#2024

A simple root n bandwidth selector

by

M. C. Jones, J. S. Marron, B. U. Park

The Open University,
University of North Carolina,
Seoul National University

April 16, 1990

ABSTRACT

The asymptotically best bandwidth selectors for a kernel density estimator currently require the use of either unappealing higher order kernel pilot estimators or related Fourier transform methods. The point of this paper is to present a methodology which allows the fastest possible rate of convergence with the use of only nonnegative kernel estimators at all stages of the selection process. The essential idea is derived through careful study of factorizations of the pilot bandwidth in terms of the original bandwidth.

Research partially supported by the National Science Foundation Grant DMS–8902973. M. C. Jones was with the IBM Research Division during the course of this work. B. U. Park's research was also supported by the Korean Science and Engineering Foundation 89–90.

AMS 1980 Subject Classification: 62G05.

key words: bandwidth factorization, bandwidth selection, density estimation, kernel estimators, rates of convergence, smoothed cross–validation

short title: bandwidth selection

1. Introduction

The problem of data based smoothing parameter selection is described and motivated in Silverman (1986), Eubank (1988), Müller (1988), and Härdle (1990). See Marron (1988) for a survey of such methods proposed up until 1987. Recently there has been quite a variety of new methods proposed.

The point of this note is to show how a simple device allows substantial improvement in the asymptotically best bandwidth selectors, in the sense of eliminating their dependence on high order kernel pilot estimators (or related Fourier transform methods). For simplicity of presentation, the explicit discussion is given in terms of kernel density estimation. However the basic ideas, the methodology and the lessons clearly apply to a variety of other settings, including nonparametric regression, intensity and hazard estimation, and other estimators, including splines, histograms, and orthogonal series.

The kernel density estimator uses a sample $X_1,\ldots,X_n$ from a density f, to estimate the curve $f(x)$ by

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x-X_i), \tag{1.1}$$

where $K_h(x) = K(x/h)/h$, K is called the kernel function, and h is called the bandwidth (i.e. smoothing parameter). Good discussion of many important practical aspects of $\hat{f}_h(x)$ may be found in Silverman (1986), including the fact that bandwidth selection is crucial to implementation.

A useful tool for comparison of various methods of bandwidth selection, see for example Park and Marron (1990), is asymptotic rate of convergence to the optimum. In this paper we take "optimum" to be the minimizer, h_0, of the Mean Integrated Squared Error

$$\text{MISE}(h) = E\int(\hat{f}_h(x) - f(x))^2,$$

for the reasons discussed in Jones (1989), although other viewpoints are possible as discussed in that paper. Hall and Marron (1989) have shown that the best possible relative rate is $n^{-1/2}$, where n is the sample size.

While many bandwidth selectors have been proposed, the great majority fall short of this very fast (according to a nonparametric frame of reference) rate of convergence. For example Least Squares Cross–Validation has a very slow $n^{-1/10}$ rate of convergence (Hall and Marron 1987a, Scott and Terrell 1987), and while there are many ways to improve this, a rate of convergence of $n^{-1/2}$ has still proven to be rather elusive.

One means of achieving this rate was presented by Hall, Sheather, Jones and Marron (1989), but that approach lacks appeal for two reasons. First it requires the use of higher order kernels (including order 6 at one point) in the estimation of preliminary quantities. Second it uses an asymptotic expansion of MISE which needs to be carried to two terms in the bias. A selector with strong ties, and similar asymptotic behavior, may be found in Gasser, Kneip and Köhler (1989) in the related setting of nonparametric regression. Chiu (1989a,b) proposes different selectors, which have a similar flavor. Instead of explicit use of higher order kernels, he works with related Fourier transform methods. In Chiu (1989a) the two term bias expansion is not done, with the result that he obtains $n^{-1/2}$ convergence, but to a bandwidth different from the optimum.

Another means of obtaining $n^{-1/2}$ convergence to the optimum is based on the idea of Smoothed Cross–Validation, as proposed by Hall, Marron and Park (1989). The idea is to use the bandwidth, $\hat{h}$, which minimizes the criterion (there is a slight technical difference between this and the version defined there, as discussed in Section 2 below)

$$SCV(h) = (nh)^{-1}R(K) + \hat{B}(h),$$

where we use the functional notation $R(K) = \int K^2$ (which will be applied to other functions, e.g. $R(f) = \int f^2$, as well). Here

$$\hat{B}(h) = \int (D_h * \hat{f}_g)^2,$$

where * denotes convolution, $D_h = K_h - K_0$ (K_0 meaning the Dirac delta function), and $\hat{f}_g$ denotes a kernel density estimator with bandwidth g and kernel L, allowed to be different from h and K. One motivation for SCV is the fact that the first term is a good approximation of the integrated variance of $\hat{f}_h$, while $\hat{B}(h)$ provides an estimate of the true bias,

$$B(h) = \int (D_h * f)^2.$$

Other motivations may be found in Hall, Marron and Park (1989). Choice of g and L are treated in that paper. One drawback of the approach used there is that again higher order kernels (as discussed above, L needs to be of order 6) are required to get an $n^{-1/2}$ rate of convergence.

