and the only change to Theorem 2.1 is that the factor $n^{3/10}$ should be replaced by $n^{(p+2)/\{2(p+4)\}}$.

2.2. General Kernels. The forms of Theorems 2.1 and 2.2 remain unchanged if we admit more general kernels. To illustrate this, we shall confine attention to the case $p=1$. Higher dimensions may be treated similarly.

If K is chosen so that $\int K=1$ and for some integer $t \geqq 2$,

$$\int z^j K(z)\,dz=0 \quad \text{for } 1 \leqq j \leqq t-1, \qquad \int z^t K(z)\,dz \neq 0,$$

then the kernel L also enjoys these properties. A version of Theorem 2.1 holds in which $n^{3/10}$ is replaced by $n^{3/\{2(2t+1)\}}$, and Theorem 2.2 holds as before.

2.3. Joint Limit Theorems. Both Theorems 2.1 and 2.2 can be expressed in the form of joint limit theorems for

$$(n^{3/10}(\hat{h}_0-h_0),\, n^{3/10}(\hat{h}_c-\hat{h}_0))$$

and

$$(n\{\Delta(h_0)-\Delta(\hat{h}_0)\},\, n\{\Delta(\hat{h}_c)-\Delta(\hat{h}_0)\})$$

respectively. In the former case the limit is bivariate normal with zero means, variances as in Theorem 2.1, and covariance

$$-8c_3^{-2}[c_0^{-3}(\textstyle\int f^2)\int(K-L)\{K*(K-L)\}+2k^2c_0^2\{\int(f'')^2f-(\int f''f)^2\}], \tag{2.8}$$

where $*$ denotes convolution and k is as in (2.1). The bivariate limit is more complex in the case of Theorem 2.2.

Many commonly-used kernels K have Fourier transforms which are nonnegative. In this circumstance it follows from (2.8) via Parseval's inequality (for the first term within square brackets) and the Cauchy-Schwarz inequality (for the second term) that the asymptotic correlation between $\hat{h}_0-h_0$ and $\hat{h}_c-\hat{h}_0$ is negative. This means that h_0 and $\hat{h}_c$ tend to error *on the same side* of the optimal window, $\hat{h}_0$.

2.4. Comparing Different Bandwidths. For the sake of simplicity we shall confine attention to the case of positive kernels and one-dimensional data. We shall adhere to our convention, discussed in Sect. 1, that "better" windows h are those which give smaller integrated square error.

Let $\hat{h}$ be a window satisfying $\hat{h}/h_0 \xrightarrow{p} 1$. Assume conditions (2.1), (2.2) and (2.7). Using the argument leading to Theorem 2.2, we obtain:

$$0 \leqq \Delta(\hat{h})-\Delta(\hat{h}_0)=\tfrac{1}{2}(\hat{h}-\hat{h}_0)^2\, c_3 n^{-2/5}\{1+o_p(1)\}. \tag{2.9}$$

We shall consider various possibilities for $\hat{h}$.

(i) We might explicitly estimate the constant c_0 in the asymptotic formula $h_0 \sim c_0 n^{-1/5}$, and take $\hat{h}$ to be the resulting window. This requires estimation of $\int(f'')^2$, perhaps by integrating the square of a kernel estimate of f''. Such an approach is really a global version of Woodroofe's [23] two-stage procedure.

Under the smoothness assumption (2.2), the rate of convergence of such an estimator can be slower than $n^{-\varepsilon}$ for any given $\varepsilon > 0$. In consequence, the error $(\hat{h}-\hat{h}_0)^2$ may converge to zero in probability at a rate slower than $n^{-2/5-2\varepsilon}$, and by (2.9), $\Delta(\hat{h})-\Delta(\hat{h}_0)$ may be no smaller than order $n^{-4/5-2\varepsilon}$. On the other hand, if $\hat{h}$ is the cross-validatory window $\hat{h}_c$ then $\Delta(\hat{h})-\Delta(\hat{h}_0)$ is as small as n^{-1} under the minimal condition (2.2).

(ii) The procedure outlined in (i) is motivated by a desire to estimate h_0. Following that philosophy, we would be doing extremely well if we actually knew the value of h_0. But according to Theorem 2.2, even if we took $\hat{h}=h_0$ we would hardly do any better than using the cross-validatory window $\hat{h}_c$, since in both cases the distance of integrated square error from the minimum would be order n^{-1}.

