11/22/85

How far are automatically chosen regression smoothing parameters from their optimum?

by

Wolfgang Härdle[1,2]
Johann Wolfgang Goethe Universität, Frankfurt

Peter Hall[1]
Australian National University

J. S. Marron[3]
University of North Carolina, Chapel Hill

Short title: Automatic regression smoothing

Key words and phrases: curve estimation, kernel regression, smoothing parameter, bandwidth selection.

1 Research supported by AFOSR Grant No. [illegible]-49620-82-C-0144.

2 Research supported by Deutche Forschungsgemeinschaft, SFB 123, "Stochastische Mathematische Modelle".

3 Research supported by NSF Grant DMS-8400602.

ABSTRACT

In the setting of nonparametric curve estimation the problem of smoothing parameter selection is addressed. The deviation between the the squared error optimal smoothing parameter and the smoothing parameters provided by a number of automatic selection methods is studied both theoretically and by simulation. The theoretical results include a central limit theorem which shows both the rate of convergence and the asymptotic distribution of the deviation. The simulations show that the asymptotic normality describes the distribution quite well for surprisingly small samples.

1. Introduction

Regression smoothing is a method for recovering the mean function from noisy data, $Y_1,\ldots,Y_n$ of the form,

$$Y_i = m(x_i) + \epsilon_i, \quad i = 1,\ldots,n,$$

where the ϵ_i are independent, identically distributed, mean zero observation errors. There are a number of methods for estimating the regression function, m, which are closely related to moving averages, i.e. to estimate m(x), average the Y_i which have X_i close to x. The width of the neighborhood over which averaging is performed, often called the bandwidth or smoothing parameter, controls the smoothness of the resulting estimate. Figure 1 shows a curve m(x), the solid line, to which simulated noise has been added, together with three weighted moving averages, given in dashed lines, with the smaller bandwidth corresponding to the smaller dashes. More details concerning Figure 1 may be found in Section 4.

[Put Figure 1 here]

It is apparent from Figure 1 that choice of the bandwidth is very important to this type of estimation. In this paper, several automated, i.e. data driven, smoothing parameter (bandwidth) selectors are considered and the amount of noise inherent to them is studied.

Proposed methods for choosing the bandwidth (window size) are based on estimates of the prediction error. For instance, the cross-validation technique provides estimates of the prediction error based on so called "leave one out" estimators of the regression function, see Clark (1975) and Härdle and Marron (1985a). A number of

other selectors are based on adjustments of the residual sum of squares which yield an unbiased estimate of the prediction error, see Craven and Wahba (1979), Rice (1984), and Härdle and Marron (1985b).

The intent of these selectors is to provide an approximate optimization of a quadratic measure of deviation. Call the "optimal bandwidth" the true minimizer of the average square error. How far is the automatically chosen bandwidth from the optimum? This question is answered by showing that the relative differences have an asymptotically normal distribution. Simulations seem to indicate that different selectors behave quite differently in approximating the optimal bandwidth, see Rice (1984), Table 1. We felt these different performances should be reflected in the asymptotic variances of these limiting distributions, and were surprised to find that each selector gave the same asymptotic variance. We reconfirm our theory by simulations which include those of Rice and demonstrate that the dramatic differences he observed seem to be due to his choice of a very small error standard deviation.

A further consequence of the theoretical results is that the order of magnitude of the difference of the optimal bandwidth and that provided by these selectors is quite large. This should not be too disappointing because this difference is of the same order as the difference between the minimizers of the average square error and the mean average square error. Hence, each of these selectors does as well as can be expected.

Section 2 contains the theoretical results. Remarks pertinent to the theoretical results are in section 3. The simulation results are in section 4. Some concluding remarks are in Section 5. The proofs are in the appendix.

2. Behavior of data-driven bandwidths

To simplify the presentation, assume the design points are equally spaced on the unit interval, i.e. $x_i = i/n$, $i = 1,\ldots,n$, and assume that the ϵ_i have common variance, σ^2. The kernel estimator proposed by Priestley and Chao (1972) is, in this setting,

$$\hat{m}_h(x) = n^{-1}h^{-1} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) Y_i,$$

where h is the bandwidth. The kernel, K, is taken here to be a symmetric, compactly supported probability density with (roughly, see the appendix for a precise formulation) a second derivative.

The optimal bandwidth, $\hat{h}_0$, is the minimizer of the average square error (ASE),

$$d_A(h) = n^{-1} \sum_{i=1}^{n} [\hat{m}_h(x_i) - m(x_i)]^2 w(x_i).$$

The weight function, w, is introduced to allow elimination of boundary effects, see Gasser and Müller (1979), by taking w to be supported on a subinterval of the unit interval. If one does not object to assuming that m is circular, i.e. m and its first two derivatives agree at the endpoints 0 and 1, then w may be taken to be identically one. Another candidate for the optimal bandwidth is h_0, the minimizer of the mean average square error (MASE),

$$d_M(h) = E[d_A(h)].$$

We call $\hat{h}_0$ the optimal bandwidth because it makes $\hat{m}_h$ as close as possible to m for the data set at hand, instead of for the average over all possible data sets.

How fast may $\hat{h}_0$ and h_0 be expected to tend to 0? If m'' is uniformly continuous, then under the assumption that the moments of the ϵ_i exist, $d_A(h)$ and $d_M(h)$ are both approximately

$$d_M^*(h) = n^{-1}h^{-1}\sigma^2 \int w \int K^2 + h^4 \left(\frac{\int u^2 K}{2}\right)^2 \int (m'')^2,$$

in the sense that

$$\sup_{h \in H_n} \left(\left| \frac{d_A(h) - d_M^*(h)}{d_M^*(h)} \right| + \left| \frac{d_M(h) - d_M^*(h)}{d_M^*(h)} \right| \right) \to 0, \tag{2.1}$$

in probability as $n \to \infty$, where $H_n = [n^{-1+\delta}, n]$, for arbitrarily small $\delta > 0$. A consequence of (2.1) is that $\hat{h}_0$ and h_0 are each roughly equal to the unique minimizer of d_M^*, $h_0^* = C_0 n^{-1/5}$, where

$$C_0 = \left(\frac{\sigma^2 (\int w)(\int K^2)}{(\int u^2 K)^2 \int (m'')^2} \right)^{1/5}, \tag{2.2}$$

that is

$$\frac{\hat{h}_0}{h_0^*}, \frac{h_0}{h_0^*} \to 1, \tag{2.3}$$

in probability. A sketch of the proof of (2.1) and (2.3) is given in the appendix.