The use of higher order kernels (and equivalently, the Fourier transform methods of Chiu) is unappealing because, while they are excellent in the limit, "quite large" sample sizes seem to be required all too often before their beneficial effects begin to appear and become dominant. Of course "quite large" is a difficult thing to quantify, but we have seen a number of examples where higher order kernels are still not dominant even when the sample size is well into the millions. Another unattractive aspect of the use of higher order kernels in pilot estimates is that then "improved" versions of kernel estimates are used in estimating MISE while only the basic kernel estimate is employed for estimating f itself.

The point of this paper is that the fast rate of $n^{-1/2}$ convergence of a selected bandwidth to the optimum is not intrinsically connected to higher order kernels and the related Fourier transform ideas. It can also be achieved by another approach. This is most simply illustrated using the SCV methodology, but essentially the same results are easily established for the methodology of Hall, Sheather, Jones, and Marron (1989). An interesting sidelight, whose details will not be given here, is that the same set of ideas do not seem to apply in the same way to "solve the equation" methods, such as that discussed in Park and Marron (1990).

The main idea consists of allowing g to depend on h, which was not considered (except in one special case) by Hall, Marron and Park (1989). The dependence considered here is the factorization:

$$g = C n^p h^m,$$

for various constants C, p and m, detailed choices of which will be considered in the next section. This form is convenient, because the asymptotically optimal choices of both g and h can be written as constants multiplied by powers of n. Also it contains, as special cases, important ideas considered by previous authors. The case $m = 0$, $p = -1/9$ was the main one considered by Hall, Marron and Park. The case $m = 1$, $p = 1/10$ was proposed (not for SCV, but a related methodology) by Gasser, Kneip and Köhler (1989). This case is important because here C is scale invariant, so auxiliary scale estimation need not be done. Another case falling into this framework is $m = 1$, $p = 0$ as developed by Taylor (1989), although it is seen in Hall, Marron and Park (1989) that this choice is very far from asymptotically optimal.

In the next section, we present a theorem which gives the surprising result that when $m = -2$, there is an important type of cancellation which is the key to $n^{-1/2}$ convergence, even in the case $L = K$ where K is nonnegative. We have not yet been able to understand at an intuitive level why $m = -2$ (i.e. g is proportional to the inverse square of h) should give such special performance.

In Section 3 some simulation results are presented which demonstrate both the effectiveness and some limitations of our proposed bandwidth selector. Also discussed are some possible improvements. Proofs of the theorems in Section 2 may be found in Section 4.

2. Asymptotic Theory

The estimation of $B(h)$ by $\hat{B}(h)$ is closely related to the problem of kernel estimation of integrated squared density derivatives. In particular it involves summation of a matrix of terms where the diagonal entries are constant. In that context, Hall and Marron (1987b) suggested deleting those term which do not use the data, but Jones and Sheather (1990) have shown that there can be a substantial advantage to leaving these terms in. For the present analysis it is easy to handle both types if we introduce an auxiliary variable Δ, which takes on the value 0 when the diagonals are omitted and 1 when they are included. Since

$$\int (D_h * \hat{f}_g)^2 = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} (D_h * D_h * L_g * L_g)(X_i - X_j),$$

define

$$\hat{B}(h) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} (D_h * D_h * L_g * L_g)(X_i - X_j)\, 1_\Delta(i,j),$$

where $1_\Delta(i,j) = 1$ for $i \neq j$ and, $1_\Delta(i,i) = \Delta$ for $\Delta = 0$ or 1.

In addition to the functional notation R introduced in Section 1, it is convenient to introduce the notation, for $j = 1, 2, \ldots$,

$$m_j(K) = \int x^j K(x)\, dx.$$

Other convenient notation, used for both K and L, is

$$\tilde{K}(x) = -K(x) - xK'(x).$$

Assumptions used here are:

(A.1) K and L are symmetric probability densities, for which

$$m_j(K) < \infty, \qquad m_j(L) < \infty,$$

for $j = 1, \ldots, 8$.

(A.2) K and L have eight bounded continuous derivatives, and

$$\lim_{|x| \to 0} |x^j K^{(j)}(x)| = 0, \qquad \lim_{|x| \to 0} |x^j K^{(j)}(x)| = 0,$$

for $j = 1,\ldots,5$.