(iii) If K is a positive kernel then $\sigma_0^2 \leqq \sigma_c^2$. This follows from the Cauchy-Schwarz inequality, $(\int gh)^2 \leqq (\int g^2)(\int h^2)$, applied with $g(z) \equiv \{K(y+z)\}^{\frac{1}{2}}$ and $h(z) \equiv \{K(y+z)\}^{\frac{1}{2}}\{K(z)-L(z)\}$. Notice that $\int L^2 = \int (W-L)^2$. (The result is true in any dimension.) In this sense, taking $\hat{h}=h_0$ does result in a marginal improvement over cross-validation. However, the improvement is not available with probability one. It is noted in Remark 2.3 that

$$(n^{3/10}(h_0-\hat{h}_0), n^{3/10}(\hat{h}_c-\hat{h}_0)) \xrightarrow{\mathscr{D}} (Z_1, Z_2)$$

say, where (Z_1, Z_2) has a joint normal distribution with $P(|Z_1|>|Z_2|)>0$. Consequently, the limit

$$\lim_{n\to\infty} P\{\Delta(h_0) > \Delta(\hat{h}_c)\}$$

exists, and is strictly positive.

3. Lemmas

The lemmas below were required for the proofs of Theorems 2.1 and 2.2. In Lemmas 3.1–3.5, we assume conditions (2.1) and (2.2). The symbols C, C_1 and C_2 denote generic positive constants.

Lemma 3.1. *For each* $0<a<b<\infty$ *and all positive integers* l,

$$\sup_{n;\, a\leqq t\leqq b} E|n^{7/10} D'(n^{-1/5}t)|^{2l} \leqq C_1(a,b,l), \tag{3.1}$$

$$\sup_{n;\, a\leqq t\leqq b} E|n^{7/10} \delta'(n^{-1/5}t)|^{2l} \leqq C_1(a,b,l). \tag{3.2}$$

Furthermore, there exists $\varepsilon_1>0$ *such that*

$$E|n^{7/10}\{D'(n^{-1/5}s)-D'(n^{-1/5}t)\}|^{2l} \leqq C_2(a,b,l)|s-t|^{\varepsilon_1 l}, \tag{3.3}$$

$$E|n^{7/10}\{\delta'(n^{-1/5}s)-\delta'(n^{-1/5}t)\}|^{2l} \leqq C_2(a,b,l)|s-t|^{\varepsilon_1 l} \tag{3.4}$$

whenever $a\leqq s\leqq t\leqq b$.

Proof. We begin by decomposing D' and δ'. Let $g_n(x|h) \equiv (nh)^{-1} \sum_i L\{(x-X_i)/h\}$, and observe that

$$\begin{aligned}-(h/2)\,\Delta'(h) &= \int (f_n - f)(f_n - g_n) \\ &= \int (f_n - Ef_n)^2 - \int (f_n - Ef_n)(g_n - Eg_n) \\ &\quad + \int (f_n - Ef_n)(2Ef_n - Eg_n - f) + \int (g_n - Eg_n)(f - Ef_n) \\ &\quad + \int (Ef_n - f)(Ef_n - Eg_n).\end{aligned}$$

Put $K_i(x) \equiv K\{(x-X_i)/h\} - EK\{(x-X_i)/h\}$, and define L_i similarly. By expanding $\int (f_n - Ef_n)^2$ as a sum of integrals of squares plus a sum of integrals of products, and expanding $\int (f_n - Ef_n)(g_n - Eg_n)$ in a similar way, we conclude that

$$D_1(h) \equiv -(h/2)\, D'(h) = S_1(h) + S_2(h) + S_3(h), \tag{3.5}$$

where $S_1 \equiv S_{11} - S_{12}$, $S_2 \equiv S_{21} + S_{22}$, $S_3 \equiv S_{31} - S_{32}$,

$$S_{11} \equiv 2(nh)^{-2} \sum\sum_{1 \leqq i < j \leqq n} \int K_i(x)\, K_j(x)\, dx,$$

