Improvement of a data based bandwidth selector

J. S. Marron

Department of Statistics, University of North Carolina,
Chapel Hill, NC 27514 USA

March 24, 1988

ABSTRACT

A recently proposed data based method for choosing the bandwidth of a kernel density estimator is considered. Intuitive and asymptotic reasons are given for why the selected bandwidth should be smaller than is appropriate. This conflicts with the results of a simulation study. The conflict is resolved through a deeper asymptotic analysis. Further simulation results investigate the issue of what sample sizes are required for the asymptotics to properly describe the situation. The analysis is extended to motivate a remedy, which leads to the widely known least-squares cross-validation, hence providing a new characterization of the latter.

Key words: bandwidth selection, cross-validation, kernel density estimation, smoothing parameter

Research partially supported by NSF grant DMS-8701201

1. Introduction

The kernel density estimator (defined mathematically in section 2) provides an effective tool for data analysis, see for example Silverman (1986). The effective performance of this estimator is crucially dependent on the value of the bandwidth or smoothing parameter. If the bandwidth is too small, a large amount of sample variability enters into the estimate, resulting in a curve that is too wiggly. On the other hand if the bandwidth is too large, a lot of bias is introduced in the form of far away observations appearing in the estimator, with the result that important features of the underlying density function can be smoothed away.

This vital dependence on the bandwidth has led to the proposal of a number of methods for using the data to choose a bandwidth. Two of the better known of these are psuedo-likelihood cross-validation and least squares cross-validation. See the survey Marron (1988) for access to a number of others.

Psuedo-likelihood cross-validation, also known as Kullback-Leibler cross-validation, was proposed by Habbema, Hermans and van den Broek (1974) and Duin (1976). It is based on a combination of the cross-validation principle and likelihood ideas. This method will not be considered further here, because it has been demonstrated, by Hall (1988a,b), that the given bandwidth is essentially minimizing the Kullback-Leibler distance between the estimate and the density. This error criterion gives a fit which is not pleasing visually, so the

resulting bandwidth is not appropriate for visual analysis, which is typically the goal of density estimation. Hall's papers provide an explanation for a number of pathologies ascribed to this selection method, see Marron (1985) for access to the literature concerning these pathologies.

Least-squares cross-validation (explicitly defined in section 2), was proposed by Rudemo (1982) and Bowman (1984). The essential idea is to provide a method of estimating the integrated squared error, which is then minimized over the bandwidth. A number of theoretical results have been established concerning this method. Several papers have shown that this method gives an asymptotically correct bandwidth, the best known of these being Stone (1984). Hall and Marron (1987a) have provided an asymptotic quantification of the sample variability of the bandwidth chosen in this manner, in particular showing that the variability will typically be quite large, but see Hall and Marron (1987b) for another side of this issue.

In a recent paper, Kappenman (1987), has proposed an alternative to these methods, which is based on the idea of choosing the bandwidth to make two different estimates of the integrated squared density the same. It is stated that the bandwidth which makes two estimators the same should be a reasonable one. A potential flaw to this reasoning is that estimation of the integrated squared density is a different goal than estimation of the curve itself. That this is a serious issue can be seen from the fact that reasonable bandwidths for the two problems will have different asymptotic orders of magnitude, as discussed in Hall and

Marron (1987c). Section 3 of the present paper shows that this does make a difference, by giving an asymptotic quantification of the type of "bias" introduced by this bandwidth selector. This bias may be viewed as the driving force behind the results of Mielniczuk and Vieu (1988), who have shown that the bandwidth selected by the Kappenman method is asymptotically suboptimal. In view of this intuition and the suboptimality result, the effective small sample performance of this bandwidth selection method, compared to the two types of cross-validation, observed by Kappenman (1987), is really quite surprising.

One goal of the present paper is to provide an explanation of the discrepancy between the above intuitive ideas, and the simulation results. This is done from an asymptotic viewpoint in Section 4, by consideration of the sample variabilities of the selectors involved. Essentially the sample variability of the Kappenman selector is much smaller than that of cross-validation, with the result that even though it is biased, it will still often give superior performance of the estimator, when measured in terms of say Mean Integrated Squared Error.

However, as all of the sample variabilities involved diminish with increasing sample size, while the bias of the Kappenman selector essentially does not, cross-validation will eventually dominate. The sample sizes required for this to happen, as well as the validity of the above conclusions, in several specific settings, are investigated in a further empirical study in Section 5.