(A.3) for $j = 1,\ldots,6$,

$$\sup_{x \in \mathbb{R}} |f^{(j)}(x)| < \infty.$$

(A.4)
$$\sup_{x \in \mathbb{R}} |x^2 f''(x)| < \infty.$$

Assumptions (A.1) – (A.4) are far from the weakest possible. They are chosen for convenience of proof and simplicity of presentation. To weaken them is straightforward, but not done here because it would add much to the length of the proof, but little to the main points being made.

Furthermore we assume that g is of the form

(A.5)
$$g = g(h,n) = C\, n^p\, h^m,$$

for constants C, p and m. It will be seen that effective choice of p and m induces a linear constraint between them, and so one of them is left as a free parameter. Hence we will think of using the results of the next theorem to indicate how C and p should be chosen, for each given value of m. Different choices of m will be considered after that.

Let $\hat{h}$ denote the minimizer of SCV(h). Asymptotic properties of $\hat{h}$ are given by:

Theorem 1: under assumptions (A.1) – (A.5),

$$(\hat{h} - h_0)/h_0 = (n^{-4/5} h_0^{\,6} g_0^{\,-9} C_{\sigma 1} + n^{-1} C_{\sigma 2})^{1/2}\, Z_n +$$
$$+ (-n^{3/5} h_0^{\,3} g_0^{\,2} C_{\mu 1} + n^{3/5} h_0^{\,3} g_0^{\,4} C_{\mu 2} + n^{-2/5} h_0^{\,3} g_0^{\,-5} \Delta C_{\mu 3})^2,$$

where $g_0 = C\, n^p\, h_0^{\,m}$, Z_n is asymptotically N(0,1), and

$$C_{\mu 1} = (1+m/2)R(f'')^{-2/5}R(f^{(3)})R(K)^{-3/5}m_2(K)^{6/5}m_2(L)/5,$$
$$C_{\mu 2} = (1+m)R(f'')^{-2/5}R(f^{(4)})R(K)^{-3/5}m_2(K)^{6/5}\{m_4(L)+3m_2(L)^2\}/60,$$
$$C_{\mu 3} = R(f'')^{-2/5}R(K)^{-3/5}m_2(K)^{6/5}[(L^*L)^{(4)}(0) + m(L^*\tilde{L})^{(4)}(0)/2]/5,$$
$$C_{\sigma 1} = 2R(f)R(f'')^{-4/5}R(K)^{-6/5}m_2(K)^{12/5}R\{(L^*L)^{(4)}+m(L^*\tilde{L})^{(4)}/2\}/25.$$
$$C_{\sigma 2} = 4\{\int (f^{(4)})^2 f \,/\, R(f'')^2 - 1\}/25.$$

Now it will be shown that Theorem 1 can be used to derive good choices of C, p and m in $g = Cn^p h^m$. Note that the "variance" and "bias" given there can be combined to give an Asymptotic (Relative) Mean Square Error (for the selected bandwidth),

$$AMSE = n^{-4/5}h_0^{\,6}g_0^{\,-9}C_{\sigma 1} + n^{-1}C_{\sigma 2} +$$
$$+ (-n^{3/5}h_0^{\,3}g_0^{\,2}C_{\mu 1} + n^{3/5}h_0^{\,3}g_0^{\,4}C_{\mu 2} + n^{-2/5}h_0^{\,3}g_0^{\,-5}\Delta C_{\mu 3})^2.$$

The rest of the analysis depends on the case under consideration. Further useful notation is $C_0 = \{R(K) \,/\, (m_2(K)^2\, R(f''))\}^{1/5}$. Note that $h_0 \sim C_0\, n^{-1/5}$.

Case 1: $m \neq -2$, $\Delta = 0$.

It is simple to check that, given $m \neq -2$, the asymptotically best choices are

$$p = m/5 - 2/13, \qquad C = \{9C_{\sigma 1}/(4C_{\mu 1}^{\,2})\}^{1/13}/C_0^{\,m}.$$

The resulting rate of convergence is:

$$(\hat{h} - h_0)/h_0 \sim n^{-4/13}.$$

In the special case $m = 0$, this gives the same answer as in Section 4 of Hall, Marron and Park (1989).

Case 2: $m \neq -2$, $\Delta = 1$.

Here, following the main idea of Jones and Sheather (1990) the asymptotically best choices come from trading off the first and third terms in the bias part,

$$p = m/5 - 1/7, \qquad C = (C_{\mu 3}/C_{\mu 1})^{1/7}/C_0^{\,m}.$$

The resulting rate of convergence is:

$$(\hat{h} - h_0)/h_0 \sim n^{-5/14}.$$

Case 3: $m = -2$, $\Delta = 0$.

The asymptotically best choices are

$$p = -44/85, \qquad C = \{9C_{\sigma 1}/(8C_{\mu 2}{}^2)\}^{1/17} C_0{}^2.$$

The resulting rate of convergence is:

$$(\hat{h} - h_0)/h_0 \sim n^{-8/17}.$$

This is by far the best rate of convergence yet obtained for a bandwidth selector which does not use higher order kernels at any stage.