$$S_{12} \equiv (nh)^{-2} \sum\sum_{1 \leqq i < j \leqq n} \int \{K_i(x)\, L_j(x) + L_i(x)\, K_j(x)\}\, dx,$$

$$S_{21} \equiv (nh)^{-1} \sum_{i=1}^{n} \int K_i(x) \{2Ef_n(x|h) - Eg_n(x|h) - f(x)\}\, dx,$$

$$S_{22} \equiv (nh)^{-1} \sum_{i=1}^{n} \int L_i(x) \{f(x) - Ef_n(x|h)\}\, dx,$$

$$S_{31} \equiv (nh)^{-2} \sum_{i=1}^{n} \int \{K_i(x)^2 - EK_i(x)^2\}\, dx,$$

$$S_{32} \equiv (nh)^{-2} \sum_{i=1}^{n} \int \{K_i(x)\, L_i(x) - EK_i(x)\, L_i(x)\}\, dx.$$

A similar argument produces the decomposition

$$\delta_1(h) \equiv (h/2)\, \delta'(h) = T_1(h) + T_2(h), \tag{3.6}$$

where $T_1 \equiv T_{11} - T_{12}$, $T_2 \equiv T_{21} - T_{22}$,

$$T_{1l} \equiv 2\{n(n-1)\,h\}^{-1} \sum\sum_{1 \leqq i < j \leqq n} \{B_l(X_i, X_j) - b_l(X_i) - b_l(X_j) + \mu_l\},$$

$$T_{2l} \equiv (nh)^{-1} \sum_{i=1}^{n} \{b_l(X_i) - \mu_l - f(X_i) + \int f^2\},$$

$B_1(x, y) = K\{(x-y)/h\}$, $B_2(x, y) = L\{(x-y)/h\}$, $b_l(x) = E\{B_l(x, X)\}$, $\mu_l = E\{b_l(X)\}$.

To prove (3.3) we shall show that for some $\varepsilon > 0$,

$$E|n^{9/10}\{S_{11}(n^{-1/5}s) - S_{11}(n^{-1/5}t)\}|^{2l} \leqq C|s-t|^{\varepsilon l}, \tag{3.7}$$

(3.8) $$E|n^{9/10}\{S_{21}(n^{-1/5}s)-S_{21}(n^{-1/5}t)\}|^{2l}\leqq C|s-t|^{\varepsilon l},$$

(3.9) $$E|n^{13/10}\{S_{31}(n^{-1/5}s)-S_{31}(n^{-1/5}t)\}|^{2l}\leqq C|s-t|^{\varepsilon l}.$$

Similar inequalities may be established for the functions S_{12}, S_{22} and S_{32}.

To verify (3.7), note that S_{11} may be written as

$$S_{11}(n^{-1/5}t)=n^{-2}\sum\sum_{1\leqq i<j\leqq n} U_t(i,j)$$

and $U_t(i,j)$ satisfies

(3.10) $$E[U_t(i,j)|X_i]=E[U_t(i,j)|X_j]=0.$$

Be the compactness of support (which without loss of generality may be taken to be $[-1,1]$) and the Hölder continuity of K, for $s,t\in(a,b)$,

$$\begin{aligned}\Big|(n^{-1/5}s)^{-2}\int K\left(\frac{x-X_i}{n^{-1/5}s}\right)K\left(\frac{x-X_j}{n^{-1/5}s}\right)dx\\ -(n^{-1/5}t)^{-2}\int K\left(\frac{x-X_i}{n^{-1/5}t}\right)K\left(\frac{x-X_j}{n^{-1/5}t}\right)dx\Big|\\ \leqq C|s-t|^{\varepsilon}(n^{-1/5}b)^{-1}1_{[-2,2]}\left(\frac{X_i-X_j}{n^{-1/5}b}\right).\end{aligned}$$

Hence,

(3.11) $$|U_s(i,j)-U_t(i,j)|\leqq C|s-t|^{\varepsilon}\left\{(n^{-1/5}b)^{-1}1_{[-2,2]}\left(\frac{X_i-X_j}{n^{-1/5}b}\right)+12\right\}.$$

