LOCALLY OPTIMAL WINDOW WIDTHS FOR KERNEL DENSITY ESTIMATION WITH LARGE SAMPLES

by

William R. Schucany

Technical Report No. SMU/DS/TR/214
Department of Statistical Science

January 1988

Department of Statistical Science
Southern Methodist University
Dallas, Texas 75275

LOCALLY OPTIMAL WINDOW WIDTHS FOR KERNEL DENSITY ESTIMATION WITH LARGE SAMPLES

William R. Schucany*
Southern Methodist University
Dallas, Texas 75275

Abstract

The smoothing parameter or window width for a kernel estimator of a probability density function at a point has been previously specified to minimize either asymptotic mean square error or asymptotic mean absolute error. In this note the ratio of these two widths is shown to be a constant for all kernels and density functions that satisfy the usual smoothness conditions. The fact that this ratio equals .985 supports recent comment that in this context these two error criteria do not yield large-sample results that differ by any meaningful amount.

Key words: *bias, mean square error, mean absolute error, smoothing parameter.*

* Research partially supported by ONR Contract N00014-85-K-0340 and performed while the author was on leave at the Institute for Advanced Studies, Australian National University.

1. Introduction

In a recent note Hall and Wand (1988) compare two asymptotically optimal window widths for kernel estimation of a probability density function, f, at a point. The two norms under investigation are mean square error (MSE) and mean absolute error (MAE). Their significant findings are that in most cases the two results differ by only a few percent. To illustrate the excellent agreement between the L_1 and L_2 coefficients of $n^{-1/5}$ they display graphs for four specific densities. In each case there are apparent singularities in the coefficients as a function of x, the point at which the value of the density is to be estimated. These occur at inflection points of f due to the fact that the formal expression for the bias vanishes. This is solely an artifact of the asymptotics and not a phenomenon that is manifest for finite samples.

In this note an alternative expression for the large sample MSE is examined briefly. The introduction of the next term of consequence in the bias expansion complicates the solution for the optimal h. A simple numerical solution permits comparisons with the closed form solution that has been utilized for large sample investigations since its introduction by Parzen (1962). The anomalous behaviour at inflection points is not present with the refined approximations. However, spurious spikes arise in other places. Rather than pursue higher-order approximations for bias and variance, a new comparison of the window widths is obtained. Their ratio is constant within the framework of exact values for asymptotic bias and variance. It follows that the relative size of these asymptotically optimal widths is the same for

every density and kernel. This result permits a succint summary of all of the numerical results in Hall and Wand (1988).

2. Asymptotic Mean Square Error

Consider a random sample of n univariate random variables, X_1, X_2, ..., X_n, from an absolutely continuous distribution with density function, f. The kernel estimator of f(x) for a single fixed x is

$$\hat{f}(x;h) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h}),$$

where K is a kernel, usually assumed to satisfy mild regularity conditions such as boundedness, $\int K(z)\, dz = 1$ and $\int |z^p K(z)|\, dz < \infty$ for some integer p such that

$$\int z^j K(z)\, dz = \begin{cases} 0, & j = 1, \ldots, p-1 \\ k_p \neq 0, & j = p \end{cases}.$$

The most widely used kernels in practice are themselves bounded, symmetric, finite-variance density functions for which p = 2 [see Silverman (1986)]. Assume that $f^{(4)}$ exists and is continuous at x and further that f(x) > 0. The familiar expression for the asymptotic MSE of $\hat{f}(x;h)$ is

$$\text{Var}\,[\hat{f}(x;h)] + \text{Bias}^2\,[\hat{f}(x;h)] = \frac{1}{nh} Q\, f(x) + [h^2 f''(x)\, k_2/2]^2, \quad (2.1)$$

where $Q = \int K^2(z)\, dz$.

Minimization of (2.1) with respect to h, assuming for the moment that $f''(x) \neq 0$, yields

$$h_2^* = \alpha_2(K)\, \beta_2(f) n^{-1/5}, \quad (2.2)$$

where $\alpha_2^5(K) = Q/k_2^2$ and $\beta_2^5(f) = f(x)/\{f''(x)\}^2$. Clearly, when $x = x_o$ is an inflection point, $f''(x_o) = 0$ and the formal expression (2.2) is infinite. When such is the case the expression for asymptotic bias in (2.1) no longer holds. Minimization of an appropriate expression for

MSE yields a different expression for the asymptotically optimal value of the smoothing parameter, h. When f"(x) vanishes

$$\text{Bias } [\hat{f}(x;h)] = h^4 f^{(4)}(x) k_4/4! + o(h^4).$$

Minimizing the corresponding MSE yields

$$h_4^* = \alpha_4(K)\ \beta_4(f) n^{-1/9}, \tag{2.3}$$

where $\alpha_4^9(K) = 72Q/k_4^2$ and $\beta_4^9(f) = f(x)/\{f^{(4)}(x)\}^2$. These solutions for h_2^* and h_4^* in (2.2) and (2.3) derive from the limiting conditions that as $n \to \infty$, $h \to 0$ and $nh \to \infty$. When one considers the behaviour of the optimal window, h, as a function of x for large finite n, intuition suggests that as $x \to x_o$ the transition from h_2^* to h_4^* is smooth rather than the abrupt change in order suggested by (2.2) and (2.3).

