#1754

Comparison of Kernel Regression Estimators

by

C. K. Chu and J. S. Marron*

University of North Carolina, Chapel Hill

February 22, 1988

ABSTRACT

For nonparametric regression, in the case of fixed design points, the two most popular methods for constructing kernel estimators, involve choosing weights either by kernel evaluation or by subinterval integration. While these estimators are very nearly the same in the case of equally spaced design points, it is shown here that the second method will typically make less efficient use of the data in the case of unequal spacing.

AMS 1980 subject classifications: Primary 62G05; Secondary 62G20.

Keywords: asymptotic variance, design points, kernel estimators, nonparametric regression.

*Research partially supported by NSF Grant DMS-8701201.

1. INTRODUCTION

Kernel estimators in fixed design, nonparametric regression settings are weighted averages of the response variables. Nadaraya (1964) and Watson (1964) proposed choosing the weights by evaluating a kernel function at the design points, and then dividing by the sum, so that the weights sum to 1. A drawback to this approach is that the estimator is difficult to handle from a technical view point. Indeed, for some kernel functions, Haerdle and Marron (1983) have shown that the moments of such an estimator will typically fail to exist, when the predictor variables are also random. To overcome these technical problems, Clark (1977, 1980) and Gasser and Mueller (1979) have proposed taking the weights to be, either implicitly or explicitly, integrals of the kernel function on small subintervals which contain the design points. If the design points are equally spaced, the two estimators are roughly the same by the integral mean value theorem. However, when the design points are not equally spaced, the estimator with integral weights has the drawback, from an intuitive point of view, that observations whose design points have very near neighbors on each side tend to be down-weighted. This means that their role in the weighted average is less than it should be, which results in an inefficient estimate.

Section 2 gives a precise formulation of the estimators. An example demonstrating the inefficiency of the integral weighted estimator in the unequally spaced case is given in Section 3. Section 4 contains a brief discussion of the implications for the random design case. Proofs are given in Section 5.

2. THE ESTIMATORS

The fixed design nonparametric regression model considered here is

$$Y_j = m(x_j) + \epsilon_j,$$

for $j = 1, 2, \ldots, n$, where the Y_j's are observed random variables, the x_j's are nonrandom design points with $0 \le x_1 \le x_2 \le \ldots \le x_n \le 1$ (without loss of generality), and the ϵ_j's are independent random variables with mean 0 and variance σ^2.

For a kernel function K, and a bandwidth h, Nadaraya (1964) and Watson (1964) introduced the following estimator of m(x), for $0 < x < 1$, based on evaluation of the kernel:

$$\hat{m}_E(x) = \left[n^{-1} \sum_{j=1}^{n} K_h(x - x_j)\, Y_j \right] / \left[n^{-1} \sum_{j=1}^{n} K_h(x - x_j) \right],$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$.

This estimator is rather tedious to analyze mathematically, because of the summation in the denominator. For this and other reasons, another type of kernel estimator, based on integration of the kernel over a subinterval, was considered by Clark (1977, 1980) and Gasser and Mueller (1979). For $0 < x < 1$, the Gasser Mueller form of the estimator is given by:

$$\hat{m}_I(x) = \sum_{j=1}^{n} Y_j \int_{s_{j-1}}^{s_j} K_h(x - t)\, dt,$$

where $s_0 = 0$, $s_n = 1$, $x_j \le s_j \le x_{j+1}$ for $j = 1, \ldots, n-1$, and K is a density function. An obvious choice of the s_j, which will be made in the following (although the conclusions are clearly not dependent on this) is

$$s_j = (x_j + x_{j+1}) / 2.$$

The Clark version of the integral weighted estimator uses different notation, but is of essentially the same form.

See the monograph Haerdle (1988) for detailed discussion of these and related estimators.