But,

(3.12) $$\begin{aligned}&E[n^{9/10}\{S_{11}(n^{-1/5}s)-S_{11}(n^{-1/5}t)\}]^{2l}\\ &=n^{-11l/5}\sum_{i_1<j_1}\sum\cdots\sum_{i_{2l}<j_{2l}}\sum E[\{U_s(i_1,j_1)-U_t(i_1,j_1)\}\\ &\quad\cdots\{U_s(i_{2l},j_{2l})-U_t(i_{2l},j_{2l})\}].\end{aligned}$$

Rearrange the terms on the right side of (3.12) into $4l$ groups where the term indexed by $i_1,j_1,\ldots,i_{2l},j_{2l}$ is put in the m-th group when there are exactly m distinct integers in the list $i_1,j_1,\ldots,i_{2l},j_{2l}$. Note that the cardinality of the m-th group is bounded by Cn^m, and by (3.10), each term is 0 in the groups $2l+1,\ldots,4l$. Hence, by (3.11) and integration by substitution,

$$\begin{aligned}&E[n^{9/10}\{S_{11}(n^{-1/5}s)-S_{11}(n^{-1/5}t)\}]^{2l}\\ &\leqq C_1n^{-11l/5}\sum_{m=2}^{2l}n^m|s-t|^{2\varepsilon l}n^{2l/5-m/10}\\ &\leqq C_2|s-t|^{2\varepsilon l},\end{aligned}$$

and the proof of (3.7) is complete.

To verify (3.8), note that by Taylor's theorem, (2.1), (2.2) and the fact that L is also symmetric and integrates to 1, for $t\in(a,b)$

$$|2Ef_n(x|n^{-1/5}t)-Eg_n(x|n^{-1/5}t)-f(x)|\leqq Cn^{-2/5}.$$

Hence S_{21} may be written

$$S_{21}(n^{-1/5}t)=n^{-1}\sum_{i=1}^{n} V_t(i),$$

where

$$E[V_t(i)]=0,$$

and where

$$|V_s(i)-V_t(i)|\leqq Cn^{-2/5}|s-t|^{\varepsilon}. \tag{3.13}$$

By a cumulant expansion of the $2l$-th moment, to show (3.8) it is enough to check that for $m=2, \ldots, 2l$,

$$|cum_m(n^{9/10}\{S_{21}(n^{-1/5}s)-S_{21}(n^{-1/5}t)\})|\leqq C|s-t|^{\varepsilon m},$$

where $cum_m(\cdot)$ denotes the m-th order cumulant. But, by the independence property of cumulants,

$$\begin{aligned}
&|cum_m(n^{9/10}\{S_{21}(n^{-1/5}s)-S_{21}(n^{-1/5}t)\})|\\
&=\left|n^{-m/10}\sum_{i=1}^{n} cum_m(V_s(i)-V_t(i))\right|\\
&\leqq Cn^{1-m/10}[n^{-2/5}|s-t|^{\varepsilon}]^m\\
&=Cn^{1-m/2}|s-t|^{\varepsilon m},
\end{aligned}$$

where the inequality follows from (3.13). This completes the proof of (3.8).

The verification of (3.9) is quite similar to that of (3.8) so only differences will be noted. Write

$$S_{31}(n^{-1/5}t)=n^{-2}\sum_{i=1}^{n} W_t(i),$$

where

$$E[W_t(i)]=0,$$

and where

$$|W_s(i)-W_t(i)|\leqq Cn^{1/5}|s-t|^{\varepsilon}.$$

Hence, for $m=2, \ldots, 2l$,

$$|cum_m(n^{9/10}\{S_{21}(n^{-1/5}s)-S_{21}(n^{-1/5}t)\})| \leqq Cn^{-9m/10+1+m/5}|s-t|^{\varepsilon m}.$$

This completes the proof of (3.9) and hence that of (3.3). The same type of argument may be used to prove (3.1), (3.2) and (3.4).