Most bandwidth selectors are based on minimization of some function of h which is related to the residual sum of squares,

$$p(h) = n^{-1} \sum_{i=1}^{n} [Y_i - \hat{m}_h(x_i)]^2 w(x_i).$$

By taking expectations, it can be seen that, as an estimator of the prediction error, p(h) is biased in such a way that its minimizer will not have desirable properties of the type described in (2.3), which is not surprising in view of the fact that p(h) is using the same set of data both to construct an estimate and to assess it. This is corrected

by multiplying p(h) by a factor $\Xi(h^{-1})$, which may be random or nonrandom. Simple examples include:

(i) Generalized Cross-validation (Craven and Wahba 1979),
$$\Xi(h^{-1}) = (1 - n^{-1}h^{-1}K(0))^{-2}.$$
(ii) Akaike's Information Criterion (Akaike 1970),
$$\Xi(h^{-1}) = \exp(2n^{-1}h^{-1}K(0)).$$
(iii) Finite Prediciton Error (Akaike 1974),
$$\Xi(h^{-1}) = (1 + n^{-1}h^{-1}K(0))/(1 - n^{-1}h^{-1}K(0)).$$
(iv) A model selector of Shibata(1981),
$$\Xi(h^{-1}) = 1 + 2n^{-1}h^{-1}K(0)$$
(v) The bandwidth selector T of Rice(1984),
$$\Xi(h^{-1}) = (1 - 2n^{-1}h^{-1}K(0))^{-1}$$

Note that each of the above has a Taylor expansion, in the variable h^{-1}, of the form

$$\Xi(h^{-1}) = 1 + 2n^{-1}h^{-1}K(0) + O(n^{-2}h^{-2}). \tag{2.4}$$

So it makes sense to define a general bandwidth selector,

$$G(h) = p(h)\cdot\Xi(h^{-1}),$$

where $\Xi(h^{-1})$ is of the form (2.4). Other bandwidth selectors are also of essentially this form but it takes more work to see this. An important example is the Cross-Validation function introduced by Clark (1975) (in this setting):

$$CV(h) = n^{-1}\sum_{i=1}^{n}[Y_i - \hat{m}_i(x_i)]^2 w(x_i),$$

where $\hat{m}_i(x_i)$ is a "leave one out" version of $\hat{m}$, i.e. the observation (x_i,Y_i) is left out in constructing $\hat{m}_i$. Priestley and Chao(1972) give a method of adapting $\hat{m}$ when the x_i are not equally spaced. It will be shown in the appendix that

$$CV(h)/p(h) = 1 + 2n^{-1}h^{-1}K(0) + O_p(n^{-2}h^{-2}), \tag{2.5}$$

uniformly over $h \in H_n$. Hence, CV(h) can also be thought of as being a special case of G(h). This last statement can easily be shown, by essentially the same method, to hold also for bandwidth selectors based on unbiased risk estimation, such as

$$\tilde{R}(h) = n^{-1} \sum_{i=1}^{n} \{[Y_i - \hat{m}_h(x_i)]^2 + n^{-1}h^{-1}K(0)[Y_i - Y_{i-1}]^2\}w(x_i),$$

see Rice (1984).

In view of the above asymptotic equivalence of these bandwidth selectors one would expect their performances to be about the same, at least for large n. Indeed it can be shown that the minimizers of all of these (let $\hat{h}$ denote a generic one) are asymptotically optimal, ie: the ratio of loss to minimum loss tends to one,

$$\frac{d_A(\hat{h})}{d_A(\hat{h}_0)} \to 1, \tag{2.6}$$

in probability, or (nearly equivalently)

$$\hat{h}/\hat{h}_0 \to 1, \tag{2.7}$$

in probability (see Rice (1984) and Härdle and Marron (1985a)). A major objective of this paper is to study how fast the convergence in (2.6) and (2.7) occurs, with a view towards trying to distinguish the various bandwidth selectors.

The first part of this is accomplished by

Theorem 1: Under the above assumptions,

$$n^{3/10}(\hat{h} - \hat{h}_0) \to N(0,\sigma_1^2), \tag{2.8}$$

$$n[d_A(\hat{h}) - d_A(\hat{h}_0)] \to C_1 \cdot \chi^2_1,$$

in distribution, where σ_1^2 and C_1 are defined in the appendix, and are seen there to be independent of the particular choice of $\hat{h}$.

Note that by (2.2), (2.3) and (2.7), all of $\hat{h}$, $\hat{h}_0$, h_0, and h_0^* are tending to 0 at the rate $n^{-1/5}$. Hence (2.8) is saying that the relative difference between $\hat{h}$ and $\hat{h}_0$ is of the very slow order $n^{-1/10}$. Although this rate seems at first glance to be excruciatingly slow, it should not be too disappointing, because it is of the same order as the difference between $\hat{h}_0$ and h_0, as demonstrated by:

Theorem 2: Under the above assumptions,

$$n^{3/10}(\hat{h}_0 - h_0) \to N(0,\sigma_2^2),$$

$$n[d_A(\hat{h}_0) - d_A(h_0)] \to C_2 \cdot \chi^2_1,$$

in distribution, where σ_2^2 and C_2 are defined in the appendix.

3. Dicussion and Remarks

Remark 3.1: An important consequence of Theorems 1 and 2 is that they imply that the "plug-in" method of choosing h (where one substitutes estimates of the unknown parts of (2.2)), even if one knew exactly the unknowns σ^2 and $\int(m'')^2$, has an algebraic rate of convergence no better than that of the $\hat{h}$'s given above. Hence the additional noise involved in estimating these unknown parts in practice, especially the second derivative part in the case where m is not very smooth, seems to cast considerable doubt on the applicability of the plug-in estimator. A further advantage of the methods of bandwidth selection proposed in this paper is that they automatically adapt to the case $m''(x) \equiv 0$.

Remark 3.2: Since the bandwidths $\hat{h}$ converge to the optimum $\hat{h}_0$ at the same algebraic rate as h_0, it is natural to compare them by studying the asymptotic variances, σ_1^2 and σ_2^2. By comparing σ_3^2 and σ_4^2 in Lemma 4 below using the Parseval Identity, we see $\sigma_2^2 \leq \sigma_1^2$, so h_0 is closer to $\hat{h}_0$ than $\hat{h}$ is in terms of asymptotic variance. However the limit theorems 1 and 2 can be put together to give a joint limit the rem. Hence

$$\liminf_{n\to\infty} P[d_A(h_0) > d_A(\hat{h}_0)] > 0,$$

i.e. for some data sets, $\hat{h}$ will perform better than h_0. Another consequence of the joint limit theorem is that the bandwdith parts of Theorems 1 and 2 can be added to give Theorem 2.3 of Rice (1984).

Remark 3.3: When we did these calculations we were surprised to note that the asymptotic variance σ_1^2 is independent of the particular

function $E(h)$, especially in view of the simulations of Rice(1984). It is seen in the next section that the phenomena observed by Rice is mostly caused by his particular setting, and often these selectors are not so different.

Remark 3.4: A technical advantage of Theorems 1 and 2 over previous results of this type, see Rice(1984) and Hall and Marron(1985a) is that the range of bandwidths under consideration has been extended from $[an^{-1/5}, bn^{-1/5}]$ to $[n^{-1+\delta}, n]$. This range is reasonble because $h \approx n^{-1}$ corresponds to no smoothing at all, and $h \approx 1$ corresponds to averaging over the entire sample.

Remark 3.5: Several extensions of Theorems 1 and 2 are very straight forward. These include:

(a) The assumption that the errors are identically distributed can be relaxed to assuming that ϵ_i has variance $V(x_i)$, where the function V is uniformly continuous. The only change in the results is in the constants, for example, in C_0 the expression $\sigma^2 \int w$ is replaced by $\int Vw$. Similar replacements are easily calculated for σ_1^2, σ_2^2, and the other expressions given in the appendix.

(b) The design points x_i need not be univariate. In the multivariate case where the x_i have dimension p, the exponents of convergence in the first parts of Theorems 1 and 2 change from 3/10 to $(p+2)/[2(p+4)]$, while the second parts remain the same except for the values of C_1 and C_2. The changes in the constants are easily calculated.

