ON NONPARAMETRIC REGRESSION WITH HIGHER-ORDER KERNELS

by

William R. Schucany

Technical Report No. SMU/DS/TR/215
Department of Statistical Science

January 1988

Department of Statistical Science
Southern Methodist University
Dallas, Texas 75275

IEEE Trans Info.

On Nonparametric Regression with Higher-Order Kernels

William R. Schucany

Abstract

Estimation of the value of an unknown function at a point of continuity using kernel estimators is considered. Optimal bandwidths for second- and fourth-order kernels are derived and their asymptotic and large-sample efficiencies are examined. For fourth-order kernels generated by the generalized jackknife combination of two quadratics with different bandwidths, the optimal bandwidth ratio is derived. In the regression context the potential advantages of a fourth-order kernel make it a viable competitor to the usual second-order kernel estimator.

The author is with the Department of Statistical Science, Southern Methodist University, Dallas TX 75275 and partially supported by ONR Contract N00014–85–K–0340. This manuscript was written while he was on research leave at the Institute for Advanced Studies, Australian National University.

I. INTRODUCTION

One version of a nonparametric regression problem is to consider modelling independent response variables, Y_i, by

$$Y_i = m(t_i) + \epsilon_i \;, \qquad i = 1, \ldots, n \;, \tag{1.1}$$

where the ϵ_i are independent identically distributed random variables with mean zero and variance σ^2, the t_i are equispaced fixed design points on some finite interval and m is an unknown function. Without assuming more than a certain amount of smoothness of m, the desire may be to estimate m at a fixed point t.

The class of kernel estimators of $m(t)$ in this context is defined by

$$\hat{m}(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{t - t_i}{h}\right) Y_i \;, \tag{1.2}$$

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

$$\int z^j K(z)\, dz = \begin{cases} 0 \;, & j = 1, \ldots, p-1 \\ k_p \neq 0 \;, & j = p \;. \end{cases} \tag{1.3}$$

The most widely-used kernels are symmetric, finite-variance probability density functions. These have $p = 2$ and are called second-order kernels (see Silverman [8]).

This extension of kernel estimators to nonparametric regression was proposed by Priestley and Chao [5]. The estimators were originally defined by Rosenblatt [6] and generalized by Parzen [4] for probability density estimation. In both settings the consistency properties depend upon the sequence of bandwidths, $h = h(n)$, and characteristics of K and m. The expression for asymptotic mean-square error (mse) of $\hat{m}(t)$ in (1.2) is presented by Härdle [3] for the model in (1.1). The results are similar to the classical derivations for density estimation, which require $h \to 0$ such that $nh \to \infty$ as $n \to \infty$. Assuming that the unknown function is reasonably smooth, e.g., $m \in C^4[a, b]$, and that the kernel is such that its fourth moment is finite, ie. $k_4 < \infty$ in (1.3), leads to a useful expression for the asymptotic mse

$$\begin{aligned} \text{mse}\,[\hat{m}(t)] &= \text{Variance} + [\text{Bias}]^2 \\ &= \sigma^2 Q/nh + [h^2 m^{(2)}(t) k_2/2 + h^4 m^{(4)}(t) k_4/4!]^2 \\ &\quad + o(1/nh) + o(h^4)\,, \end{aligned} \tag{1.4}$$

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

Schucany and Sommers [7] introduced a generalized jackknife combination of two estimators of the type (1.2) designed to eliminate the h^2 term from the bias expansion in (1.4). If the same kernel is used for both estimators, then Härdle [3] and others have shown that the resulting estimator is again of the type (1.2). The kernel

$$K^*(z) = [K(z) - c^3 K(cz)]/(1 - c^2)\,,$$

where c is fixed and is in effect the ratio of the two bandwidths, is a fourth order kernel. In other words, it has been previously demonstrated that one can begin with a second-order kernel, K, and create a fourth-order kernel, K^*, so that

$$m^*(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_1} K^* \left(\frac{t - t_i}{h_1} \right) Y_i \tag{1.5}$$

has asymptotic mean-square error

$$\text{mse}\,[m^*(t)] = \sigma^2 Q^*/nh_1 + [h_1^4 m^{(4)}(t) k_4^*/4!]^2 + o(1/nh_1) + o(h_1^8)\,, \tag{1.6}$$

where Q^* and k_4^* are the quantities introduced in (1.4) and (1.3), respectively, evaluated at K^* rather than K.