3. AN EXAMPLE

An example which shows clearly that the integral weighted estimator, $\hat{m}_I$, will typically be inefficient in the unequally spaced design case will now be constructed. Recall that the intuition is that points with nearest neighbors too close will be downweighted. This effect can be studied by starting with an equally spaced design, and considering consecutive triples of points. For each triple move the first and the third towards the center. The amount of shift to the center can be parametrized by a value $\alpha \in [0,1]$, which results in the design,

$$x_j = \begin{cases} ((3\ell+2)-\alpha) / n & \text{if } j = 3\ell+1 \\ (3\ell+2) / n & \text{if } j = 3\ell+2, \\ ((3\ell+2)+\alpha) / n & \text{if } j = 3\ell+3 \end{cases}$$

for $\ell = 0, 1, \ldots, (n/3)-1$, where it has been assumed that n is a multiple of 3. Note that $\alpha = 1$ gives the usual equally spaced design, while $\alpha = 0$ gives a design which is also essentially equally spaced with three replications at each point. Such a design is clearly artificial, and it is difficult to think of a practical situation where it would arise, but it is considered here because it provides a clear and simple illustration of the point being made. The effects described will obviously also be present in more realistic unequally spaced designs.

The inefficiency of the integral type estimator can now be seen at an intuitive level by considering the effect on the weights given to the center observation of each triple, as α varies between 0 and 1. Note that the weights given to these points by the kernel evaluation estimator, $\hat{m}_E$, is nearly independent of α, while the weights assigned by the subinterval integration estimator, $\hat{m}_I$, are nearly proportional to α. Hence, for α close to 0, the weight on the center obaservation is essentially 0, so the weighted average, $\hat{m}_I$, is making use of only 2/3 of the available observations. The extent of the inefficiency caused by this can be quantified by studying the asymptotic variance.

Assume that the kernel function K is a density function with support contained in the interval [-1, 1], and that the kernel function K and regression function m are Hoelder continuous. Then as $n \to \infty$, $h \to 0$, with $nh \to \infty$, for $0 < x < 1$, it is shown in Section 5 that

$$\text{(3.1)} \qquad \mathrm{Var}(\hat{m}_E(x)) = n^{-1}h^{-1}\sigma^2\int K^2 + O(n^{-2}h^{-2}),$$

$$\text{(3.2)} \qquad \mathrm{Var}(\hat{m}_I(x)) = C(\alpha)\, n^{-1}h^{-1}\sigma^2\int K^2 + O(n^{-2}h^{-2}),$$

where $C(\alpha) = (\alpha^2 - 2\alpha + 3) / 2$, $0 \leq \alpha \leq 1$.

Observe that for $\alpha = 1$, the estimators have essentially the same performance, which, as remarked above, is to be expected from the integral mean value theorem. However, in the opposite case of $\alpha = 0$, note that $\hat{m}_I$ will have 3/2 times the variance of $\hat{m}_E$, which, in view of the above intuition, is also to be expected, because then $\hat{m}_I$ is only using 2/3 of the available data. Of course, if one really had three replications at each design point (as we have when $\alpha = 0$), the obvious thing to do is pool, by working with the average of the observations at these points. But it is a compelling feature of $\hat{m}_E$ that it makes this adjustment automatically, as $\alpha \to 0$, while $\hat{m}_I$ has a disturbing tendency

to delete an observation.

While it is the variance that quantifies the inefficiency of the integral weighted estimators, as with any smoothing method attention must also be paid to the bias. Both methods are, at least asymptotically, the same in the following sense. Under the same assumptions as made for (3.1) and (3.2), as $n \to \infty$, $h \to 0$, with $nh \to \infty$, for $0 < x < 1$, it is shown in Section 5 that

$$(3.3)\quad \text{Bias}(\hat{m}_E(x)) = \int K_h(x - t)\,(m(t) - m(x))\,dt + O(n^{-1}h^{-1}),$$

$$(3.4)\quad \text{Bias}(\hat{m}_I(x)) = \int K_h(x - t)\,(m(t) - m(x))\,dt + O(n^{-1}).$$

Note that the first terms in the above representations of the biases for the two estimators are the same. It is in the second terms that a difference shows up, however when the squared bias is combined with the variance to form the mean squared error, the second terms are both seen to be of lower order. Note that the integral weighted estimator has better properties with respect to these lowere order terms, which quantifies the remark made in the introduction that the kernel weighted estimator is technically more difficult to analyze.