An interesting question is whether or not the Kappenman selector

can be modified to overcome its bias problems. A means of doing this is investigated in Section 6. The result, in the case of a Gaussian kernel, is to provide a new derivation of least squares cross-validation.

Sketches of proofs of the asymptotic results are given in section 7.

2. Notation and Bandwidth Selectors

The kernel density estimator, of a density $f(x)$, is given by

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

where $X_1,\ldots,X_n$ is a random sample from f, where $K_h(\cdot) = K(\cdot/h)/h$ for a function K which is often a probability density, and where the amount of smoothness of the resulting curve estimate is determined by the bandwidth, or smoothing parameter, h. Silverman (1986) gives a good discussion of many important applied aspects of this estimator.

The least squares cross-validation method for using the data in selection of h is to take the minimizer, denoted $\hat{h}_{CV}$, of the cross-validation score function,

$$CV(h) = \int \hat{f}_h^2 - 2n^{-1} \sum_{j=1}^{n} \hat{f}_{h,j}(X_j),$$

where $\hat{f}_{h,j}$ denotes the "leave one out" estimator,

$$\hat{f}_{h,j}(x) = (n-1)^{-1} \sum_{i \neq j} K_h(x-X_i).$$

The motivation for the bandwidth $\hat{h}_{CV}$ is provided by the fact that $CV(h)$ is an estimate of a vertical shift of the Integrated Squared Error,

$$ISE(h) = \int (\hat{f}_h - f)^2,$$

whose minimizer will be denoted h_{ISE}.

The bandwidth proposed in Kappenman (1987) is the solution of the equation

$$\int \hat{f}_h^2 = n^{-1} \sum_{j=1}^{n} \hat{f}_{h,j}(X_j), \tag{2.1}$$

which will be denoted here as $\hat{h}_K$. Observe that both sides of this equation provide reasonable estimators of

$$\int f^2 = Ef(X).$$

Kappenman points out that, at least at an empirical level, there appears to be no problem with existence and uniqueness, and this was supported in the present study. It seems that a proof of existence and uniquenesss could probably be obtained for the Gaussian kernel using "total positivity" ideas, see Silverman (1981). Various properties of the two estimators of $\int f^2$ have been described in a number of papers, for example Hall and Marron (1987c).

Widely accepted error criteria (although see Devroye and Gyorfi 1984 for another viewpoint), for studying the performance of $\hat{f}_h$ and of automatically selected bandwidths such as $\hat{h}_{CV}$ and $\hat{h}_K$, include the ISE and its expected value, the Mean Integrated Squared Error,

$$MISE = E(ISE(h)),$$

whose minimizer will be denoted by h_{MISE}. There is some controversy concerning which of h_{ISE} and h_{MISE} should be called the "correct" bandwidth, see Haerdle, Hall and Marron (1988) and the following discussion. The fact that there is, for all reasonable sample sizes, a very considerable difference between these goals has been established by

Hall and Marron (1987a). In this paper, MISE is used as the error criterion, because of its sample stability, and because of some numerical problems with ISE described in Section 5.

A compelling feature of MISE is that it admits, under certain technical assumptions, the asymptotic representation,

$$\text{(2.2)} \qquad \text{AMISE} = n^{-1}h^{-1}\int K^2 + h^4\int (f'')^2 (\int x^2 K)^2/4,$$

as $n \to \infty$, $h \to 0$, with $nh \to \infty$, see for example (3.20) of Silverman (1986). This representation has the attractive feature of very clearly showing how the bandwidth is crucial to the performance of $\hat{f}_h$. In particular note that h small causes a large penalty in the first term, while h large has the same effect on the second term. For this reason AMISE is typically considered to provide a convenient representation of the trade-off that is inherent to the density estimation problem.

However it is important to keep in mind that AMISE is only an approximation to MISE. While this approximation is often fairly good, an example where it is disastrous has been given by Scott (1986). The AMISE $\simeq$ MISE approximation has the very annoying feature that it is quite good for h small, and becomes bad for increasing h, which can create havoc when one considers minimizers. The fact that the approximation is terrible for h large is easily seen from observing that in the limit as $h \to \infty$, MISE $\to \int f^2$, while AMISE tends to infinity at the rate h^4. Figure 1 below demonstrates what this implies for a particular example, discussed in Section 5. Because of this uncertainty, the error criterion used in the simulations of this paper is the actual MISE, and not AMISE.