Härdle [3] alerts potential users of m^* to the pitfalls of indiscriminate selection of the constants c and h_1. He illustrates with a specific example application that it is possible to choose c and h_1 in a manner that will produce a larger approximate mse for $m^*(t)$

than for $\hat{m}(t)$. Of course, in practice one has little ***a priori*** guidance on the best choices for either h or c and h_1. What is needed is a reliable data dependent procedure for selecting bandwidths. Local squared error cross validation [2] is a possible approach to this. However, until such procedures are refined and evaluated by convincing simulation experiments, it is of interest to compare $\hat{m}$ and m^* at their respective optimal bandwidths. In the next section expressions for the asymptotically optimal bandwidths are derived. For the estimator m^* the constant c is taken to be fixed. In subsequent numerical examples the sensitivity of the relative efficiency to the choice of c is examined and the optimal value is found.

II. OPTIMAL BANDWIDTHS AND EFFICIENCIES

Using only the dominant term of the bias component in expression (1.4), the asymptotic mse of $\hat{m}(t)$ can be readily minimized with respect to h. The solution is

$$h_{opt} = [\sigma^2 Q/(k_2 m^{(2)}(t))^2 n]^{1/5} \ . \tag{2.1}$$

Recall that the user selects the kernel, K, and thus can evaluate both Q and k_2. The noise variance, σ^2, and the value of the second derivative of the function are usually unknown.

Proceeding in a similar manner with (1.6) the asymptotically optimal bandwidth for $m^*(t)$ is

$$h^*_{opt} = [\sigma^2 Q^*/8n(k^*_4 m^{(4)}(t)/24)^2]^{1/9} \ . \tag{2.2}$$

As in (2.1) there are two unknown quantities, σ^2 and a higher order derivative of m. Henceforth, a bandwidth for $m^*(t)$ in (1.5) will be denoted by h^* rather than h_1 as in the previous section.

Evaluating the mse expressions (1.4) and (1.6) at their respective optimal bandwidths shows that mse$[\hat{m}(t)] = O(n^{-4/5})$ and mse$[m^*(t)] = O(n^{-8/9})$. Indeed all that is necessary for these rates to hold is that h be of order $n^{-1/5}$ and h^* in (1.5) be of order $n^{-1/9}$. Consequently, even without the information necessary to evaluate the coefficients in (2.1) and (2.2), the asymptotic relative efficiency (ARE) can be evaluated for any second-order kernel estimator relative to any fourth-order kernel estimator, as long as each have their bandwidth sequences decreasing at their proper rates. In other words, taking $h = an^{-1/5}$ and $h^* = a^* n^{-1/9}$ we have

$$ARE[\hat{m}(t), m^*(t)] = \lim_{n\to\infty} \frac{\text{mse}\,[m^*(t)]}{\text{mse}\,[\hat{m}(t)]} = 0\ .$$

Naturally, a question of greater practical importance is the value of this ratio of mse's in finite samples.

III. FINITE SAMPLE EFFICIENCIES

The specific choice of kernel for all subsequent illustrations is the one introduced by Epanechnikov [1]. This quadratic $K(z) = .75(1 - z^2)$, for $|z| \leq 1$ and zero otherwise, has been shown to have certain optimality properties over the class of second-order kernels. Even though the choice of kernel is not as critical to good performance as the selection of bandwidth, there are clearly theoretical and practical advantages to one with compact support. It follows that $Q = .6$, $k_2 = .2$ and $k_4 = 3/35$. For the associated fourth-order kernel, K^*, straightforward evaluation yields

$$k_4^* = -k_4/c^2$$

and

$$Q^* = .3(3c^3 + 6c^2 + 4c + 2)/(c+1)^2\ . \tag{3.1}$$

A reassessment of a specific example in [3] which compares the leading terms in (1.4) and (1.6) for $n = 100$, $\sigma^2 = 1$ and $m(t) = \sin(t)$ at $t = \pi/4$ is presented in Table I. The optimal values of h and h^* are evaluated from (2.1) and (2.2) and these substituted into the large sample approximations for variance and bias^2. A range of values of c is used to illustrate the dependence of h^*_{opt} on c, as well as the relative insensitivity of optimal mse $[m^*(t)]$. The column labelled relative efficiency is the ratio of mse $[\hat{m}(t)]$ to mse $[m^*(t)]$.