Case 4: $m = -2$, $\Delta = 1$.

Again using the Jones – Sheather idea, this time trading off the second and third terms in the bias part,

$$p = -23/45, \qquad C = (-C_{\mu 3}/C_{\mu 2})^{1/9} C_0{}^2.$$

The minus sign is not a problem, because for $m = -2$, $C_{\mu 2} < 0$. The resulting rate of convergence is:

$$(\hat{h} - h_0)/h_0 \sim n^{-1/2}.$$

As noted above, this very fast rate of convergence has been shown to be the best possible, even when the underlying density is known to be very smooth, in Hall and Marron (1989).

3. Simulations

A hurdle to actual use of the methodology discussed above is that optimal performance requires knowledge of the constants $C_{\mu 1}$, $C_{\mu 2}$, $C_{\mu 3}$ and $C_{\sigma 1}$. These in turn involve the unknown density in the form of $R(f^{(j)})$, for various j. Of course if only

asymptotic rate of convergence is of interest, then any value may be substituted for these, but obviously for any fixed n it is crucial to pay careful attention to the values of these constants. The most simple approach to this problem is to replace f by a reference distribution. An often used reference distribution is $N(0, \hat{\sigma}^2)$, where $\hat{\sigma}^2$ is some scale estimate.

This was implemented and tested on a variety of densities for $n = 100$ and 1000, with 500 replications in each case, using the sample variance for the scale estimate, for $\Delta = 0$, with $m = 0, 1, -2$. For computational speed the binned implementation ideas discussed in Härdle and Scott (1989) were used, and also for all calculations described below. Only the main conclusions from this are described here. When f was the Gaussian density (or not too far away in terms of shape), all the SCV methods were much better in all respects than Least Squares Cross–Validation. The faster rate of convergence for the case $m = -2$ is easily seen when n goes from 100 to 1000.

However, for densities which were somewhat further from the Gaussian in terms of shape, the performance was not so good. For example for the bimodal normal mixture density, .75 N(0,1) + .25 N(3/2,1/9) with $n = 100$, Figure 1, shows MISE(h), overlaid with kernel density estimates (using a normal reference distribution bandwidth, because of the limiting distributions) of the distributions of the $\hat{h}$'s. Note that for LSCV, the distribution is very spread, which is an artifact of the slow $n^{-1/10}$ rate of convergence. The selectors corresponding to the direct use of the normal reference distribution are shown as SCV1, for the case $m = 0$ and $\Delta = 0$ (i.e. Case 1), and SCV2 for $m = -2$ and $\Delta = 0$ (i.e. Case 3). These $\hat{h}$'s have much tighter distributions than LSCV, however there is substantial bias toward larger values (although, as predicted by the theory, this is substantially less severe for Case 3). This bias is large enough in the present case to give the $\hat{h}$'s worse average (over simulations) MISE($\hat{h}$) performance. One approach to this problem is to use $\Delta = 1$. This was tried, but while there was some improvement in average MISE, it was not sufficiently large that it seemed worth separately reporting (not

surprising from the fairly close rates of convergence in the various cases).

[put figure 1 about here]

A more promising method for handling this bias is motivated by the fact that the normal reference distribution, used in the choice of $C_{\mu 1}$, $C_{\mu 2}$, $C_{\mu 3}$ and $C_{\sigma 1}$, was too far from being accurate. To investigate this, we tried substituting kernel estimates of the unknowns, with their bandwidths chosen by the method of Jones and Sheather (1990). The resulting bandwidth distributions are shown as SCV3 in the case $m = 0$ and $\Delta = 1$, and as SCV4 in the case $m = -2$ and $\Delta = 1$. Observe that the bias is cut down somewhat (in the sense that the mass is moved closer to h_0), in the SCV3 distribution, as compared to SCV1, and also in SCV4 compared with SCV2. However, there in both cases there is a slight increase in the variability, which is due to the extra noise added by the additional estimation step. When assessing these in terms of average MISE($\hat{h}$), SCV3 is still worse than LSCV, but SCV4 is now better.

We also tried some densities which were extremely far from the normal in shape. Here all of the $\hat{h}$'s seemed unacceptable, with the above bias problem being so bad that the selected bandwidth was often more than three times the size of h_0. LSCV however, was as expected very variable, but gave better average performance because of its lack of bias.

4. Proofs

Here and below, for any function ℓ, $\ell^{[i]}$ denotes the i^{th} derivative of ℓ with respect to h. The randomness of $\hat{h}$ is driven by the variability of $\hat{B}(h)^{[1]}$, which is quantified in the following lemmas.