However, there are still important lessons to be learned from AMISE. In particular, we will soon use the fact that the minimizer of (2.2) is,

$$h_{AMISE} = C(f,K)\, n^{-1/5}, \tag{2.3}$$

where

$$C(f,K) = \left[\frac{\int K^2}{\int (f'')^2 \, (\int x^2 K)^2} \right]^{1/5},$$

In Hall and Marron (1987a) it is seen that, under some mild technical assumptions,

$$h_{MISE} / h_{AMISE} \to 1,$$

and

$$h_{ISE} / h_{AMISE} \to 1,$$

in probability, as $n \to \infty$

3. Bias in Bandwidth Selection

A compelling feature of the cross-validated bandwidth is that it is, at least asymptotically, unbiased. In particular, Hall and Marron (1987a) have demonstrated that, under some reasonable technical assumptions (stated precisely in section 7 of this paper), $\hat{h}_{CV}$ has a limiting normal distribution centered around h_{MISE}.

Unfortunately, the bandwidth $\hat{h}_K$ does not share this property. In particular it is seen in section 7 (under the technical assumptions stated there) that while $\hat{h}_K$ also has a limiting normal distribution, it is centered instead at h_{EK}, the root of the expected value of (2.1). The asymptotic behavior of h_{EK} can be analyzed by use of the expansion,

(3.1) $$E[\int \hat{f}_h^2 - n^{-1}\Sigma_j \hat{f}_{h,j}(X_j)] = n^{-1}h^{-1}\int K^2 - (1/2)h^2\int (f')^2 \int x^2 K,$$

which is derived in section 7. In particular note that h_{EK} has essentially the same asymptotic behavior as

(3.2) $$h_{AK} = C^*(f,K)\, n^{-1/3},$$

where

$$C^*(f,K) = \left[\frac{2\int K^2}{\int (f')^2 \int x^2 K}\right]^{1/3}.$$

A disturbing property of h_{AK} is that it asymptotically diverges from the above goals, in the sense that its ratio with either h_{MISE} or h_{AMISE} tends to zero. In other words, at least asymptotically, h_{AK}, h_{EK}, $\hat{h}_K$ will all provide a choice of the bandwidth which is too small. More precisely, the bandwidth $\hat{h}_K$ essentially tends to zero at the rate $n^{-1/3}$, while the appropriate rate is $n^{-1/5}$.

4. Sample Variability in Bandwidth Selection

Hall and Marron (1987a) have quantified the sample variability in the least squares cross-validated bandwidth by showing, under typical technical assumptions (stated precisely in section 7), that as $n \to \infty$,

(4.1) $$(\hat{h}_{CV} / h_{MISE}) - 1 = O_p(n^{-1/10}).$$

The very slow rate of convergence given here is certainly very disturbing, and goes a long way towards explaining why least squares cross-validation has produced some rather undistinctive results in simulation studies (such as that of Bowman 1984 and Kappenman 1987), and also why it has not become widely accepted in practice.

In section 7, it is shown that, under similar technical assumptions,

$$(4.2) \qquad (\hat{h}_K / h_{EK}) - 1 = O_p(n^{-1/6}).$$

The implication of (4.1) and (4.2) is that the sample variability of the least squares cross-validated bandwidth, $\hat{h}_{CV}$, can be expected to be substantially larger than the Kappenman bandwidth, $\hat{h}_K$. Hence, when the relative effects of these bandwidths are measured by plugging them into MISE, the negative effect of the bias of $\hat{h}_K$ can be swamped by the very large variability of $\hat{h}_{CV}$. The fact that this is exactly what caused the superior performance of $\hat{h}_K$, in the simulations of Kappenman (1987), is demonstrated in the simulation study in the next section.

5. Simulation Results

To investigate the implications of the above intuitive and asymptotic ideas, an empirical study was done. The underlying density functions that were used were:

1. a standard normal distribution,

   N(0,1),

2. a mixture of normals with different means,

   .5N(-1.5,1) + .5N(1.5,1),

3. 2 mixtures of normals with different variances,

   .5N(0,1) + .5N(0,.1),

   .5N(0,1) + .5N(0,.01).

The first two were considered in the study of Kappenman (1987). The advantage of normal mixtures is that they greatly facilitate the exact computation of MISE (recall the discussion in Section 2 above), using an

obvious extension of the ideas of Fryer (1976). For this same reason only the Gaussian kernel function was considered here.