The results of this section may be generalized in a straight forward fashion, to the case of forming clusters of k points, instead of just three as done above. When this is done, all of the above results remain the same, except (3.2) becomes

$$(3.5)\quad \text{Var}(\hat{m}_I(x)) = C_k(\alpha)\, n^{-1}h^{-1}\sigma^2\int K^2 + O(n^{-2}h^{-2}),$$

where $C_k(\alpha) = ((k-2)\alpha^2 - 2(k-2)\alpha + k) / 2$, $0 \le \alpha \le 1$. Note that the downweighting effect of the integral weighted estimator can be made arbitrarily bad here, subject of course to the fact these asymptotics describe only the situation where $nh >> k$.

4. RANDOM DESIGN CASE

The main point illustrated by the above example, that observations whose nearest neighbors are too close will be down weighted by the integral weighted estimator, has implications in the random design case (i.e. where the x_j are chosen by some random mechanism) as well. In particular, in that context, just by chance, some design points will certainly have nearest neighbors closer than others. In view of the ideas illustrated by the example of the previous section one would expect intuitively that, when expectations are taken with respect to the randomness of both the responses and the design points, the integral weighted estimator should have higher variance. This has indeed been observed by Jennen-Steinmetz and Gasser (1987).

This does not mean that the kernel weighted estimator is superior to the integral weighted estimator in this setting because the biases are different, and it is clear that sometimes one will have less bias, and sometimes the other. An interesting feature of the integral weighted estimator is that the design density does not appear in the bias, although it is not clear that this is an advantage, because the design certainly is a part of the estimation setting.

5. PROOFS

For the proofs of (3.1) and (3.3), we first derive an asymptotic expression of the denominator of $\hat{m}_E(x)$. Since the design points are

clustered in groups of three, rearrange

$$n^{-1} \sum_{j=1}^{n} K_h(x-x_j) = A_1 + B_1 + C_1,$$

where

$$A_1 = n^{-1} \sum_{\ell=0}^{(n/3)-1} K_h(x - x_{3\ell+1}),$$

$$B_1 = n^{-1} \sum_{\ell=0}^{(n/3)-1} K_h(x - x_{3\ell+2}),$$

$$C_1 = n^{-1} \sum_{\ell=0}^{(n/3)-1} K_h(x - x_{3\ell+3}).$$

Each of these summations is over an equally spaced grid, with width $3/n$, so A_1, B_1, and C_1 are all Riemann sums for

$$(1/3) \int_0^1 K_h(x - t)\, dt.$$

Hence, by the Hoelder continuity of K,

$$n^{-1} \sum_{j=1}^{n} K_h(x - x_j) = 1 + O(n^{-1}h^{-1}). \tag{5.1}$$

Using (5.1), the variance and bias of $\hat{m}_E(x)$ can be expressed as

$$\mathrm{Var}(\hat{m}_E(x)) = n^{-2} \sigma^2 \sum_{j=1}^{n} \left[K_h(x - x_j) \right]^2 \left[1 + O(n^{-1}h^{-1})\right],$$

$$\mathrm{Bias}(\hat{m}_E(x)) = n^{-1} \sum_{j=1}^{n} K_h(x - x_j)\, (m(x_j) - m(x))\, (1 + O(n^{-1}h^{-1})).$$

The Riemann summation methods used above show that

$$n^{-2} \sum_{j=1}^{n} \left[K_h(x - x_j) \right]^2 = n^{-1}h^{-1} \int K^2 + O(n^{-2}h^{-2}), \tag{5.2}$$

and

$$n^{-1} \sum_{j=1}^{n} K_h(x-x_j)\,(m(x_j)-m(x)) = \int K_h(x-t)\,(m(t)-m(x))\,dt + O(n^{-1}h^{-1})$$

asymptotically. This completes the proof of (3.1), (3.3).

The proof of (3.4) is immediate from equation (6) of Gasser and Mueller (1984).