Lemma 4.1: under the assumptions (A.1) – (A.5),

$$E(\hat{B}(h)^{[1]}) = B(h)^{[1]} - h^3 g^2 C_{\mu 1} C_D + h^3 g^4 C_{\mu 2} C_D + n^{-1} h^3 g^{-5} \Delta C_{\mu 3} C_D + O(n^{-1}h^3 + n^{-1}h^4 g^{-6} + h^3 g^6 + h^5 g^2),$$

where

$$C_D = n^{-3/5} R(f'')^{2/5} R(K)^{3/5} m_2(K)^{4/5}.$$

The proof of Lemma 4.1 comes after it is shown how the lemmas imply Theorem 1.

Lemma 4.2: under the assumptions (A.1) – (A.5),

$$\mathrm{var}(\hat{B}(h)^{[1]}) = n^{-2} h^6 g^{-9} C_{\sigma 1} C_D{}^2 + n^{-11/5} C_{\sigma 2} C_D{}^2 + o(n^{-2}h^6 g^{-9}),$$

The proof of Lemma 4.2 follows that of Lemma 4.1.

Lemma 4.3: under the assumptions (A.1) – (A.5),

$$\{\hat{B}(h)^{[1]} - E(\hat{B}(h)^{[1]}) \,/\, \mathrm{var}(\hat{B}(h)^{[1]}\}^{1/2} \xrightarrow{d} N(0,1).$$

The proof of Lemma 4.3 is omitted because it follows from Lemmas 4.1 and 4.2 by a standard martingale argument, see for example Hall and Marron (1987a).

These Lemmas give Theorem 1 also by standard arguments as in Hall and Marron (1987a), for example. Hence only a rough sketch is given here. Since

$$0 = \mathrm{MISE}^{[1]}(\hat{h}) + \{(\mathrm{SCV} - \mathrm{MISE})^{[1]}(\hat{h})\}$$
$$= \mathrm{MISE}^{[2]}(h^*)(\hat{h} - h_0) + \{(\mathrm{SCV} - \mathrm{MISE})^{[1]}(\hat{h})\},$$

where h^* is between $\hat{h}$ and h_0, it follows that

$$(\hat{h} - h_0) = \frac{(\mathrm{SCV} - \mathrm{MISE})^{[1]}(\hat{h})}{\mathrm{MISE}^{[2]}(h^*)}.$$

Theorem 1 follows by suitable preliminary results showing $\hat{h}/h_0 \to 1$, and the representations

$$h_0 = R(K)^{1/5}\, m_2(K)^{-2/5}\, R(f'')^{-1/5}\, n^{-1/5} + O(n^{-2/5}),$$
$$MISE^{[2]}(h_0) = n^{-2/5}\, 5\, R(f'')^{3/5}\, R(K)^{2/5}\, m_2(K)^{6/5} + O(n^{-3/5}),$$

together with the fact that the dominant part of $SCV - MISE$ is $\hat{B}(h) - B(h)$.

Before proving the Lemmas, some convenient representations for $B(h)^{[1]}$ and $\hat{B}(h)^{[1]}$ will be established. Using the notation $D = K - K_0$ from section 1, note that

$$B(h)^{[1]} = 2 \int \{\int D_h(x-s)\, f(s)\, ds\} \{\int D_h^{[1]}(x-t)\, f(t)\, dt\}\, dx$$

(4.A)
$$= 2 \iint (D_h * D_h^{[1]})(s-t)\, f(s)\, f(t)\, ds\, dt$$
$$= (2/h) \iint (D * \tilde{K})_h(s-t)\, f(s)\, f(t)\, ds\, dt,$$

From

$$\int \{(D_h * \hat{f}_g)^2\}^{[1]} = 2 \int (D_h * \hat{f}_g)\, (D_h^{[1]} * \hat{f}_g + D_h * \hat{f}_g^{[1]}) =$$
$$2\, n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} (D_h * D_h^{[1]} * L_g * L_g + D_h * D_h * L_g^{[1]} * L_g)(X_i - X_j),$$

it follows that

(4.B)
$$\hat{B}(h)^{[1]} = 2\, n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} T_{ij},$$

where

$$T_{ij} = [\tfrac{1}{h}\{(D^*\tilde{K})_h * (L^*L)_g\} + \tfrac{m}{h}\{(D^*D)_h * (L^*\tilde{L})_g\}](X_i - X_j) 1_\Delta(i,j).$$

Proof of Lemma 4.1: Useful facts about the functions appearing in (4.B) are:

$$m_0(D^*\tilde{K}) = m_2(D^*\tilde{K}) = 0; \quad m_4(D^*\tilde{K}) = 12 m_2(K)^2;$$
$$m_0(L^*L) = 1; \quad m_2(L^*L) = 2m_2(L); \quad m_4(L^*L) = 2m_4(L) + 6m_2(L)^2;$$
$$m_0(D^*D) = m_2(D^*D) = 0; \quad m_4(D^*D) = 6m_2(K)^2;$$
$$m_0(L^*\tilde{L}) = 0; \quad m_2(L^*\tilde{L}) = 2m_2(L); \quad m_4(L^*\tilde{L}) = 4m_4(L) + 12m_2(L)^2.$$