An approach that permits an examination of this issue is simply to use both of the terms in the bias expansion that will come into play. Letting $b_j = k_j\ f^{(j)}(x)/j!$ for $j = 2,4$, the bias may be expressed as

$$\text{Bias } [\hat{f}(x;h)] = h^2(b_2 + h^2 b_4) + O(h^4).$$

Substituting this in the expression for MSE changes the optimization problem to

$$\min_h \{Ah^4 + Bh^6 + Ch^8 + Dh^{-1}\}, \tag{2.4}$$

where $A = b_2^2 = k_2^2 \{f"(x)\}^2/4$,

$B = 2b_2b_4 = k_2k_4\ f"(x)\ f^{(4)}(x)/24$,

$C = b_4^2 = \{k_4\ f^{(4)}(x)/24\}^2$ and $D = Qf(x)/n$.

Differentiating the objective with respect to h and equating to zero yields

$$M(h) = h^5(4A + 6Bh^2 + 8Ch^4) - D = 0. \tag{2.5}$$

Since A and C are both non-negative, when $B \geq 0$ there is clearly only one real root of (2.5) and it corresponds to the desired minimum. There are regions in which $B < 0$ resulting in 3 real roots. Extraneous positive roots corresponding to relative maxima can be identified by the sign of

$M'(h)$. In practice the solution, h', is found by Newton's method. Select the initial value $h_1 = \min\{h_2^*, h_4^*\}$ and iterate the relation

$$h_{i+1} = h_i - M(h_i)/M'(h_i), \quad i = 1, 2, \ldots$$

until a specified tolerance is acheived. The sequence is well behaved and with this starting value convergence to 4 significant digits was generally realized within 3-6 iterations for all of the cases considered here. There are some numerical instabilities for some combinations of n and x; but this routine is not being proposed for any practical implementation rather only to illustrate the smoothness of $h'(x)$ in contrast to that of h_2^*. Some of the difficulties posed by multiple roots may be avoided by computing for a sequence of neighboring x values and succesively initializing with $h_1(x_{j+1}) = h'(x_j)$.

3. Numerical Results for Minimum MSE

One table of values is sufficient to illustrate the smoothness of h' as a function of both n and x. Take K to be the Gaussian kernel for which $Q = 1/\sqrt{4\pi}$, $k_2 = 1$ and $k_4 = 3$. This choice permits comparisons with $c_2^*(x)$, as displayed in Hall and Wand (1988), where $h_2^* = c_2^*(x)n^{-1/5}$. For the standard normal density $f(x) = \varphi(x)$ the derivatives that are required for the coefficients in (2.5) are $f''(x) = (x^2 - 1)\varphi(x)$ and $f^{(4)}(x) = (x^4 - 6x^2 + 3)\varphi(x)$. Table 1 contains values of c', an equivalent multiplier of $n^{-1/5}$. In other words, after finding h', the root of (2.5), numerically, the equivalent coefficient $c' = h'n^{1/5}$ is calculated for comparison with c_2^* obtained from (2.2).

Table 1

Asymptotically minimum-MSE window widths. Values in the table are c', the equivalent coefficient of $n^{-1/5}$. The entries for $n = \infty$ are c_2^*. Normal kernel and density.

n \ x	.75	.80	.85	.90	.95	1.00
50	1.363	1.400	1.437	1.475	1.515	1.556
100	1.366	1.419	1.475	1.533	1.593	1.654
500	1.370	1.452	1.548	1.657	1.778	1.909
1,000	1.371	1.462	1.573	1.705	1.858	2.030
10,000	1.373	1.482	1.628	1.829	2.109	2.491
100,000	1.373	1.491	1.655	1.902	2.315	3.057
∞	1.374	1.497	1.675	1.966	2.591	∞

Clearly, the singularity that occurs in $c_2^*(x)$ at $x = 1$ is not present in the more refined approximation for any sample size of practical interest. There are values of x for which the global minimum of (2.4) differs markedly from (2.2). The higher-order approximation has its own regions of peculiar behavior. They are simply relocated away from the inflection points. For this specific example it is possible to observe the overall smoothness in the true solution using exact expressions for bias and variance (see Fryer (1976)). Indeed, Table 1 and the exact results confirm that a good approximation to the optimal h is $cn^{-1/5}$ even for those values of x where the effective coefficient of h^2 in the bias vanishes.

4. Minimizing Absolute Error

With the aforementioned large-sample approximation in mind, consider the approximations

$$\text{Bias } [\hat{f}(x;h)] = b_x(h^2 + o(h^2))$$

and (4.1)

$$\text{Var } [\hat{f}(x;h)] = \sigma_x^2(\frac{1}{nh} + O(n^{-1})),$$

where b_x and σ_x^2 are non-vanishing smooth functions of x. The usual argument about balancing variance and bias squared dictates the choice of $h = cn^{-1/5}$, which implies that $MSE[\hat{f}(x;h)] = m_x n^{-4/5} + O(n^{-1})$. The dominant and second order terms may be comparable in size for very large n. This is a reminder of the caution that should be exercised in relying too heavily on the conventional asymptotics for this problem.