(c) The kernel K can be allowed to take on negative values to

exploit the well known higher rates of convergence possible in that case. In particular, if K is of order k, ie.

$$\int K = 1, \quad \int xK = \ldots = \int x^{k-1}K = 0, \quad \int x^{k}K > 0,$$

and if m has a uniformly continuous k-th derivative, then the exponents of convergence in the first parts of Theorems 1 and 2 change from 3/10 to 3/2(2k+1), while again the second parts are essentially unchanged, and again the new constants are easily calculated.

(d) The Priestly-Chao $\hat{m}$ may be replaced by a number of other kernel type estimators, including those of Nadaraya (1964), Watson (1964), and Gasser and Müller (1979).

Remark 3.6: Note that by estimating the unknown parts in the expressions in the appendix for σ_1^2, Theorem 1 can be used to provide approximate confidence intervals for $\hat{h}_0$, which can be useful for suggesting a reasonable range of bandwidths to consider for choosing the smoothing parameter by an interactive trial and error approach. Of course the comments in remark 3.1 serve to put some substantial limitations on this approach.

Remark 3.7: It is conjectured that the slow relative rate of convergence of $\hat{h}$ to $\hat{h}_0$ that was observed in Section 2 is in fact the best possible in the minimax sense. This is made precise, in the related density estimation setting, by Hall and Marron (1985b). The implication of this is that while all of the procedures given in this paper give a slow rate of convergence, there is no point in searching for a procedure which gives a faster rate.

4. Simulations

Following section 4 of Rice(1984), we generated 100 samples of n = 75 psuedo-random normal variables, ϵ_i, with mean 0 and standard devation $\sigma = 0.0015$. These were then added to the regression curve $m(x) = x^3(1-x)^3$, which has the nice effect of allowing a circular design (i.e. when estimating near i = 1, for $i \leq 0$, let $x_i = x_{75-i}$, and similarly at the other end) to eliminate boundary effects. The kernel function was taken to be

$$K(x) = (15/8)(1 - 4x^2)^2 1_{[-0.5,0.5]}(x).$$

Table 1 contains the results when the selectors introduced in Section 2 were used to find $\hat{h}$. The entries show the number of times out of 100 that either the ratio of Mean Average Square Errors,

$$d_M(\hat{h})/d_M(\hat{h}_0),$$

or the ratio of Average Square Errors,

$$d_A(\hat{h})/d_A(h_0),$$

exceeded the value of the column head.

[put Table 1 here]

The rows indexed "Rice" are copied from the study of Rice(1984), who only worked with d_M, and are included to provide some assurance against programming errors and to allow some understanding of how things change when one works with a different data set. The rows indexed "MASE" shows our reproduction of Rice's simulations. Note that they correspond about as one might expect, except for the selectors AIC, FPE, S. The reason these are much different is that these selectors all have a trivial minimum at $h = n^{-1}K(0) = 0.025$, the "no smoothing" point

where $\hat{m}(x_i) = Y_i$. We believe this was not a disaster in Rice's study because the iterative minimizer that he used would typically go to a local minimimum which provided a reasonable bandwidth estimate (except when the curve G(h) had no local minima, which occured, in our study, about the same number of times as Rice reported a ratio exceeding 8). We could not duplicate this because we used a grid search minimizer (motivated by concern over local minima, which we did indeed discover), and this always gave the no smoothing point as the minimum. To make these selectors work at all well, we minimized them over only the interval $h \in [0.1,1]$, where this interval was chosen by examining the functions of h and asking "what range will make them work well on the average?" Of course this can not be done in practice, but we feel it is instructive to see what happens when we give these selectors their best chance. This local minima problem may also contribute to the difference between Rice's results and ours for the other selectors as well.

The rows indexed "ASE" shows how the same set of selected bandwidths performed when the d_A ratio is use instead of the d_M ratio. Note that while the same rough ordering of selectors is preserved, the relative differences are much less.

We feel that the main reason the simulation results of Rice (1984) showed such a big difference in the performance of these selectors is that σ was chosen to be only 0.0015. When this is plugged into asymptotic formulae such as d_M^*, it indicates that this setting requires a very small amount of smoothing, i.e. reasonable h's tend to be quite small. Thus the setting chosen by Rice is one where the asymptotics of

the theory presented in this paper take a rather long time to take effect. To see if they take effect for reasonable sample sizes in settings which are not slanted toward undersmoothing, we repeated the above study with $\sigma = 0.011$ (chosen because it makes the minimizer of d_M^* roughly 1/2). The results are contained in Table 2, whose format is similar to that of Table 1.

[put Table 2 here]

While the same general ordering of selectors that was observed by Rice still holds up here (except that both here and in Table 1, FPE and AIC have traded places, and in Table 2 GCV seems slightly better than $\tilde{R}$), we feel our asymptotic result that the performance of these selectors is roughly the same holds up quite well in the present setting.

Figure 1 gives an indication of what our results mean in terms of the actual curves, for one of our 100 data sets. The curve made of short dashes is $\hat{m}_h(x)$ with h = .26, the minimizer of S for that data set. The curve made of medium dashes is $\hat{m}_h(x)$ with h = .39, the minimizer of ASE. The curve made of long dashes is $\hat{m}_h(x)$ with h = .66, the minimizer of the other automatic selectors. This particular dataset was chosen because, while it was far from the worst, most of the other data sets gave better performance of the automatic selectors.

To investigate just how well our Central Limit Theorems were describing the situation in the finite sample case, Epanechnikov kernel density estimates, see Rosenblatt(1971), with bandwidth chosen by the cross-validation method of Rudemo(1982) and Bowman(1984), were

constructed based on the samples from the distributions of (i) $\hat{h}_0 - h_0$, (ii) $\hat{h}_T - \hat{h}_0$, (iii) $\hat{h}_T - h_0$, where $\hat{h}_T$ is the minimizer of Rice's T(h). Figures 2, 3, and 4 show these curves as solid lines with the dashed lines showing a parametric normal fit for each of these sets of observations.

[put Figures 2, 3, and 4 here]

Figures 2 and 3 demonstrate that even for only n = 75, the asymptotic normality of Theorems 1 and 2 hold to what we feel is a remarkable degree. Furthermore these also do a good job of illustrating the departure of the distribution of these data from normality, in particular there is a skewness in the direction of a slightly heavy right tail (the height of the peak is also lower than the parametric fit, which is to be expected from a kernel density estimate).

Figure 4 is remarkable both for the shape of the left side and because it is actually taller than the mode of the parametric normal fit, indicating substantial lepto-kurtosis. This shows that the normal asymptotic distribution which can be computed for $\hat{h}_T - h_0$ takes a much larger sample size to realistically describe what is happening. In view of the computations in the appendix, this is not so surprising because the terms that drive the limiting distrbutions of $\hat{h}_0 - h_0$ and $\hat{h}_T - \hat{h}_0$ have a simple structure as a sum of uncorrelated martingales, while their sum, $\hat{h}_T - h_0$, has a much more complicated structure.

To allow a more conventional analysis of the departures from normality of the above three distributions, Table 3 summarizes the usual statistics.

[put Table 3 here]

These were computed by the SAS procedure UNIVARIATE. Observe that the pictures of Figures 2, 3, and 4, show quite clearly the computed skewness and kurtosis. The third column gives the Kolmogorov distance to the best Gaussian fit for each difference. The fourth column contains the observed significance of the Kolmogorov test of the hypothesis that the data are indeed normal. Here again the statement made above, from looking at the pictures, that the data sets of Figures 2 and 3 are much closer to normally distributed, is backed up by the computations.