Given symmetric functions a and b, for $i \neq j$,

$$E\{(a_h * b_g)(X_i - X_j)\} = \iint a_h(x-u) \{\int b_g(u-y)\, f(y)\, dy\}\, f(x)\, dx\, du.$$

But

$$\int b_g(u-y)\, f(y)\, dy \;=\; \sum_{k=0}^{2} g^{2k} f^{(2k)}(u)\, m_{2k}(b) / (2k)! + O(g^6),$$

and so

$$\begin{aligned} &E\{(a_h * b_g)(X_i - X_j)\} \;= \\ &= \sum_{k=0}^{2} g^{2k} \frac{m_{2k}(b)}{(2k)!} \iint a_h(x-u)\, f^{(2k)}(u)\, f(x)\, du\, dx + \\ &\qquad + \iint a_h(x-u)\, O(g^6)\, f(x)\, dx\, du, \end{aligned}$$

where the factor $O(g^6)$ is uniform over u, and independent of x. Next observe that for $k = 1,2,\ldots,5$.

$$\iint a_h(x-u)\, f^{(2k)}(u)\, f(x)\, du\, dx \;=\; \sum_{\ell=0}^{5} h^{2\ell} \frac{m_{2\ell}(a)}{(2\ell)!} (-1)^{k-\ell} R(f^{(k+\ell)}) + O(h^{12-2k}),$$

$$\iint a_h(x-u)\, O(g^6)\, f(x)\, dx\, du \;=\; m_0(a)O(g^6) + m_2(a)O(h^2 g^6) + O(h^4 g^6).$$

Applying these results in the above cases and simplifying yields, for $i \neq j$,

$$\begin{aligned} (4.C) \qquad E\, T_{ij} \;&=\; B(h)^{[1]}/2 - h^3 g^2 (1+m/2)\, R(f^{(3)})\, m_2(K)^2\, m_2(L) / 2 \\ &+ h^3 g^4 (1+m)\, R(f^{(4)})\, m_2(K)^2 \{m_4(L)+3m_2(L)^2\} / 24 \;+\; O(h^3 g^6 + h^5 g^2). \end{aligned}$$

The above provides a convenient representation for the $i \neq j$ terms in the double sum. To handle the remaining terms, note that again for symmetric a and b,

$$(a_h * b_g)(0) \;=\; \sum_{k=0}^{2} h^{2k} g^{-1-2k}\, m_{2k}(a)\, b^{(2k)}(0) / (2k)! \;+\; O(h^5 g^{-6}).$$

Using this in the above cases gives

$$T_{ii} \;=\; h^3 g^{-5}\, m_2(K)^2 \{(L*L)^{(4)}(0)/2 + m(L*\tilde{L})^{(4)}(0)/4\}\, \Delta + O(h^4 g^{-6}).$$

Lemma 1 follows from using this together with (4.C) applied to (4.B).

Proof of Lemma 4.2: Given symmetric, eight times uniformly continuously differentiable functions a, b, c, d with

$$m_0(a) = m_2(a) = m_0(c) = m_2(c),$$

we have

$$\int c(w)\, d(t+h(s-w)/g)\, dw \;=\; \sum_{\ell=0}^{8} \Lambda_\ell s^\ell + o(h^8 g^{-8}),$$

where

$$\Lambda_\ell \;=\; \sum_{k=\ell}^{8} h^k g^{-k}\, m_{k-\ell}(c)\, d^{(k)}(t)\, \binom{k}{\ell} / k!.$$

But for $\ell > 4$, $\Lambda_\ell = 0$, and for $\ell \le 4$, $\Lambda_\ell = O(h^{\ell+4} g^{-\ell+4})$, and

$$f(y+hs+gt) \;=\; \sum_{\ell'=0}^{3} (hs)^{\ell'}\, f^{(\ell')}(y+gt) / \ell'! \;+\; O(h^4),$$

and so

$$\iint a(s)\, c(w)\, d(t+h(s-w)/g)\, f(y+hs+gt)\, dw\, ds \;=$$
$$= h^8 g^{-8}\, f(y+gt)\, d^{(8)}(t)\, m_4(a)\, m_4(c)\, \binom{8}{4}/8! \;+\; o(h^8 g^{-8}).$$

Hence

$$\iiint a(s)\, c(w)\, d(t+h(s-w)/g)\, f(y+hs+gt)\, f(y)\, dw\, ds\, dy \;=$$
$$= h^8 g^{-8}\, R(f)\, d^{(8)}(t)\, m_4(a)\, m_4(c)\, \binom{8}{4}/8! \;+\; o(h^8 g^{-8}).$$