Using the notation in (4.1) the standard result for minimum MSE in (2.2) may be given an alternate expression

$$h_2 = c_2 n^{-1/5}, \tag{4.2}$$

where $c_2^{5/2} = \sigma_x/2b_x$. For purposes of comparison with the analogous result for minimum MAE let $v_2 = c_2^{5/2}$. This implies that for minimum MSE the coefficient in (4.2) must satisfy

$$\xi_2 = \frac{v_2 b_x}{\sigma_2} = \frac{1}{2}. \tag{4.3}$$

The scaled variable, vb/σ, can be recognized as the fundamental quantity in the development of minimum MAE by Hall and Wand (1988). In their notation, except that b_x and σ_x^2 are to be viewed as more accurate assessments of bias and variance, the equation that must be solved is

$$4(\frac{v_1 b_x}{\sigma_x})[\Phi(\frac{v_1 b_x}{\sigma_x}) - \frac{1}{2}] - \varphi(\frac{v_1 b_x}{\sigma_x}) = 0, \tag{4.4}$$

where, as before, $v_1 = c_1^{5/2}$ and $h_1 = c_1 n^{-1/5}$. In other words, to find

the window width that minimizes the asymptotic MAE, one must find the unique root $\xi_1 = v_1 b_x/\sigma_x$ of

$$4\xi_1[\Phi(\xi_1) - \tfrac{1}{2}] - \varphi(\xi_1) = 0. \qquad (4.5)$$

Given the value of ξ_1 the value of c_1 depends upon K and the unknown f(x) through b_x and σ_x.

The solutions to (4.3) and (4.5) are $\xi_2 = .500$ and $\xi_1 = .481$, respectively. The ratio c_1/c_2 does not depend upon b_x and σ_x and therefore holds for every kernel, density and x. Hence

$$\frac{h_1}{h_2} = \frac{c_1}{c_2} = \left(\frac{\xi_1}{\xi_2}\right)^{2/5} = \left(\frac{.481}{.500}\right)^{.4} = .985,$$

indicating that the asymptotically optimal window width for L_1 norm is uniformly smaller than that for the L_2 norm by 1.5%. This relationship is evident throughout the Table and Figures 3.1 given by Hall and Wand (1988). Their examples all involve the normal kernel, but clearly, their very interesting result holds with greater uniformity and more generally than the examples convey.

Remarks

A. An alternative view of (4.1) that is the basis for (4.4) is the representation

$$\hat{f}(x;h) - f(x) = n^{-2/5}(b_x - \sigma_x Z) + o_p(h^2),$$

where Z is asymptotically standard normal. This is perhaps also more satisfactory as a description of the essential difference between the current approach and the classical one. A problem with the conventional approach is that the approximation Bias (h) = $k_2 h^2 f''(x)/2$ is identically zero at points of inflection.

This does not occur in reality. Slightly more realistic results derive from the simple introduction of $b_x > 0$ for every x.

B. The comparison of c_1 and c_2 for minimizing the (global) L_1 and L_2 distances is given in Hall and Wand (1989). These results do not yield the constant ratio presented here for optimal estimation at a single fixed point. Nevertheless, the two coefficients do not differ by more than a few percent.

C. The usual regularity conditions that are sufficient to ensure the validity of these asymptotic comparisons are that one is using a non-negative second-order kernel, that $f(x) > 0$ and that f and f" are bounded and continuous at x. With less reliance here on the Taylor series expansion of f to get an expression for the bias, the constraints on f" are not necessary. However, smoothness of f' is, and as a practical matter there seem to be very few densities that possess one level of smoothness and not the other.

D. If one is truly focussing on f at a inflection point x_o, and $0 < f^{(4)}(x_o) < \infty$, then the optimal smoothing parameter is given in (2.3) for MSE. Similarly, minimum MAE is also produced by a h of order $n^{-1/9}$. The appropriate ratio, ξ_1/ξ_2, is a different constant deriving from altering the 2 to $\sqrt{8}$ in (4.3) and the 4 to an 8 in (4.5). The ratio of the associated widths follows as before with the exponent being 2/9 rather than 2/5.

References

Fryer, M.J. (1976). Some errors associated with the nonparametric estimation of density functions. *J.Inst.Maths.Applics.*, **18**, 371-380.

Hall, P. and Wand, M.P. (1988). On the minimization of absolute distance in kernel density estimation. *Statist.Prob.Letters*, To appear.

Hall, P. and Wand, M.P. (1989). Minimizing L_1 distance in nonparametric density estimation. *J.Multiv.Anal.*, To appear.

Parzen, E. (1962). On the estimation of a probability density function and the mode. *Ann.Math.Statist.* **40**, 1065-1076.

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