[Put Table 4 here]

Table 4 adds some insight into how the different selectors compare with each other for the data of Table 2. The first two columns contain the sample mean and standard deviation of the bandwidths which minimize the quantity listed at the left side. Note that the mean for the automatically selected bandwidths is nearly a decreasing function of the ordering given in table 2. Also the selectors whose mean matches up best with $\hat{h}_0$ is the rather poorly performing FPE, which is not surprising in view of the comment of Rice that our error criteria d_A and d_M penalize more heavily for h small, or in other words the good performance of the selector T shows up quite well in Table 4 as a bias towards $\hat{h}$ too big. Another interesting feature is that the standard deviation of the selected bandwidths increases as a function of Table 2 performance.

The last 2 columns of Table 4 shows the sample correlation coefficients for the selected bandwidth with $\hat{h}_0$ and, $\hat{h}_{GCV}$, the minimizer of GCV (chosen because it seemed the most representative) respectively. In interpreting these numbers it should be kept in mind that all minima were computed for the same 100 sets of 75 observations. Note that the negative correlations of $\hat{h}_0$ with the $\hat{h}$ cause the ASE values to be much worse than the MASE values in Tables 1 and 2. The high correlation between $\hat{h}_{GCV}$ and each $\hat{h}$ is of course expected because of the similar character of these bandwidth selectors.

5. Conclusions

We believe the most important lesson to be learned from these results is that while automatic smoothing methods contain a good deal of useful information, they are subject to quite a bit of noise. Hence a reasonable procedure seems to be to first choose a bandwidth by a method such as Rice's T and then look at plots of the estimated regression function for that bandwidth and also ones on either side (the confidence intervals described in Remark 3.6 could be useful here).

For the problem of which of these bandwidth selectors one should use, we have a very slight preference for T, but the statement of Rice that "these results are suggestive, but far from conclusive" seems pertinent here as well. One recommendation that clearly can be made is that for kernel regression estimation, FPE, AIC, and S should not be used because of their trivial minimum at the no smoothing point (note that these were designed for model selection and not kernel regression estimation).

Appendix

Proof of Theorems 1 and 2: Assume that K has a Hölder continuous second derivative. The proof of Theorem 2 is based on the expansion

$$(A.1)\qquad 0 = d_A'(\hat{h}_0) = d_M'(\hat{h}_0) + D'(\hat{h}_0) = (\hat{h}_0 - h_0)d_M''(h^*) + D'(\hat{h}_0),$$

where h^* is between $\hat{h}_0$ and h_0, where $D(h) = d_A(h) - d_M(h)$ and where D', d_A', and d_M', denote the derivatives with respect to h of D, d_A, and d_M, respectively. Whereas the proof of Theorem 1 is based on the expansion

$$(A.2)\qquad G(h) = [d_A(h) + \hat{\sigma}^2 + \delta_1(h)][1 + 2n^{-1}h^{-1}K(0) + O_p(n^{-2}h^{-2})],$$

where

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}[m(x_i) - Y_i]^2 w(x_i),$$

$$\delta_1(h) = 2n^{-1}\sum_{i=1}^{n}[\hat{m}_h(x_i) - m(x_i)][m(x_i) - Y_i]w(x_i).$$

Let

$$\delta_2(h) = \delta_1(h) + 2n^{-1}h^{-1}K(0)\hat{\sigma}^2.$$

To analyze the expressions (A.1) and (A.2), we use the following lemmas. Notation used there includes:

$$r_n(h) = n^{-1}h^{-1} + h^4$$

$$L(u) = -uK'(u)$$

$$K_h(u) = h^{-1}K(u/h), \quad L_h(u) = h^{-1}L(u/h),$$

$$b_h(x) = n^{-1}\sum_{i=1}^{n} K_h(x-x_i)m(x_i) - m(x),$$

$$c_h(x) = n^{-1}\sum_{i=1}^{n} L_h(x-x_i)m(x_i) - m(x).$$

Lemma 1: For $\ell = 1,2,\ldots$ there is a constant C_4 so that

$$(A.3)\qquad \sup_{h\in H_n} E\,|\, r_n(h)^{-1}h^{1/2}D'(h)\,|^{2\ell} \le C_4,$$

(A.4) $$\sup_{h\in H_n} E \mid r_n(h)^{-1}h^{1/2}\delta_2'(h) \mid^{2\ell} \leq C_4,$$

furthermore there is an $\eta_1 > 0$ and a constant C_5 so that

(A.5) $$E \mid r_n(h)^{-1}h^{1/2}[D'(h) - D'(h')] \mid^{2\ell} \leq C_5(\frac{|h'-h|}{h})^{\eta_1 \ell},$$

(A.6) $$E \mid r_n(h)^{-1}h^{1/2}[\delta_2'(h) - \delta_2'(h')] \mid^{2\ell} \leq C_5(\frac{|h'-h|}{h})^{\eta_1 \ell},$$

whenever $h,h' \in H_n$, with $h \leq h'$ and $|\frac{h-h'}{h}| \leq 1$.

Lemma 2: For any $\eta_2 \in (0,1/10)$,

(A.7) $$\sup_{h\in H_n} \{r_n(h)^{-1}h^{1/2}[|D'(h)| + |\delta_2'(h)|]\} = o_p(n^{\eta_2}),$$

furthermore if $h_1 n^{1/5}$ tends to a constant, then

(A.8) $$\sup_{|h-h_1|\leq n^{-1/5-\eta_2}} r_n(h)^{-1}h^{1/2}[|D'(h)-D'(h_1)| + |\delta_2'(h)-\delta_2'(h_1)|] = o_p(1).$$

Lemma 3: For some $\epsilon > 0$,

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

Lemma 4:

$$n^{7/10}D'(h_0) \to N(0, \sigma_3^2),$$
$$n^{7/10}\delta_2'(h_0) \to N(0, \sigma_4^2),$$

in distribution, where (letting * denote convolution)

$$\sigma_3^2 = \frac{4}{C_0^3}\sigma^4[\int w^2][\int(K*K-K*L)^2] + 4C_0^2\sigma^2[\int u^2K]^2[\int(m'')^2w^2],$$

$$\sigma_4^2 = \frac{4}{C_0^3}\sigma^4[\int w^2][\int(K-L)^2] + 4C_0^2\sigma^2[\int u^2K]^2[\int(m'')^2w^2].$$

Lemma 5: For any constants a and b,

$$a n^{-1/2} \sup_{\leq h \leq bn^{-1/2}} |D''(h)| = o_p(n^{-2/5}).$$

To finish the proof of the first part of Theorem 2, note first that

$$n^{2/5}d_M''(h^*) \to C_3, \tag{A.9}$$

where

$$C_3 = \frac{2}{C_0^3}\sigma^2[\int K^2][\int w] + 3C_0^2[\int u^2K]^2[\int (m'')^2 w].$$