It follows from this that for $i \neq j$,

$$E\{(a_h * b_g)(X_i - X_j)\, (c_h * d_g)(X_i - X_j)\} \;=$$
$$= \iiint\int a(s)\, b(t)\, c(w)\, d(t+h(s-w)/g))\, f(y+hs+gt)\, f(y)/g\, ds\, dt\, dw\, dy$$
$$= h^8 g^{-9}\, R(f)\, m_4(a)\, m_4(c) \int b^{(4)} d^{(4)}\, \binom{8}{4}/8! \;+\; o(h^8 g^{-9}).$$

But by calculations of the type done in the proof of Lemma 4.1

$$E(a_h * b_g)(X_i - X_j) \;=\; O(h^4),$$

and similarly for $(c_h * d_g)$. Thus,

$$\mathrm{cov}\{(a_h * b_g)(X_i - X_j), (c_h * d_g)(X_i - X_j)\} \;=$$
$$= h^8 g^{-9}\, R(f)\, m_4(a)\, m_4(c) \int b^{(4)} d^{(4)}\, \binom{8}{4}/8! \;+\; o(h^8 g^{-9}),$$

and hence

$$(4.D)\quad \mathrm{var}(T_{ij}) \;=\; h^6 g^{-9}\, R(f)\, m_2(K)^4 \int \{(L^*L)^{(4)} + m(L^*\tilde{L})^{(4)}/2\}^2/4 + o(h^6 g^{-9}).$$

Next, for i, j, k all different

$$E\{(a_h * b_g)(X_i - X_j)\, (c_h * d_g)(X_i - X_k)\} \;=$$
$$= \iiiint\int a(s)\, b(u)\, c(t)\, d(v)\, f(x)\, f(x-hs-gu)\, f(x-ht-gv)\, ds\, dt\, du\, dv\, dx$$

$$= h^8 m_4(a)m_4(c)m_0(b)m_0(d)\int (f^{(4)})^2 f/(4!)^2 + o(h^8),$$

and so

$$\mathrm{cov}(T_{ij}, T_{ik}) = h^6 m_2(K)^4 \{\int (f^{(4)})^2 f - R(f'')^2\}/4 + o(h^6).$$

Thus from (4.B), Lemma 4.2 follows from (4.D) and

$$\mathrm{var}\{\hat{B}(h)^{[1]}\} = 4n^{-4}\{2 \sum_{i \neq j} \sum \mathrm{var}(T_{ij}) + 4 \sum_{i \neq j \neq k} \sum \sum \mathrm{cov}(T_{ij}, T_{ik})\}.$$

References

Chiu, S.–T. (1989a) An asymptotically optimal plug–in bandwidth estimate for density estimation and some remarks, unpublished manuscript.

Chiu, S.–T. (1989b) A stabilized bandwidth selection procedure for density estimation, unpublished manuscript.

Eubank, R. L. (1988) Spline smoothing and nonparametric regression, Dekker, New York.

Gasser, T., Kneip, A. and Köhler, W. (1989) A flexible and fast method for automatic smoothing and differentiation, unpublished manuscript.

Härdle, W. (1990) Applied nonparametric regression, Oxford University Press, Boston.

Härdle, W. and Scott, D. W. (1989) Smoothing in low and high dimensions by weighted averaging using rounded points, unpublished manuscript.

Hall, P. and Marron, J. S. (1987a) Extent to which least–squares cross–validation minimises integrated square error in nonparametric density estimation, Prob. Th. Rel. Fields, 74, 567–581.

Hall, P. and Marron, J. S. (1987b) Estimation of integrated squared density derivatives, Statist. Prob. Letters, 6, 109–115.

Hall, P. and Marron (1989) Lower bounds for bandwidth selection in density estimation, unpublished manuscript.

Hall, P., Marron, J. S., and Park, B. U. (1989) Smoothed cross–validation, unpublished manuscript.

Hall, P., Sheather, S. J., Jones, M. C. and Marron, J. S. (1989) On optimal data based bandwidth selection in kernel density estimation, unpublished manuscript.

Jones, M. C. (1989) The roles of ISE and MISE in density estimation, unpublished manuscript.

Jones, M. C. and Sheather, S. J. (1990) Using nonstochastic terms to advantage in kernel–based estimation of integrated squared density derivatives, unpublished manuscript.

Marron, J. S. (1988) Automatic smoothing parameter selection: a survey, Empirical Economics, 13, 187–208.

Müller, H.–G. (1988) Nonparametric analysis of longitudinal data, Springer Verlag, Berlin.

Park, B. U. and Marron, J. S. (1990) Comparison of data driven bandwidth selectors, J. Amer. Statist. Assoc., 85, 66–72.