Lemma 3.2. *For some* $\varepsilon>0$ *and any* $0<a<b<\infty$,

$$\sup_{a\leqq t\leqq b}\{|D'(n^{-1/5}t)|+|\delta'(n^{-1/5}t)|\}=O_p(n^{-3/5-\varepsilon}). \tag{3.14}$$

Furthermore, for any $\varepsilon_2>0$ *and any non-random* h_1 *asymptotic to a constant multiple of* $n^{-1/5}$,

$$\sup_{|t-n^{1/5}h_1|\leqq n^{-\varepsilon_2}} n^{7/10}\{|D'(n^{-1/5}t)-D'(h_1)|+|\delta'(n^{-1/5}t)-\delta'(h_1)|\}\xrightarrow{p}0. \tag{3.15}$$

Proof. We give a proof only for D'. The proof for δ' is similar. To check (3.15), note that using the decomposition (3.5) of D', the Hölder continuity of K and L, and the fact that both of these functions have compact support, there is an $\alpha>0$ sufficiently large that

$$\sup_{\substack{a\leqq s\leqq t\leqq 2b \\ |s-t|\leqq n^{-\alpha+1/5}}} |D'(n^{-1/5}s)-D'(n^{-1/5}t)|=O(n^{-1}). \tag{3.16}$$

For $a<\lim n^{1/5}h_1<b$, suppose

$$n^{1/5}h_1-n^{-\varepsilon_2}=t_0<t_1<\ldots<t_{m-1}\leqq n^{1/5}h_1+n^{-\varepsilon_2}<t_m,$$

where $t_i-t_{i-1}=n^{-\alpha}$ for each i. In view of (3.16), to finish the proof of (3.15) it suffices to check that

$$\sup_{(t_i,t_j)\in\mathscr{T}} n^{7/10}|D'(n^{-1/5}t_i)-D'(n^{-1/5}t_j)| \xrightarrow{p} 0,$$

where $\mathscr{T}$ is the set of all pairs (t_i, t_j) with $0<t_i-t_j\leqq n^{-1/5-\varepsilon_2}$ and $i\leqq m$. For any $\eta>0$,

$$\begin{aligned} &P\{\sup_{(t_i,t_j)\in\mathscr{T}} n^{7/10}|D'(n^{-1/5}t_i)-D'(n^{-1/5}t_j)|>\eta\} \\ &\leqq \sum_{(t_i,t_j)\in\mathscr{T}} E\{\eta^{-1}n^{7/10}|D'(n^{-1/5}t_i)-D'(n^{-1/5}t_j)|\}^{2l} \\ &\leqq C\eta^{-2l}n^{2(\alpha-\varepsilon_2-1/5)}(n^{-1/5-\varepsilon_2})^{\varepsilon_1 l}, \end{aligned} \tag{3.17}$$

using (3.3) and the fact that the number of elements in $\mathscr{T}$ is of order $n^{2(\alpha-\varepsilon_2-1/5)}$. By choosing l sufficiently large we may ensure that the term in (3.17) converges to zero as $n\to\infty$. This proves (3.15). A similar partitioning argument may be used to prove (3.14).

Lemma 3.3. *For some* $\varepsilon>0$,

$$|\hat{h}_0-h_0|+|\hat{h}_c-h_0|=O_p(n^{-1/5-\varepsilon}).$$

Proof. First we treat $|\hat{h}_0$-$h_0|$. It is not difficult to prove, using techniques of Hall [11] (p. 1160), that $\hat{h}_0/h_0 \xrightarrow{p} 1$. Therefore by Lemma 3.2,

$$\Delta'(h_0)=\Delta'_0(h_0)-\Delta'(\hat{h}_0)=M'(h_0)-M'(\hat{h}_0)+O_p(n^{-3/5-\varepsilon}).$$

Also by Lemma 3.2, $\Delta'(h_0)=D'(h_0)=O_p(n^{-3/5-\varepsilon})$, and so

$$O_p(n^{-3/5-\varepsilon})=M'(h_0)-M'(\hat{h}_0)=(h_0-\hat{h}_0)M''(h^*), \tag{3.18}$$

where h^* lies inbetween h_0 and $\hat{h}_0$. As in Sect. 2, $M''(h^*)=c_3n^{-2/5}+o_p(n^{-2/5})$. Using this estimate in (3.18) we conclude that $h_0-\hat{h}_0=O_p(n^{-1/5-\varepsilon})$, as required.