It follows from (A.1), (A.8), and Lemma 3 that

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

Hence, by Lemma 4,

$$n^{7/10}D'(\hat{h}_0) \to N(0,\sigma_3^2).$$

Thus, applying Lemma 2 and (A.9) to (A.1) gives

$$0 = (\hat{h}_0 - h_0)C_3n^{-2/5} + D'(h) + o_p(n^{-7/10}), \tag{A.10}$$

from which it follows that

$$n^{7/10}(\hat{h}_0 - h_0) \to N(0,\sigma_2^2),$$

where

$$\sigma_2^2 = \sigma_3^2/C_3^2.$$

To prove the second part of Theorem 2, note that by Lemma 5,

$$d_A(h_0) - d_A(\hat{h}_0) = \frac{1}{2}(h_0 - \hat{h}_0)^2 d_A''(h^*) =$$
$$= \frac{1}{2}(h_0 - \hat{h}_0)^2 d_M''(h^*) + o_p(n^{-1}),$$

where h^* is between h_0 and $\hat{h}_0$. Hence,

$$n[d_A(h_0) - d_A(\hat{h}_0)] \to C_2 \cdot \chi^2_1,$$

in distribution, where

$$C_2 = C_3\sigma_2^2/2.$$

The proof of the first part of Theorem 1 takes slightly more work than the proof of the first part of Theorem 2. For $h \in [an^{-1/5}, bn^{-1/5}]$ (where a and b are arbitrary constants), differentiating (A.2) gives

$$0 = G'(\hat{h}) = [d_A'(\hat{h}) + \delta_1'(\hat{h})][1 + O_p(n^{-4/5})] + \tag{A.11}$$

$$+ [d_A(\hat{h}) + \hat{\sigma}^2 + \delta_1(\hat{h})][-2n^{-1}\hat{h}^{-2}K(0) + O_p(n^{-7/5})],$$

which may be written

$$0 = d_A'(h_0) + \delta_2'(h_0) + o_p(n^{-7/10}). \tag{A.12}$$

Working on (A.12) as in (A.1) and (A.10) gives

$$0 = (\hat{h}_0' - h_0)C_3 n^{-2/5} + D'(h_0) + \delta_2'(h_0) + o_p(n^{-7/10}),$$

which after subtracting (A.10) yields

$$\delta_2'(h_0) = (\hat{h} - \hat{h}_0)C_3 n^{-2/5} + o_p(n^{-7/10}).$$

Hence, by Lemma 4,

$$n^{3/10}(\hat{h} - \hat{h}_0) \to N(0, \sigma_1^2),$$

where $\sigma_1^2 = \sigma_4^2/C_3^2$.

The proof of the second part of Theorem 2 is so similar to the above, that only the result is given:

$$n[d_A(\hat{h}) - d_A(\hat{h}_0)] \to C_1 \cdot \chi^2_1,$$

in distribution, where $C_1 = C_3\sigma_2^2/2$.

Proof of Lemma 1: From here on, for notational simplicity we take $w(x) \equiv 1$. Note that $D_1(h) = -(h/2)D'(h)$, can be expanded into

$$D_1(h) = S_1(h) + S_2(h) + S_3(h), \tag{A.13}$$

where

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

$$S_{11}(h) = 2n^{-2} \sum_{i<j} \sum [n^{-1} \sum_{\ell=1}^{n} K_h(x_\ell - x_i)K_h(x_\ell - x_j)]\epsilon_i\epsilon_j,$$

$$S_{12}(h) = n^{-2} \sum_{i<j} \sum \{n^{-1} \sum_{\ell=1}^{n} [K_h(x_\ell - x_i)L_h(x_\ell - x_j) +$$

$$L_h(x_\ell - x_i)K_h(x_\ell - x_j)]\}\epsilon_i\epsilon_j,$$

$$S_{21}(h) = n^{-1} \sum_{i=1}^{n} [n^{-1} \sum_{\ell=1}^{n} K_h(x_\ell - x_i)[2b_h(x_\ell) - c_h(x_\ell)]\epsilon_i,$$

$$S_{22}(h) = n^{-1} \sum_{i=1}^{n} [n^{-1} \sum_{\ell=1}^{n} L_h(x_\ell - x_i)[-b_h(x_\ell)]\epsilon_i,$$

$$S_{31}(h) = n^{-2} \sum_{i=1}^{n} [n^{-1} \sum_{\ell=1}^{n} K_h(x_\ell - x_i)^2](\epsilon_i^2 - \sigma^2),$$

$$S_{32}(h) = n^{-2} \sum_{i=1}^{n} [n^{-1} \sum_{\ell=1}^{n} K_h(x_\ell - x_i)L_h(x_\ell - x_i)](\epsilon_i^2 - \sigma^2),$$

Note that

$$-\delta_2(h)/2 = n^{-1} \sum_{i=1}^{n} \epsilon_i [n^{-1} \sum_{j=1}^{n} K_h(x_i - x_j)\epsilon_j + b_h(x_i) - n^{-1}h^{-1}K(0)\epsilon_i] =$$

$$= 2n^{-2} \sum_{i<j} \sum K_h(x_i - x_j)\epsilon_i\epsilon_j + n^{-1} \sum_{i=1}^{n} b_h(x_i)\epsilon_i,$$

Now, as above write

$$\delta_3(h) = (h/2)\delta_2'(h) = T_1 + T_2, \tag{A.14}$$

where

$$T_1 = 2n^{-2} \sum_{i<j} \sum [K_h(x_i - x_j) - L_h(x_i - x_j)]\epsilon_i\epsilon_j,$$

$$T_2 = n^{-1} \sum_{i=1}^{n} [b_h(x_i) - c_h(x_i)]\epsilon_i,$$

The proof of (A.3) is very similar in spirit to that of (A.5), but is easier so only the proof of (A.5) will be given. First write

$$S_{11}(h) = n^{-2} \sum_{i \neq j} \sum A_{ij}(h)\epsilon_i\epsilon_j,$$

where

$$A_{ij}(h) = n^{-1} \sum_{\ell=1}^{n} K_h(x_\ell - x_i)K_h(x_\ell - x_j).$$

Note that the sum over ℓ ranges only over at most a multiple of nh indices, due to the compactness of support of K. Standard arguments show that,

$$|A_{ij}(h) - A_{ij}(h')| \leq Ch^{-1}\left|\frac{h-h'}{h}\right|,$$

where here and below C denotes a generic constant. By Theorem 2 of

Whittle(1960), for a generic constant C,

$$E\{[r_n(h)^{-1}h^{-1/2}|S_{11}(h) - S_{11}(h')|]^{2\ell}\} \leq$$
$$\leq Cr_n(h)^{-2\ell}h^{-\ell}n^{-4\ell}[\sum_{i\neq j}\sum|A_{ij}(h) - A_{ij}(h')|^2]^{\ell} \leq$$
$$\leq Cr_n(h)^{-2\ell}h^{-\ell}n^{-4\ell}(n^2h'|\frac{h-h'}{h}|^2h^{-2})^{\ell} \leq$$
$$\leq C\,|\frac{h-h'}{h}|^{\eta_1\ell}.$$

By similar bounds on the other terms in the decomposition of $D_1(h)$, it may be seen that

$$E\,|\,r_n^{-1}(h)h^{1/2}[D_1(h) - D_1(h')]\,|^{2\ell} \leq C(\frac{|h'-h|}{h})^{\eta_1\ell},$$

from which (A.5) is an easy consequence. The proofs of (A.4) and (A.6) are the same in spirit and hence are omitted.