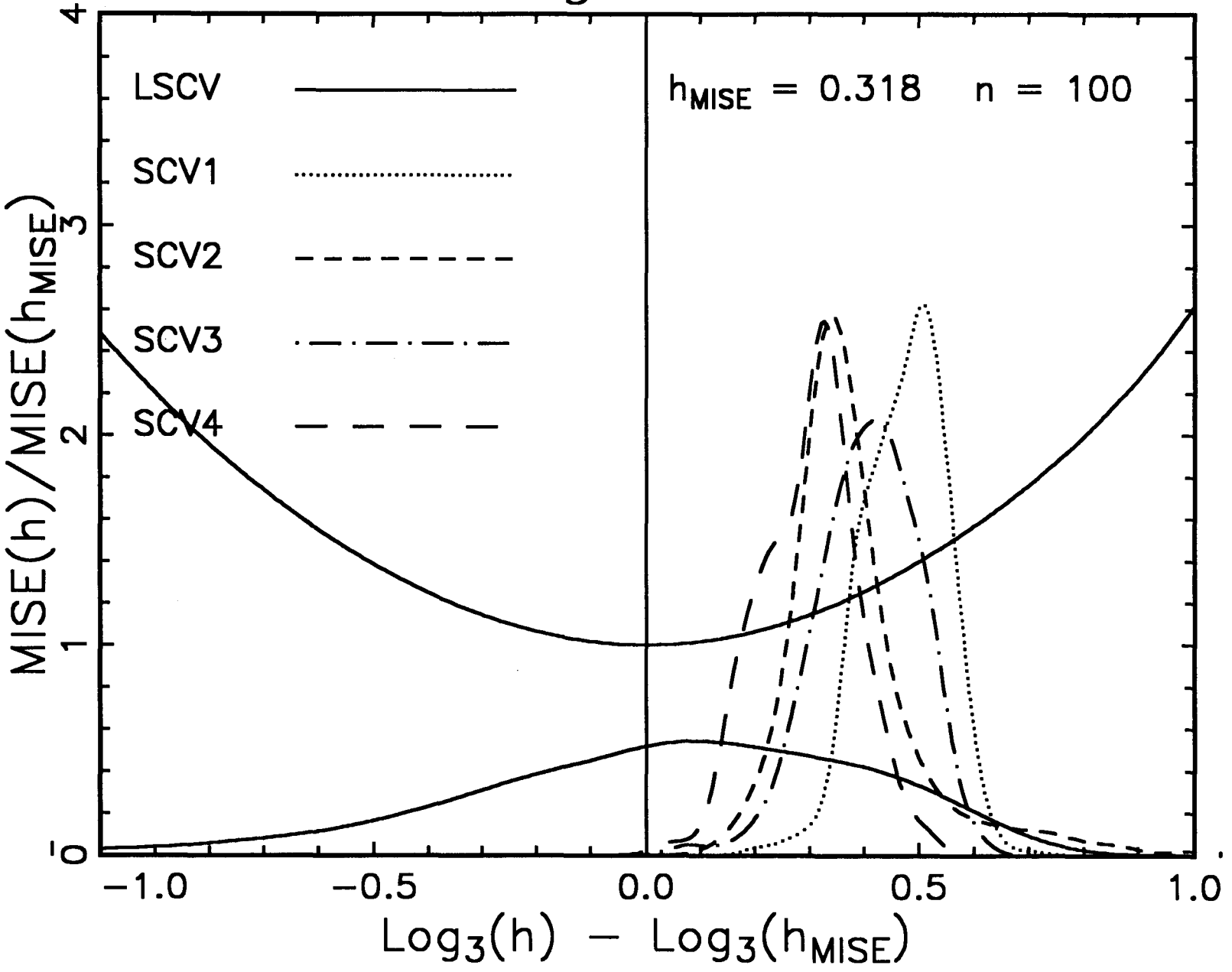
Scott, D. W. and Terrell, G. R. (1987) Biased and unbiased cross–validation in density estimation, J. Amer. Statist. Assoc., 82, 1131–1146.

Silverman, B. W. (1986) Density estimation for statistics and data analysis, Chapman and Hall, London.

Taylor, C. C. (1989) Bootstrap choice of the smoothing parameter in kernel density estimation, Biometrika, 76, 705–712.

Caption for Figure 1: MISE(h) together with kernel estimates of densities of selected bandwidths, for n = 100, bimodal mixture density. SCV1 is m = 0, SCV2 is m = –2, when standard normal reference distribution used. SCV3 is m = 0, SCV4 is m = –2, when pilot estimates are used.

Figure 1


LSCV
SCV1
SCV2
SCV3
SCV4
h_{MISE} = 0.318 n = 100
MISE(h)/MISE(h_{MISE})
0
1
2
3
4
−1.0
−0.5
0.0
0.5
1.0
$Log_3(h) - Log_3(h_{MISE})$