To treat $|\hat{h}_c$-$h_0|$, notice that $\hat{h}_c/h_0 \xrightarrow{p} 1$. Therefore

$$\begin{aligned} CV'(h_0)&=CV'(h_0)-CV'(\hat{h}_c)=\Delta'(h_0)-\Delta'(\hat{h}_c)+\delta'(h_0)-\delta'(\hat{h}_c) \\ &=M'(h_0)-M'(\hat{h}_c)+O_p(n^{-3/5-\varepsilon}), \end{aligned}$$

again using Lemma 3.2. But $CV'(h_0)=M'(h_0)+O_p(n^{-3/5-\varepsilon})$, and so as before it follows that $h_0-\hat{h}_c=O_p(n^{-1/5-\varepsilon})$.

Lemma 3.4. $n^{7/10}D'(h_0)\xrightarrow{\mathscr{D}}N(0,\sigma_0^2)$.

Proof. We shall start from decomposition (3.5), and prove that $n^{9/10}D_1(h_0)\xrightarrow{\mathscr{D}}N(0,c_0^2\sigma_0^2/4)$. Now, the argument leading to (3.9) gives $E\{S_3^2(h_0)\}=O(n^{-13/5})$, and so $S_3(h_0)=o_p(n^{-9/10})$. Therefore by (3.5), it suffices to show that

$$(n^{9/10}S_1, n^{9/10}S_2)\xrightarrow{\mathscr{D}}(Z_1,Z_2), \tag{3.19}$$

where $S_i=S_i(h_0)$ and Z_1 and Z_2 are independent normal variables with zero means and variances adding up to $c_0^2\sigma_0^2/4$.

Our route to (3.19) uses the argument of Hall [12], and so we omit many details. The variables S_1 and S_2 are uncorrelated.

If we write $S_1=\sum\sum_{j<i}A(X_i,X_j)$ and $S_2=\sum_i a(X_i)$, then $E\{A(X_i,X_j)|X_j\}=0$ almost surely for each $j<i$, and so for any real c and d, the variables

$$Y_i\equiv c\sum_{j=1}^{i-1}A(X_i,X_j)+d\,a(X_i),\qquad 1\leqq i\leqq n,$$

are zero-mean martingale differences with respect to the σ-fields $\mathscr{F}\{X_1,\ldots,X_i\}$. In this sense, $cS_1+dS_2\equiv\sum_{i=1}^{n}Y_i$ is a martingale. The argument leading to Hall's [12] Theorem 1 shows that cS_1+dS_2 is asymptotically normally distributed with variance $c^2\operatorname{var}(S_1)+d^2\operatorname{var}(S_2)$. This property, together with the Cramér-Wold device, permits us to complete the proof of (3.19) by showing that

$$n^{9/5}\operatorname{var}(S_1)\to 2c_0^{-1}(\textstyle\int f^2)\int[\int K(y+z)\{K(z)-L(z)\}\,dz]^2\,dy, \tag{3.20}$$

$$n^{9/5}\operatorname{var}(S_2)\to 4k^2c_0^4\{\textstyle\int(f'')^2f-(\int f''f)^2\}. \tag{3.21}$$

Let

$$\gamma_1(x,y)=E\left[\left\{K\left(\frac{x-X}{h_0}\right)-EK\left(\frac{x-X}{h_0}\right)\right\}\left\{K\left(\frac{y-X}{h_0}\right)-EK\left(\frac{y-X}{h_0}\right)\right\}\right],$$

$$\gamma_2(x,y)=E\left[\left\{L\left(\frac{x-X}{h_0}\right)-EL\left(\frac{x-X}{h_0}\right)\right\}\left\{L\left(\frac{y-X}{h_0}\right)-EL\left(\frac{y-X}{h_0}\right)\right\}\right],$$

$$\gamma_3(x,y)=E\left[\left\{K\left(\frac{x-X}{h_0}\right)-EK\left(\frac{x-X}{h_0}\right)\right\}\left\{L\left(\frac{y-X}{h_0}\right)-EL\left(\frac{y-X}{h_0}\right)\right\}\right],$$

and $\gamma_4(x,y)=\gamma_3(y,x)$. Then

$$\operatorname{var}(S_1)=(nh_0)^{-4}n(n-1)\iint(2\gamma_1^2-2\gamma_1\gamma_3-2\gamma_1\gamma_4+\gamma_1\gamma_2+\gamma_3\gamma_4).$$