Proof of Lemma 2: By Hölder Continuity and compactness of the support of K and L, there is a $\rho > 0$ large enough so that

$$\sup_{|\frac{h-h'}{h}|\leq n^{-\rho}} |D'(h) - D'(h')| = O(n^{-1}).$$

Hence, it is sufficient to restrict the supremum in the statement of Lemma 2 to a set H_n', which is a subset of H_n so that $\#(H_n') \leq n^{\rho+1}$ and so that for any $h \in H_n$ there is an $h' \in H_n'$ with $|\frac{h-h'}{h}| \leq n^{-\rho}$. By Bonferroni's inequality, Whittle's inequality, and (A.3)

$$P[\sup_{h\in H_n'}|r_n(h)^{-1}h^{1/2}n^{-\eta_2}D'(h)| > \epsilon] \leq$$
$$\leq \#(H_n')\cdot\sup_{h\in H_n} E|\epsilon^{-1}r_n^{-1}(h)h^{1/2}n^{-\eta_2}D'(h)|^{2\ell} \leq$$
$$\leq Cn^{\rho+1}(n^{-\eta_2})^{2\ell} \to 0,$$

by taking ℓ sufficiently large, which proves the D part of (A.7). The proofs of the δ_2 part of (A.7) and of (A.8) use the same type of

partitioning argument with (A.4), (A.5), and (A.6) respectively.

Proof of Lemma 3: By (2.3) and (A.7)

$$d_A'(h_0) = d_A'(h_0) - d_A'(\hat{h}_0) = d_M'(h_0) - d_M'(\hat{h}_0) + o_p(n^{-7/10}).$$

But by (A.7),

$$d_A'(h_0) = D'(h_0) + O_p(n^{-7/10+\eta_2}).$$

Thus, setting $\epsilon = -\eta_2+1/10$,

$$d_M'(h_0) - d_M'(\hat{h}_0) = O_p(n^{-3/5-\epsilon}).$$

But

$$d_M'(h_0) - d_M'(\hat{h}_0) = (h_0 - \hat{h}_0)d_M''(h^*),$$

and so, by (A.9),

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

By the same method it can be shown that

$$|\hat{h} - h_0| = O_p(n^{-1/5-\epsilon}).$$

Proof of Lemma 4: This proof is very close to the proofs of Lemmas 3.4 and 3.5 of Hall and Marron (1985a). The major difference is that the variances of the terms in the expansion (A.13) satisfy:

$$n^2h_0\text{var}(S_1(h_0)) \to \sigma^4[\int(K*K - K*L)^2][\int w^2],$$

$$nh_0^{-4}\text{var}(S_2(h_0)) \to \sigma^2[\int u^2K]^2[\int(m''w)^2],$$

$$\text{var}(S_3(h_0)) = O_p(n^{-13/5}),$$

and for (A.14),

$$n^2h_0\text{var}(T_1(h_0)) \to \sigma^4[\int(K - L)^2][\int w^2],$$

$$nh_0^{-4}\text{var}(T_2(h_0)) \to \sigma^2[\int u^2K]^2[\int(m''w)^2].$$

Also a little care must be taken with the martingale structure in the case that w is not identically 1, but this case is handled by writing, for example in the part involving S_{11},

$$\sum_{i\neq j}\sum = \sum_{i<j}\sum + \sum_{i>j}\sum.$$

Proof of (2.1) and (2.3): The proof of (2.1) follows by an argument which is easier than that used in the proof of Lemma 2. A consequence of (2.1) is that

$$\left(\left|\frac{d_M^*(\hat{h}_0) - d_M^*(h_0^*)}{d_M^*(h_0^*)}\right| + \left|\frac{d_M^*(h_0) - d_M^*(h_0^*)}{d_M^*(h_0^*)}\right|\right) \to 0,$$

from which (2.3) follows.

Proof of (2.5): Write

$$CV(h) = p(h) + I + II, \tag{A.15}$$

where

$$I = 2n^{-1}\sum_{i=1}^{n}[Y_i - \hat{m}(x_i)][\hat{m}(x_i) - \hat{m}_i(x_i)]w(x_i),$$

$$II = n^{-1}\sum_{i=1}^{n}[\hat{m}(x_i) - \hat{m}_i(x_i)]^2 w(x_i).$$

But it is straight forward to verify that

$$p(h) = \sigma^2[\int W] + O_p(n^{-1}h^{-1}),$$

$$I = 2n^{-1}h^{-1}K(0)\sigma^2[\int w] + O_p(n^{-2}h^{-2}),$$

$$II = O_p(n^{-2}h^{-2}),$$

uniformly over $h \in H_n$. Hence (A.15) gives (2.5).

References

Akaike, H. (1970), "Statistical predictor information," *Annals of the Institute of Statistical Mathematics*, 22, 203-217.

Akaike, H. (1974), "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, AC19, 719-723.

Bowman, A. (1984), "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, 65, 521-528.

Clark, R. M. (1975), "A calibration curve for radio carbon dates," *Antiquity*, 49, 251-266.

Craven, P. and Wahba, G. (1979), "Smoothing noisy data with spline functions," *Numerische Mathematik*, 31, 377-403.

Gasser, T. and Müller, H.G. (1979), "Kernel estimation of regression functions," in *Smoothing Techniques in Curve Estimation*, *Lecture Notes in Mathematics* 757, 23-68.

Hardle, W. and Marron, J. S. (1985a), "Optimal bandwidth selection in nonparametric regression function estimation," *Annals of Statistics*, 12, to appear.

Hardle, W. and Marron, J. S. (1985b), "Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression," *Biometrika*, 72, 481-484.

Hall, P. and Marron, J. S. (1985a), "Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation," Center for Stochastic Processes Technical Report No. 94.

Hall, P. and Marron, J. S. (1985b), "The amount of noise inherent in bandwidth selection for a kernel density estimator, " Center for Stochastic Processes Technical Report No. 100.

Nadaraya, E. A. (1964), "On estimating regression," *Theory of Probability and its Application*, 9, 141-142.

Priestley, M. B. and Chao, M. T. (1972), "Non-parametric function fitting," *Journal of the Royal Statistical Society*, series B, 34, 385-392.

Rice, J. (1984), "Bandwidth choice for nonparametric regression," *Annals of Statistics*, 12, 1215-1230.

Rosenblatt, M. (1971), "Curve estimates," *Annals of Mathematical Statistics*, 42, 1815-1842.

Rudemo, M. (1982), "Empirical choice of histograms and kernel density estimators," *Scandanavian Journal of Statistics*, 9, 65-78.

Shibata, R. (1981), "An optimal selection of regression variables," *Biometrika*, 68, 45-54.

Watson, G. S. (1964), "Smooth Regression Analysis," *Sankhyā*, series A, 26, 359-372.

Whittle, P. (1960), "Bounds for the moments of linear and quadratic forms in independent variables," *Theory of Probability and Its Applications*, 5, 302-305.