The sample sizes considered were 25, 50, and 100, as considered by Kappenman (1987), and also, because the asymptotics in sections 3 and 4 seem to need larger samples to describe the situation, 400 and 1600, for the first two densities. For the last two, only 25, 100 and 1600 were run. Because of computer time constraints, exact calculations were done only for the sample sizes up to 100. For the larger sample sizes, Fourier Transform approximations, of the type described in section 3.5 of Silverman (1986) were used. The Gaussian kernel is very convenient for this as well. These approximations were checked in the case of $n = 100$ against the exact calculations, and the errors introduced were certainly noticeable, but seemed generally acceptable. One place where the error did not seem acceptable, was in the calculation of the ISE, and that is why ISE results are not presented here. Only the exact ISE calculations are deemed trustworthy, and the lessons there were not substantially different from what can be seen in terms of MISE.

To calculate the minimizers of ISE and the score function $CV(h)$, an exact calculation of their derivatives was set equal to zero. The roots of these equations, together with the roots of (2.1) were calculated by evaluation on a logarithmic grid of 11 h values. The endpoints of the grids were different for the different settings, and chosen to contain essentially all the bandwidths of interest. After evaluation on the grid, a one step Newton-Raphson improvement was done, with the result taken as the selected bandwidth. In the case of

multiple roots, which was an occasional problem for CV(h), the algorithm chose the larger root (this choice was arbitrary). Experimentation indicated that this method of root finding typically gave errors of about 1% to 5% in the reported results, measured with respect to evaluation over a grid of 200 h values.

A less compelling feature of the study in Kappenman (1987) is that rather few Monte Carlo data sets were employed. Of course the number of data sets required for any simulation study depends completely on the context, but it seems that bandwidth selection is a setting where the distributions involved are quite "heavy tailed" in the sense that a large number of Monte Carlo runs are required to get a reasonable amount of accuracy. To see this consider Table 1. This shows, for the sample sizes and densities indicated, the left and right endpoints of standard asymptotic normal theory 95% confidence intervals for the mean of the distribution of ISE evaluated at the indicated bandwidths. The ones corresponding to 100, 50 and 25 Monte Carlo data sets were calculated from the means and standard deviations reported in Table 2 of Kappenman (1987). Note that, especially for the larger values of n, the intervals overlap far too much to allow drawing many conclusions as to how the various bandwidths compare. To attempt to alleviate this problem, 500 Monte Carlo runs were used in the present study. The effect of this is shown in Table 1 in the colmun indexed "NSIM = 500". Note that the lengths of the intervals are now such that meaningful comparison of the estimators can be done. Of course pairwise comparisons should really be done here to make this precise, but this needs more information than is

available, and the essential idea is clear from the univariate intervals.

[Put Table 1 about here]

The choice, by Kappenman (1987), of simply using the value of ISE, evaluated at the bandwidths under consideration, was probably not making the best use of the available information, because of the rather large amount of vertical sample variability of the whole curve ISE(h) (see Hall (1984) for a theoretical quantification of this). This variablility is demonstrated for the current examples in Table 2. Note that a very significant component of the standard deviations of $ISE(\hat{h}_{CV})$ and $ISE(\hat{h}_{K})$ can be attributed to vertical shift of the ISE curves, as described by the standard deviations of the height of the curves at their minima, as given in the first column. This additional noise of course reduces the effectiveness of ISE for the purpose of comparison of bandwidth selectors,

[Put Table 2 about here]

To overcome this problem, in the present study, MISE results are reported, instead of those for ISE. Another reason, for using MISE, is that for the larger sample sizes the Fourier transform approximation was not considered to be sufficiently accurate (even giving negative ISE's in one case!). For meaningful comparison across sample sizes, it seems best to assess the performances of the automatically selected bandwidths by use of the relative error,

$$MISE(\hat{h}) \;/\; MISE(h_{MISE}).$$

To get confidence intervals for the relative MISE's, use the fact that a

suitable normalization of,

(5.1) $$(\mathrm{MISE}(\hat{h}_{CV}) \,/\, \mathrm{MISE}(h_{\mathrm{MISE}})) - 1$$

has a limiting distribution which is $C \cdot \chi^2_1$, for a suitable constant C, as demonstrated by Hall and Marron (1987a). A similar result can be obtained for $\hat{h}_{CV}$ replaced by h_{ISE}. This motivates using the estimated value of C to compare the various bandwidths. Since the χ^2_1 distribution has mean 1, a reasonable estimate of C is $\hat{C}$, the sample average of the quantities (5.1). To obtain confidence intervals for C, use the fact that $\hat{C}$ has an asymptotic (as the number of Monte Carlo simulations grows) Normal$(C, 2C^2)$ distribution to derive the pivoted 95% confidence intervals

(5.2) $$(\hat{C}/(1+T), \hat{C}/(1-T)),$$

where

$$T = 1.96 \cdot (2/\mathrm{NSIM})^{1/2},$$

for NSIM the number of Monte Carlo runs. The limiting chi square distribution does not hold up for $\hat{h}_{CV}$ replaced by $\hat{h}_K$ because of the bias of the latter described in Section 3, however there is still considerable insight to be gained concerning the effectiveness of $\hat{h}_K$ by considering intervals of the type (5.2).