TABLE I

Large Sample Approximations for Second-Order Kernel and Several Fourth-Order Kernels.

kernel	optimal h	variance	bias^2	mse	relative efficiency
K	.786	.00763	.00191	.00954	
$K^*: c$					
.4	1.155	.00630	.00079	.00709	1.346
.5	1.286	.00609	.00076	.00685	1.392
.6	1.406	.00601	.00075	.00676	1.412
.7	1.518	.00600	.00075	.00675	1.415
.8	1.624	.00603	.00075	.00678	1.407
.9	1.725	.00609	.00076	.00685	1.392

It is clear that m^* has the potential to be about 40% more efficient than the best achievable by $\hat{m}$. This comparison for this specific example becomes even more favourable for m^* with larger sample sizes or better signal-to-noise ratio ($\sigma^2 < 1$). For $\sigma^2 = .5$ (which is equivalent to maintaining $\sigma^2 = 1$ but increasing n to 200) the figure is about 50%, at $\sigma^2 = .2$ it is approximately 60% and at $\sigma^2 = .1$, about 70%. The same insensitivity to c persists for the alternative values of σ^2 and n and at $t = \pi/8$ and $t = \pi/2$.

There is an obvious problem with the formal expressions (2.1) and (2.2) at points where the second or fourth derivatives of the function vanish. This is a familiar problem for density estimators. Clearly, the leading term for the bias in (1.4) vanishes at a point of inflection and the optimal rate for $\hat{m}$ may be improved for that single special case. Near 0 (or π) for this specific example both $m^{(2)}(t)$ and $m^{(4)}(t)$ are small and m^* has relatively greater difficulty than $\hat{m}$. With $\sigma^2 = .1$ and $n = 200$ the comparison did not favor $\hat{m}$ unless $|t| < .03$. To a great extent this difficulty is only an artifice of the asymptotics. For most functions, when $m^{(2)}(t)$ vanishes, it is not the case that the bias $[\hat{m}(t)]$ is identically zero. In practice with finite samples an appropriate adaptive procedure will deal with estimates of variance and bias witout regard to any terms of a Taylor expansion.

These comparisons were also made for the functions $m(t) = e^{2t}$ and $m(t) = t^{-1}$ with similar results for the fixed values of t where the functions are changing. In general m^* did not exhibit better performance than $\hat{m}$ in regions where the function is quite flat, e.g. for e^{2t} when $t < -2$. It is clear that this is due to the fact that $\hat{m}$ has a small bias in such cases and the bias reduction feature of m^* is of little use.

IV. OPTIMAL BANDWIDTH RATIO

The construction of the fourth-order kernel, K^*, was originally motivated as a linear combination of two second-order kernels with the same K and bandwidths h_1 and h_2 such that $c = h_1/h_2$. It may be observed in Table I, as in other numerical examples, that the best value for the bandwidth ratio, c, is in the neighbourhood of .7. An analytical approach to the selection of c to maximize the ARE is possible. This is equivalent to minimizing $\text{mse}\,[m^*(t)]$ at h^*_{opt} with respect to c. This process yields the minimum with respect to h^* and c simultaneously.

By substituting (2.2) into (1.6) one can show that the optimal mse to be minimized with respect to c is

$$\text{mse}\,[m^*(t)] = (\sigma^2/n)^{8/9}(m^{(4)}/24)^{29}(8^{1/9} + 8^{-8/9})\left[(Q^*)^{8/9}(k_4^*)^{2/9}\right] . \tag{4.1}$$

Hence an equivalent objective function is proportional to $(Q^*)^2/c$. For the Epanechnikov kernel, Q^* from (3.1) implies that an equivalent objective is

$$\frac{3c^3 + 6c^2 + 4c + 2}{c^{1/2}(c+1)^2} . \tag{4.2}$$

Differentiation of (4.2) with respect to c leads to c_{opt} as the unique positive real root of

$$c^5 + 4c^4 + 5c^3 - \tfrac{8}{3}c - \tfrac{2}{3} = 0 .$$

By Newton's method the solution is $c_{opt} = .670854$. This is the recommended choice for a specific member of the class of estimators defined by (1.5). The remaining task for the user of either $\hat{m}$ or m^* is the proper selection of h or h^*, respectively.