The functions γ_i are covariances, and each may be expressed in the form $E(UV)-E(U)E(V)$ for variables U and V. A little algebra shows that the term $E(U)E(V)$ makes a negligible contribution, and in fact

$$\text{(3.22)} \qquad \operatorname{var}(S_1) = n^{-2} h_0^{-1} (\textstyle\int f^2) \int (2\beta_1^2 - 2\beta_1\beta_3 - 2\beta_1\beta_4 + \beta_1\beta_2 + \beta_3\beta_4) + o(n^{-2} h_0^{-1}),$$

where

$$\beta_1(y) = \int K(z) K(y+z)\, dz, \qquad \beta_2(y) = \int L(z) L(y+z)\, dz,$$
$$\beta_3(y) = \int K(z) L(y+z)\, dz, \qquad \beta_4(y) = \int L(z) K(y+z)\, dz = \beta_3(y).$$

Since $\int \beta_1 \beta_2 = \int \beta_3^2$ then

$$\int (2\beta_1^2 - 2\beta_1\beta_3 - 2\beta_1\beta_4 + \beta_1\beta_2 + \beta_3\beta_4) = 2 \int (\beta_1 - \beta_3)^2,$$

and so (3.20) is immediate from (3.22).

To prove (3.21), observe that

$$\text{(3.23)} \qquad \operatorname{var}(S_2) = (n h_0)^{-2} n (v_2 - v_1^2),$$

where

$$v_i = E\left(\left|\int \left[K\left(\frac{x-X}{h}\right) \{2 E f_n(x|h) - E g_n(x|h) - f(x)\} + L\left(\frac{x-X}{h}\right) \{f(x) - E f_n(x|h)\}\right] dx\right|^i\right).$$

As $h \to 0$,

$$E\{f_n(x|h)\} - f(x) = k h^2 f''(x) + o(h^2),$$
$$E\{g_n(x|h)\} - f(x) = 3 k h^2 f''(x) + o(h^2).$$

Estimates of this type give:

$$v_1 = -k h^2 \int\int \left\{K\left(\frac{x-y}{h}\right) + L\left(\frac{x-y}{h}\right)\right\} f''(x) f(y)\, dy\, dx + o(h^3)$$
$$= -2 k h^3 \int f'' f + o(h^3),$$
$$v_2 = (k h^2)^2 \int \left[\int \left\{K\left(\frac{x-y}{h}\right) + L\left(\frac{x-y}{h}\right)\right\} f''(x)\, dx\right]^2 f(y)\, dy + o(h^6)$$
$$= 4 k^2 h^6 \int (f'')^2 f + o(h^6).$$

Result (3.21) now follows from (3.23).

Lemma 3.5. $n^{7/10} \delta'(h_0) \xrightarrow{\mathscr{D}} N(0, \sigma_c^2)$.

Proof. The martingale methods and Cramér-Wold device used to prove Lemma 3.4, are also applicable here. The argument is based on (3.6) instead of (3.5). We shall prove only the analogue of (3.20):

$$\text{(3.24)} \qquad n^{9/5} \operatorname{var}(T_1) \to 2 c_0^{-1} (\textstyle\int f^2) \int (K - L)^2.$$

The analogue of (3.21), which declares that $n^{9/5} \operatorname{var}(T_2)$ converges to the same limit as in (3.21), follows as before.

To prove (3.24), notice that with $B=B_1-B_2$, $b=b_1-b_2$ and $\mu=\mu_1-\mu_2$,

$$\begin{aligned}\operatorname{var}(T_1)&=2\{n(n-1)\,h_0\}^{-1}\,n(n-1)\,E\{B(X_1,X_2)-b(X_1)-b(X_2)+\mu\}^2\\&=2\{n(n-1)\,h_0^2\}^{-1}\,E\{B^2(X_1,X_2)-2\,b^2(X_1)+\mu^2\}\\&\sim 2\,n^{-2}\,h_0^{-2}\,E\{B^2(X_1,X_2)\}\\&=2\,n^{-2}\,h_0^{-2}\int\!\!\int\left\{K\left(\frac{x-y}{h_0}\right)-L\left(\frac{x-y}{h_0}\right)\right\}^2 f(x)f(y)\,dx\,dy\\&\sim 2\,n^{-2}\,h_0^{-1}(\textstyle\int f^2)\int(K-L)^2.\end{aligned}$$