Table 3 contains the main results of the simulation study. For each distribution and sample size, it shows the 95% confidence intervals for the comparison number, C, given in (5.2). This allows comparison of the various bandwidths, and at the same time gives an idea of the Monte Carlo variability involved.

[put Table 3 about here]

Notice that for all 4 distributions, $\hat{h}_K$ is always the best for the small sample sizes, and eventually becomes dominated by $\hat{h}_{CV}$. The reason for this is exactly the explanation given in Sections 3 and 4. In particular, $\hat{h}_K$ has a bias which gets worse for larger sample sizes, while $\hat{h}_{CV}$ suffers from a large amount of sample variance, which for large enough sample sizes, is eventually less debilitating than the bias of $\hat{h}_K$. Note that the point where the tradeoff occurs depend on the particular distribution involved, but except for the standard normal, it seems to happen typically for sample sizes bigger than 100.

The effectiveness of this explanation could be verified by reporting the sample means and variances of these bandwidths, but instead a more visual method of illustrating this point is given in Figure 1. Another benefit to this approach is that additional insight into how the bias and variabilities of the automatically selected bandwidths affect MISE can be obtained.

[Put Figures 1a,b,c,d about here]

Figures 1a,b,c,d are for the sample sizes n = 25, 100, 400, 1600 respectively. All concern the mean mixture distribution, chosen because of the interesting behavior of the different bandwidths observed in Table 3. These figures each show an overlay of the MISE and AMISE curves, together with kernel density estimates for the distribution of the bandwidths $\hat{h}_K$ and $\hat{h}_{CV}$. Because the central limit theory of Hall and Marron (1987a) and section 7 indicates that the distribution of these quantities should be roughly normal, the bandwidths of these kernel estimates was chosen by the oversmoothing method of Terrell and

Scott (1985), and the kernel was taken to be Gaussian. For further comparison purposes, the location of the bandwidths h_{MISE}, h_{AMISE} and h_{AK} are indicated by vertical lines.

The fact that the the smaller sample variability of $\hat{h}_K$ explains it superior Table 3 performance, for $n = 25$ and 100, can now be clearly seen from Figures 1a and b. Similarly Figures 1c and 1d show that the inferior Table 3 performance of $\hat{h}_K$, for $n = 400$ and 1600, is explained by the fact that the centerpoint of its distribution is too small.

Another thing which can be seen from these figures is exactly what the rates of convergence in (4.1) and (4.2) imply in actual estimation settings. These rates show up as the fact that the $\hat{h}_{CV}$ distributions are much more spread out than the $\hat{h}_K$ distributions.

Recall that the vertical lines represent the asymptotic centerpoints for the bandwidth distribution. Figure 1 also shows how well they represent the true centers. As expected the representation improves with larger sample sizes. Another thing that can be seen from the vertical lines is the implications of the bias of $\hat{h}_K$ as asymptotically quantified in (2.3) and (3.2).

Recall that in Section 2, concern was expressed concerning the $AMISE(h) \simeq MISE(h)$ approximation. What this means in the present context can also be seen from Figure 1. The remark, made in Section 2, that the difference is worse for larger h, is clearly seen to be true. As expected, when the sample size increases, the low points on the curves get smaller and move to the left, also h_{MISE} gets closer to

h_{AMISE}. Note also that the changes in the curves for different sample sizes show up mostly on the left side of the curves. This is because the bias (which is the dominant part of the curve for large h values) is independent of the sample size.

One conclusion that follows from these results is that for large sample sizes, the bandwidth $\hat{h}_K$ can not be recommended. While its small sample performance is certainly superior to $\hat{h}_{CV}$ for the examples considered here, it does not seem safe to use in general. In particular, there may be distributions where the bias of $\hat{h}_K$ may cause problems for very small sample sizes. The main lesson to be learned from this study seems to be that there is a lot of room for improvement of $\hat{h}_{CV}$. See Marron (1988) for discussion of some possibilities along this line.