TABLE 1

		1.05	1.1	1.2	1.4	1.6	1.8	2	4	6	8
	Rice	28	17	4	1	0	0	0	0	0	0
T	MASE	30	13	3	1	0	0	0	0	0	0
	ASE	63	51	29	13	10	7	2	0	0	0
	Rice	33	22	7	1	0	0	0	0	0	0
CV	MASE	38	22	3	2	0	0	0	0	0	0
	ASE	73	53	30	14	10	7	3	0	0	0
	Rice	36	21	6	3	1	0	0	0	0	0
R	MASE	32	19	8	6	4	3	3	0	0	0
	ASE	65	52	32	17	12	10	8	2	0	0
	Rice	33	21	8	4	1	1	0	0	0	0
GCV	MASE	34	17	12	7	5	5	4	0	0	0
	ASE	64	50	35	19	14	11	10	2	1	0
	Rice	46	38	28	25	22	21	21	21	18	18
FPE	MASE	40	24	20	11	8	7	0	0	0	0
	ASE	65	51	38	20	16	12	12	2	0	0
	Rice	46	27	18	16	14	13	11	4	4	4
AIC	MASE	40	24	20	14	11	9	0	0	0	0
	ASE	64	51	38	22	17	12	12	2	1	0
	Rice	66	57	50	43	42	42	41	41	19	19
S	MASE	68	63	56	47	45	43	0	0	0	0
	ASE	75	69	62	56	43	32	25	5	1	0

TABLE 2

		1.05	1.1	1.2	1.4	1.6	1.8	2	4	6	8
T	MASE	49	32	20	7	4	3	3	0	0	0
	ASE	70	59	48	33	27	22	16	5	5	4
CV	MASE	49	35	26	6	5	5	5	2	2	0
	ASE	75	64	53	35	29	24	17	6	6	4
GCV	MASE	48	36	24	13	9	7	7	3	2	0
	ASE	75	65	50	37	30	25	19	8	8	6
R	MASE	50	37	28	15	11	11	10	4	0	0
	ASE	73	63	49	36	32	26	21	12	10	8
FPE	MASE	50	39	31	19	16	13	10	2	0	0
	ASE	76	65	49	46	32	27	22	12	9	9
AIC	MASE	50	41	33	20	17	13	10	4	0	0
	ASE	76	65	50	42	33	28	22	12	10	9
S	MASE	63	57	48	37	34	31	30	20	0	0
	ASE	82	70	60	54	48	43	38	23	14	12

TABLE 3

	SKEWNESS	KURTOSIS	D:NORMAL	PROB>D
$\hat{h}_0 - h_0$	0.29018	-0.47242	0.08307	0.088
$\hat{h}_T - \hat{h}_0$	-0.07598	-0.45908	0.04796	>0.15
$\hat{h}_T - h_0$	-0.80269	0.91670	0.11471	<0.01

TABLE 4

$\hat{h}$	$\mu_n(\hat{h})$	$\sigma_n(\hat{h})$	$\rho_n(\hat{h},\hat{h}_0)$	$\rho_n(\hat{h},\hat{h}_{GCV})$
ASE	0.51000	0.10507	1.00000	-0.46602
T	0.56035	0.13845	-0.50654	0.85076
CV	0.57287	0.15411	-0.47494	0.87105
GCV	0.52929	0.16510	-0.46602	1.00000
R	0.52482	0.17852	-0.40540	0.83565
FPE	0.49790	0.17846	-0.45879	0.76829
AIC	0.49379	0.18169	-0.46472	0.76597
S	0.39435	0.21350	-0.21965	0.52915

SAS

FIGURE 1

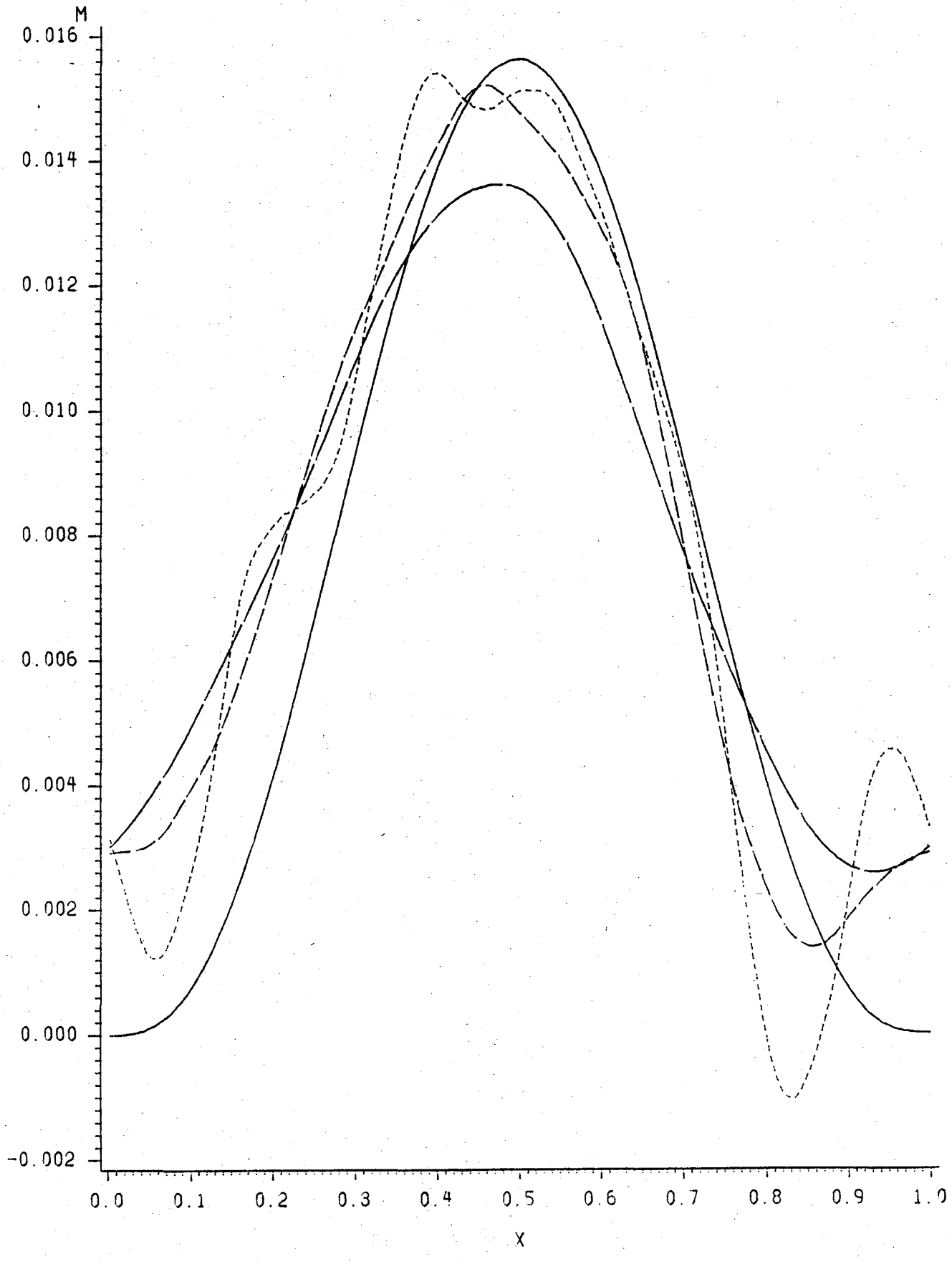
M
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
-0.002
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
X