A general expression may be derived for the degradation of the efficiency of m^* if the optimal bandwidth h^*_{opt} in (2.2), is not the value actually used. Neglecting higher order terms the notation in (1.6) may be simplified to

$$\text{mse}\,[m^*(t)] = A/h^* + B(h^*)^8 \,. \tag{4.3}$$

When $h^* = h^*_{opt}$, it can be seen from (4.1) that the ratio of A/h^*_{opt} to $B(h^*_{opt})^8$ is 8.0. It follows then from (4.3) that

$$M_0 = \text{mse}\,[m^*(t; h^*_{opt})] = 9B(h^*_{opt})^8 \,. \tag{4.4}$$

Next consider letting $h^*_\epsilon = (1+\epsilon)h^*_{opt}$, where ϵ may be positive or negative and represents a possible relative error in the selection of the bandwidth. Substituting this in (4.3) and using the same relationship that led to (4.4), produces an expression for mse that depends on ϵ and will be denoted by M_ϵ. It is straightforward to obtain

$$M_\epsilon = \text{mse}\,[m^*(t; h^*_\epsilon)] = B(h^*_{opt})^8[8/(1+\epsilon) + (1+\epsilon)^8]$$

and then

$$\frac{M_0}{M_\epsilon} = \frac{9(1+\epsilon)}{8+(1+\epsilon)^9} \,. \tag{4.5}$$

Evaluation of this simple expression (4.5) at some typical values, such as $\epsilon = .10$ and .20, raises two pertinent points. First, if one selects a bandwidth within 10% of its optimal, then the mse of m^* will be within 4.5% of its minimum value, M_0. Second, it is better to have $h^* < h^*_{opt}$ than to exceed h^*_{opt} by the same amount. For example when $\epsilon = -.20$ the efficiency is 11.5% below optimum and yet at $\epsilon = .20$ it is almost 18% below.

Note that in general the ratio of the variance term to bias^2 term in the minimized expression for mse is equal to $2p$. It follows that the ratio M_0/M_ϵ, similar to that derived at (4.5), is $5(1+\epsilon)/[4+(1+\epsilon)^5]$ for the second-order kernel estimator. This tends to indicate a degradation of $\hat{m}$ due to an error in assessing h_{opt} that is approximately half of that occurring due to a comparable error in h^*.

V. DISCUSSION

The better asymptotic rate for fourth-order kernels is widely recognized. However, in the context of probability density estimation they have not been employed to a great extent because they do not produce only non-negative estimates. In the regression setting this is obviously not generally a flaw. In Section 3 specific illustrations show the potential advantages of using a fourth-order kernel over the best second-order kernel. Clearly, the estimator m^* is not uniformly superior to $\hat{m}$, but the gains are substantial at points where the unknown regression function is not approximately constant.

In Section 4 for the specific case of the Epanechnikov kernel the optimal bandwidth ratio is found to be approximately 0.671. Using this value for c^* in the kernel in (1.5), reduces the difficulty of employing a fourth-order kernel to the same that exists for second-order kernels. In either case a bandwidth must be selected. To the extent that this may be accomplished without reliance upon subjective values, the question remains one of relative stability. Finite sample experimentation with adaptive estimates of h and h^* will shed more light upon the practical effectiveness of m^* for nonparametric regression.

REFERENCES

[1] V.A. Epanechnikov, "Nonparametric estimation of a multivariate probability density," *Theory Prob. Appl.*, vol.14, pp.153–158, 1969.

[2] P. Hall and W.R. Schucany, "A local cross-validation algorithm", submitted to *Statist. Prob. Letters*, 1988.

[3] W. Härdle, "A note on jackknifing kernel regression function estimators," *IEEE Trans. Inform. Theory*, vol.IT-32, pp.298–300, 1986.

[4] E. Parzen, "On estimation of a probability density function and its mode," *Ann. Math. Statist.*, vol.33, pp.1065–1076, 1962.

[5] M.B. Priestley and M.T. Chao, "Nonparametric function fitting," *J. Roy. Statist. Soc. B*, vol.34, pp.385–392, 1972.

[6] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Ann. Math. Statist.*, vol.26, pp.832–837, 1956.

[7] W.R. Schucany and J.P. Sommers, "Improvement of kernel-type density estimators," *J. Amer. Statist. Ass.*, vol.72, pp.420–423, 1977.

[8] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall, 1986.