6. Bias Elimination

This section explores a means of removing the bias inherent to the Kappenman bandwidth selector, as quantified in Section 3. The essential idea is to replace (3.2) with an equation which has h_{AMISE} as a root. In section 7, it is seen that, by considering estimates of $\int(f')^2$, this may be accomplished in a practical fashion by replacing the equation (2.1) by

$$(6.1) \quad \int(\hat{f}_h')^2 - n^{-1}h^{-3}(\int(K')^2 - \int K^2/2\int x^2K) = -n^{-1}\sum_{j=1}^{n}\hat{f}_{h,j}''(X_j).$$

A very interesting feature of the equation (6.1) is that, for K standard normal (as in the simulation studies of Kappenman (1987) and of

section 5), it is the same as the equation

(6.2) $$CV'(h) / 2h = 0.$$

This is also verified in section 7. The implication of (6.2) is that solving the equation (6.1) is equivalent to minimizing the least squares cross-validation score, CV(h), so this modification of Kappenman's ideas provides a new way of motivating the least squares cross-validated bandwidth $\hat{h}_{CV}$. See Diggle and Marron (1987) for another characterization of $\hat{h}_{CV}$.

7. Assumptions and Proofs

The technical assumtions that are required for the new results in this paper, as well as the other results that are used are:

1. the underlying density has two Hoelder continuous derivatives.
2. the kernel function is a twice continuously differentiable probability density.
3. All asymptotics are as $n \to \infty$, $h \to 0$, with $nh \to \infty$.

To prove (3.1), allowing $*$ to denote convolution, note that

$$\int \hat{f}_h^2 = n^{-1}(K*K)_h(0) + n^{-2} \sum_{i \neq j} \sum (K*K)_h(X_i - X_j),$$

so

$$E \int \hat{f}_h^2 = n^{-1}h^{-1} \int K^2 + R,$$

where

(7.1)
$$\begin{aligned} R &= \iint (K*K)_h(x-t)\, f(x)\, f(t)\, dx\, dt \\ &= \int [\int K*K(u)\, f(x-hu)\, du]\, f(x)\, dx \\ &= \int [f(x) + h^2 f''(x) (2\int x^2 K) /2 + o(h^2)]\, f(x)\, dx \end{aligned}$$

$$= \int f^2 - h^2 \int (f')^2 \int x^2 K + o(h^2)$$

A similiar but easier calculation gives

$$E\, n^{-1} \sum_{j=1}^{n} \hat{f}_{h,j}(X_j) = \int f^2 - (1/2)h^2 \int (f')^2 \int x^2 K + o(h^2) + o(n^{-1}).$$

Straightforward algebra yields (3.1).

The fact that $\hat{h}_K$ has a limiting normal distribution is established using methods of proof very similar to those used in Hall and Marron (1987a), so only an outline is given here. Note that $\hat{h}_K$ is a root of the equation

$$0 = D(h) = \int \hat{f}_h^{\,2} - n^{-1} \sum_{j=1}^{n} K_{h,j}(X_j).$$

Note that h_{AK} will have essentially the same behavior as h_{EK} which is the root of

$$0 = E\, D(h).$$

The key to the proof is the expansion

$$0 = ED(h_{EK}) = ED(\hat{h}_K) + (h_{EK} - \hat{h}_K)(ED)'(\tilde{h}),$$

where $\tilde{h}$ is between h_{EK} and $\hat{h}_K$. This gives

$$h_{EK} - \hat{h}_K = \frac{-ED(\hat{h}_K)}{(ED)'(\tilde{h})} = \frac{D(\hat{h}_K) - ED(\hat{h}_K)}{(ED)'(\tilde{h})}.$$

The limiting normal distribution stated in section 3, and at (4.2), now follows from

$$n^{5/6}(D(\hat{h}_K) - ED(\hat{h}_K)) \rightarrow N(0,\sigma^2),$$

$$(ED)'(\tilde{h}) = O(n^{-1/3}).$$

The proofs of these may be obtained in a straightforward fashion using the techniques of Hall and Marron (1987a).

To see why (6.1) is a reasonable modification of (2.1), again take expected values, and consider asymptotic approximations. Calculating as

at (7.1) gives

$$E\int(\hat{f}_h')^2 = n^{-1}h^{-3}\int(K')^2 + \int(f')^2 - h^2\int(f'')^2\int x^2K + o(h^2),$$

$$E\ -n^{-1}\sum_{j=1}^{n}\hat{f}_{h,j}''(X_j) = \int(f')^2 - (1/2)h^2\int(f'')^2\int x^2K + o(h^2).$$

It follows from these that the expected value of (6.1) has the asymptotic representation

$$\int(f')^2 - h^2\int(f'')^2\int x^2K + n^{-1}h^{-3}\int K^2/2\int x^2K = \int(f')^2 - (1/2)h^2\int(f'')^2\int x^2K.$$

Note that h_{AMISE} is a root of this equation.