For the proof of (3.2), note that by the integral mean value theorem,

$$\begin{aligned} \mathrm{Var}(\hat{m}_I(x)) &= \sigma^2 \sum_{j=1}^{n} \left[\int_{s_{j-1}}^{s_j} K_h(x-t)\,dt \right]^2 \\ &= \sigma^2 \sum_{j=1}^{n} \left[s_j - s_{j-1} \right]^2 \left[K_h(x-t_j) \right]^2, \end{aligned}$$

for some $t_j \in [\, s_{j-1}, s_j]$, $j = 1, 2, \ldots, n$. So by the Hoelder continuity of K,

$$\begin{aligned} \mathrm{Var}(\hat{m}_I(x)) &= \sigma^2 \sum_{j=1}^{n} \left[s_j - s_{j-1} \right]^2 \left[K_h(x-x_j) \right]^2 + O(n^{-2}h^{-2}) \\ &= \sigma^2 \{ \alpha^2 A_2 + \left[\frac{3-\alpha}{2}\right]^2 B_2 + \left[\frac{3-\alpha}{2}\right]^2 C_2 \} + O(n^{-2}h^{-2}), \end{aligned}$$

where, for $\ell = 0, 1, \ldots, (n/3)-1$,

$$s_j - s_{j-1} = \begin{cases} \alpha/n & \text{if } j = 3\ell+2, \\ ((3-\alpha)/2n) & \text{if } j = 3\ell+3,\ 3\ell+4 \end{cases}$$

and where

$$A_2 = n^{-2} \sum_{\ell=0}^{(n/3)-2} \left[K_h(x - x_{3\ell+2}) \right]^2,$$

$$B_2 = n^{-2} \sum_{\ell=0}^{(n/3)-2} \left[K_h(x - x_{3\ell+3}) \right]^2,$$

$$C_2 = n^{-2} \sum_{\ell=0}^{(n/3)-2} \left[K_h(x - x_{3\ell+4}) \right]^2.$$

Since A_2, B_2, C_2 all have the same asymptotic expression,

(1/3) $n^{-1}h^{-1}\sigma^2\int K^2$, which follows from (5.2), we conclude that $\mathrm{Var}(\hat{m}_I(x))$ has an asymptotic expression

$$\sigma^2 \{ \alpha^2 + \left[\frac{3-\alpha}{2}\right]^2 + \left[\frac{3-\alpha}{2}\right]^2 \} (1/3)\, n^{-1}h^{-1}\sigma^2\int K^2 + O(n^{-2}h^{-2})$$
$$= C(\alpha)\, n^{-1}h^{-1}\sigma^2\int K^2 + O(n^{-2}h^{-2}).$$

The proof of (3.2) is complete.

The proof of (3.5), the case of clustering k points, can be easily derived by the same procedure as for the case k = 3.

References

Clark, R. M. (1977). Nonparametric Estimation of a Smooth Regression function. Journal of the Royal Statistical Society, Series B, 39, 107-113.

Clark, R. M. (1980). Calibration, cross-validation and carbon-14, II. Journal of Royal Statistical Society, Series A, 143, 177-194.

Gasser, T. and Mueller, H. G. (1979). Kernel estimation of regression functions. In Smoothing techniques for curve estimation. Lecture Notes in Math. 757,23-68, New York: Springer-Verlag.

Gasser, T. and Mueller, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. Scand. J. Statist. 11, 171-185.

Haerdle, W. (1988). Applied Nonparametric Regression, unpublished manuscript.

Haerdle, W. and Marron, J. S. (1983). The nonexistence of moments of some kernel regression estimators. North Carolina Institute of Statistics, Mimeo Series No.1537.

Jennen-Steinmetz and Gasser (1987). A unifying approach to nonparametric regression estimation. unpublish manuscript.

Nadaraya, E. A. (1964). On estimating regression. Theory Probab. Appl. 9, 141-142.

Watson, G. S. (1964). Smooth regression analysis. Sankhya Ser. A 26, 359-372.