FIGURE 2

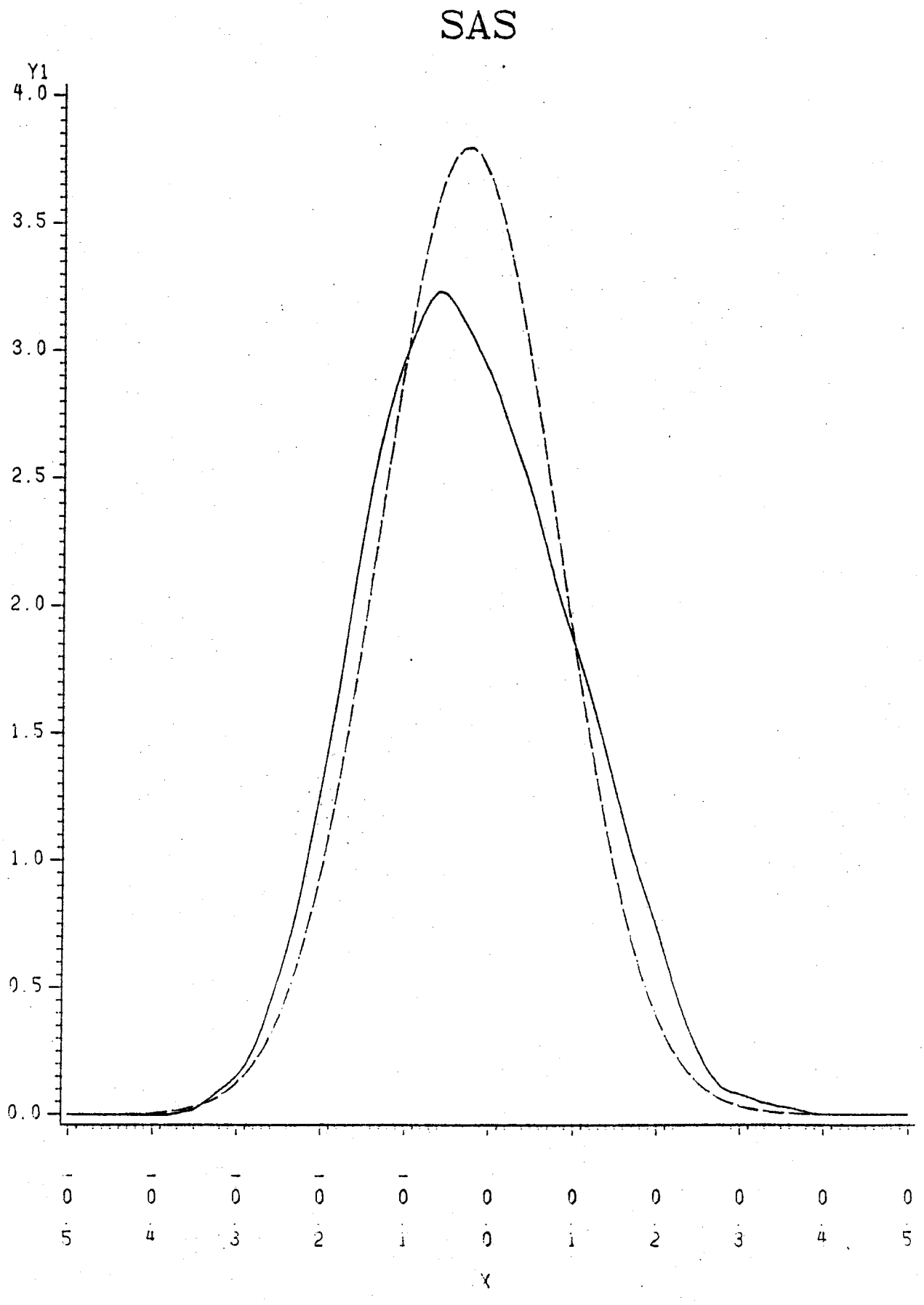

FIGURE 3

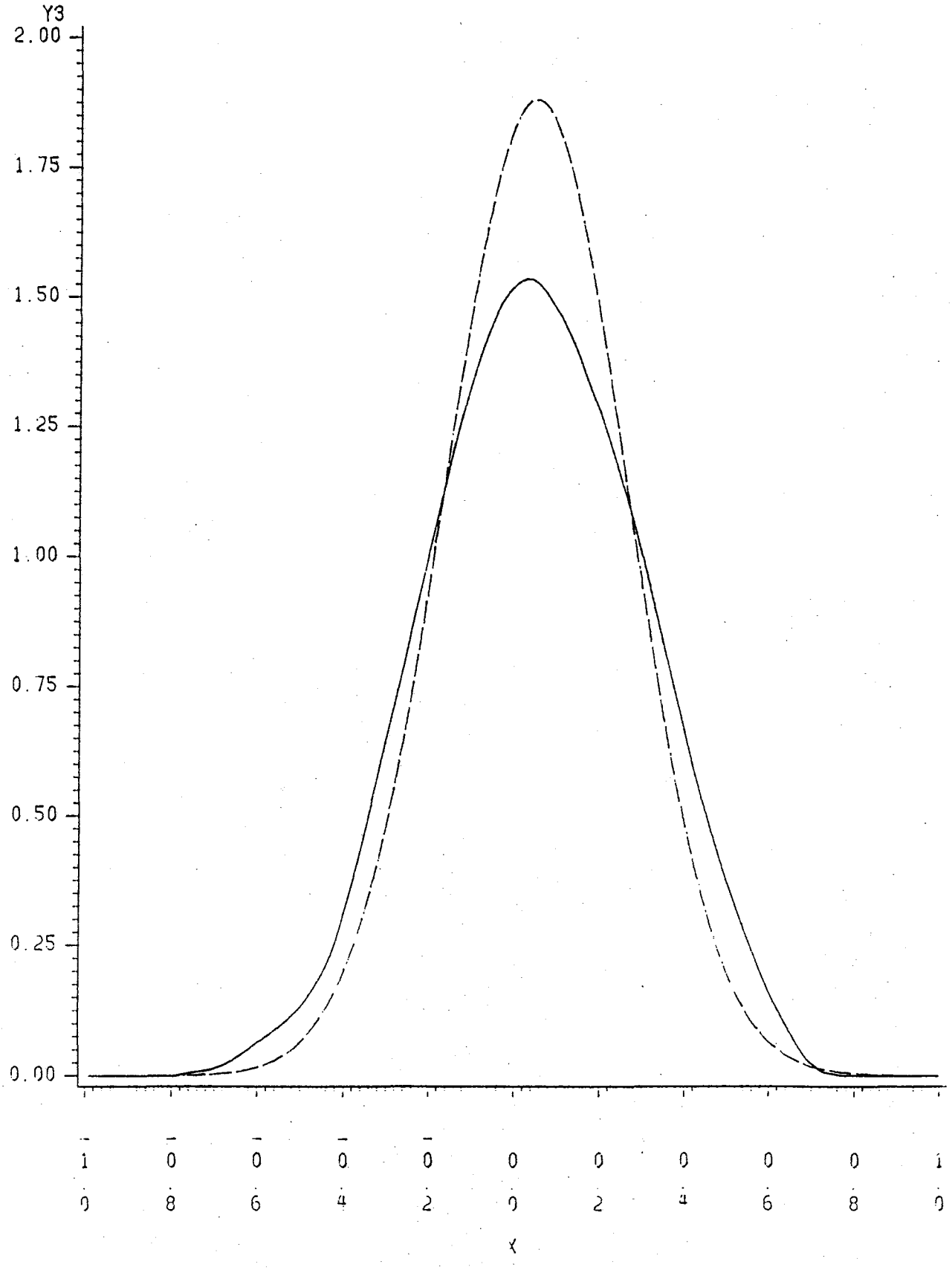

FIGURE 4