To check that (6.1) is the same as (6.2) in the case of K Gaussian, note first that

$$CV(h) = n^{-1}h^{-1}\int K^2 + n^{-1}\sum_{i\neq j}\sum[n^{-1}(K*K)_h(X_i-X_j) - 2(n-1)^{-1}K_h(X_i-X_j)].$$

For K a standard normal density, recall that $K*K = K_{\sqrt{2}}$, and observe that

$$\frac{d}{dh}K_h(x) = h^{-1}(K'')_h(x).$$

It follows that

$$CV'(h) = -n^{-1}h^{-3}\int K^2 + n^{-1}\sum_{i\neq j}\sum[n^{-1}h^{-1}(K'')_{h\sqrt{2}}(X_i-X_j) - 2(n-1)^{-1}h^{-1}(K'')_h(X_i-X_j)].$$

The fact that (6.1) is equivalent to (6.2) is now a consequence of

$$\int(\hat{f}_h')^2 - n^{-1}h^{-3}(\int(K')^2 = -(1/2)n^{-2}h^{-2}\sum_{i\neq j}\Sigma(K'')_{h\sqrt{2}}(X_i-X_j)$$

and

$$-n^{-1}\sum_{j=1}^{n}\hat{f}_{h,j}''(X_j) = -n^{-1}(n-1)^{-1}h^{-2}\sum_{i\neq j}\Sigma(K'')_h(X_i-X_j).$$

References

Bowman, A. (1984), "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, 71, 353-360.

Diggle, P. and Marron. J. S. (1988) "Equivalence of smoothing parameter selectors in density and intensity estimation", to appear *Journal of the American Statistical Association.*

Duin, R. P. W. (1976), "On the choice of smoothing parameters fo Parzen estimators of probability density functions," *IEEE Trnasactions on Computers*, C-25, 1175-1179.

Fryer, M. J. (1976), "Some errors associated with nonparametric estimation of density functions," *Journal of the Institute of Mathematics and its Applications*, 18, 371-380.

Habbema, J. D. F., Hermans, J. and van den Broek, K. (1984), "A stepwise discrimination analysis program using density estimation," *Compstat 1974: Proceedings in Computational Statistics*, 101-110, Physica Verlag, Vienna.

Härdle, W., Hall, P. and Marron, J. S. (1988), "How far are automatically chosen regression smoothers from their optimum?," to appear with discussion, *Journal of the American Statistical Association.*

Hall, P. (1984a), "Central limit theorem for integrated square error of multivariate nonparametric density estimators," *Journal of Multivariate Analysis* 14, 1-16.

Hall, P. (1988a), "On the estimation of probability densities using compactly supported kernels," unpublished manuscript.

Hall, P. (1988b), "On Kullback-Leibler loss and density estimation," unpublished manuscript.

Hall, P. and Marron, J. S. (1987a), "Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation," *Probability Theory and Related Fields*, 74, 567-581.

Hall, P. and Marron, J. S. (1987b), "On the amount of noise inherent in bandwidth selection for a kernel density estimator, " *Annals of Statistics*, 15, 163-181.

Hall, P. and Marron, J. S. (1987c), "Estimation of integrated squared density derivatives", *Statistics and Probability Letters*, 6, 109-115.

Kappenman, R. F. (1987), "A nonparametric data based univariate function estimate," *Computational Statistics and Data Analysis*, 5, 1-7.

Marron, J. S. (1985), "An asymptotically efficient solution to the bandwdith problem of kernel density estimation," *Annals of Statistics*, 13, 1011-1023.

Marron, J. S. (1988), "Automatic smoothing parameter selection: A survey", North Carolina Institiute of Statistics, Mimeo Series #1746.

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

Scott, D. W. (1986), "Handouts for ASA short course in density estimation," Rice University Technical Report 776-331-86-2.

Silverman, B. W. (1981), "Using density estimates to investigate unimodality", *Journal of the Royal Statistical Society, Series B*, 43, 97-99.

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

Stone, C. J. (1984), "An asymptotically optimal window selection rule for kernel density estimates," *Annals of Statistics*, 12, 1285-1297.