Lemma 3.6. *Under conditions* (2.1), (2.2) *and* (2.7), *and for any* $0<a<b<\infty$,

$$\sup_{a\leqq t\leqq b}|D''(n^{-1/5}t)|=o_p(n^{-2/5}). \tag{3.25}$$

Proof. First derive an analogue of (3.1), using an almost identical argument:

$$\sup_{n;\,a\leqq t\leqq b} E|n^{1/2}\,D''(n^{-1/5}t)|^{2l}\leqq C(a,b,l).$$

Then follow the proof of Lemma 3.2, to conclude that (3.25) holds, and in fact the right-hand side equals $O_p(n^{-2/5-\varepsilon})$ for some $\varepsilon>0$.

References

1. Bickel, P.J., Rosenblatt, M.: On some global measures of the deviations of density estimates. Ann. Stat. **1**, 1071–1095 (1973)
2. Bowman, A.W.: A comparative study of some kernel-based nonparametric density estimates. Manchester-Sheffield School of Probability and Statistics Research Report 84/AWB/1 (1982)
3. Bowman, A.W.: An alternative method of cross-validation for the smoothing of density estimates. Biometrika **71**, 353–360 (1984)
4. Bowman, A.W., Hall, P., Titterington, D.M.: Cross-validation in nonparametric estimation of probabilities and probability densities. Biometrika **71**, 341–352 (1984)
5. Chow, Y.S., Geman, S., Wu, L.D.: Consistent cross-validated density estimation. Ann. Stat. **11**, 25–38 (1983)
6. Csörgő, M., Révész, P.: Strong approximations in probability and statistics. New York: Academic Press 1981
7. Devroye, L., Györfi, L.: Nonparametric density estimation: the L_1 view. New York: Wiley 1985
8. Duin, R.P.W.: On the choice of smoothing parameters for Parzen estimators of probability density functions. IEEE Trans. Comput. **C-25**, 1175–1179 (1976)
9. Habbema, J.D.F., Hermans, J., van den Broek, K.: A stepwise discriminant analysis program using density estimation. Compstat 1974, ed.: G. Bruckman 101–110. Vienna: Physica 1974
10. Hall, P.: Limit theorems for stochastic measures of the accuracy of density estimators. Stochastic Processes Appl. **13**, 11–25 (1982)
11. Hall, P.: Large sample optimality of least squares cross-validation in density estimation. Ann. Stat. **11**, 1156–1174 (1983)
12. Hall, P.: Central limit theorem for integrated square error of multivariate nonparametric density estimators. J. Multivariate Anal. **14**, 1–16 (1984)
13. Hall, P.: Asymptotic theory of minimum integrated square error for multivariate density estimation. In: Krishnaiah, P.R. (ed.) Proc. Sixth Internat. Sympos. Multivariate Analysis, pp. 289–309. Amsterdam: North Holland 1985

14. Marron, J.S.: An asymptotically efficient solution to the bandwidth problem of kernel density estimation. Ann. Stat. **13**, 1011–1023 (1985).
15. Marron, J.S.: A comparison of cross-validation techniques in density estimation. N.C. Inst. of Statist. Mimeo Series #1568 (1985)
16. Rice, J.: Bandwidth choice for nonparametric regression. Ann. Stat. **12**, 1215–1230 (1984)
17. Rosenblatt, M.: Curve estimates. Ann. Math. Stat. **42**, 1815–1842 (1971)
18. Rosenblatt, M.: A quadratic measure of deviation of two-dimensional density estimates and a test of independence. Ann. Stat. **3**, 1–14 (1975)
19. Rudemo, M.: Empirical choice of histogram and kernel density estimators. Scand. J. Stat. **9**, 65–78 (1982)
20. Silverman, B.W.: Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. Ann. Stat. **6**, 177–184 (1978)
21. Stone, C.J.: An asymptotically optimal window selection rule for kernel density estimates. Ann. Stat. **12**, 1285–1297 (1984)
22. Titterington, D.M.: Common structure of smoothing techniques in statistics. Int. Stat. Rev. **53**, 141–170 (1985)
23. Woodroofe, M.: On choosing a delta sequence. Ann. Math. Stat. **41**, 1665–1671 (1970)

Received March 16, 1985; in final form July 15, 1986