Terrell, G. R. and Scott, D. W. (1985), "Oversmoothed density estimates," *Journal of the American Statistical Association*, 80, 209-214.

Table 1

Endpoints of 95% asymptotic confidence intervals for the mean of the distribution of the Integrated Squared Error, evaluated at the bandwidths h_{AMISE}, $\hat{h}_{CV}$, and $\hat{h}_{K}$. The number of Monte Carlo data sets is represented by NSIM. The sample size of each Monte Carlo data set is given by n.

n = 25		NSIM = 100	NSIM = 500
Standard Normal	AMISE	(0.0114,0.0166)	(0.0113,0.0131)
	CV	(0.0213,0.0371)	(0.0201,0.0239)
	K	(0.0165,0.0255)	(0.0173,0.0205)
Mean Mixture	AMISE	(0.0119,0.0151)	(0.0106,0.0116)
	CV	(0.0155,0.0205)	(0.0168,0.0196)
	K	(0.0123,0.0157)	(0.0141,0.0159)
n = 50		NSIM = 50	NSIM = 500
Standard Normal	AMISE	(0.0065,0.0103)	(0.0072,0.0082)
	CV	(0.0089,0.0141)	(0.0120,0.0144)
	K	(0.0088,0.0132)	(0.0106,0.0122)
Mean Mixture	AMISE	(0.0070,0.0106)	(0.0068,0.0074)
	CV	(0.0097,0.0139)	(0.0103,0.0117)
	K	(0.0078,0.0122)	(0.0084,0.0094)
n = 100		NSIM = 25	NSIM = 500
Standard Normal	AMISE	(0.0028,0.0062)	(0.0045,0.0051)
	CV	(0.0034,0.0062)	(0.0073,0.0085)
	K	(0.0041,0.0067)	(0.0070,0.0080)
Mean Mixture	AMISE	(0.0037,0.0061)	(0.0044,0.0046)
	CV	(0.0043,0.0079)	(0.0065,0.0073)
	K	(0.0039,0.0063)	(0.0054,0.0060)

Table 2

Sample standard deviations for Integrated Squared Errors evaluated at the bandwidths h_{ISE}, h_{AMISE}, $\hat{h}_{CV}$, $\hat{h}_{K}$. From 500 Monte Carlo datasets.

		ISE	AMISE	CV	K
Standard Normal	n = 25	0.0100	0.0107	0.0219	0.0178
	n = 100	0.0037	0.0039	0.0074	0.0055
Mean Mixture	n = 25	0.0061	0.0096	0.0157	0.0105
	n = 100	0.0027	0.0031	0.0048	0.0035
Variance Mixture 1	n = 25	0.0240	0.0273	0.0520	0.0363
	n = 100	0.0094	0.0094	0.0134	0.0101
Variance Mixture 2	n = 25	0.0689	0.0759	0.1335	0.1154
	n = 100	0.0236	0.0237	0.0314	0.0270

Table 3

95% Monte Carlo Confidence intervals for the mean of the distribution of the relative Mean Integrated Squared Errors minus one, see (5.1), for the bandwidths h_{ISE}, h_{AMISE}, $\hat{h}_{CV}$, $\hat{h}_{K}$, h_{AK}. From 500 Monte Carlo datasets.

		Standard Normal		Mean Mixture
n = 25				
	AMISE	(0.01757,0.02254)	AK	(0.01711,0.02195)
	AK	(0.02807,0.03601)	K	(0.06974,0.08947)
	ISE	(0.07514,0.09640)	ISE	(0.07856,0.10079)
	K	(0.17049,0.21873)	AMISE	(0.11363,0.14579)
	CV	(0.29432,0.37761)	CV	(0.25094,0.32196)
n = 50				
	AMISE	(0.01066,0.01368)	AK	(0.01679,0.02155)
	AK	(0.07065,0.09065)	K	(0.05670,0.07274)
	ISE	(0.07309,0.09377)	AMISE	(0.05970,0.07659)
	K	(0.18222,0.23379)	ISE	(0.06203,0.07958)
	CV	(0.24561,0.31511)	CV	(0.23758,0.30482)
n = 100				
	AMISE	(0.00641,0.00822)	AK	(0.02909,0.03733)
	ISE	(0.07130,0.09147)	AMISE	(0.03248,0.04167)
	AK	(0.13136,0.16854)	ISE	(0.05170,0.06633)
	CV	(0.21243,0.27255)	K	(0.06117,0.07848)
	K	(0.24536,0.31480)	CV	(0.21859,0.28045)
n = 400				
	AMISE	(0.00226,0.00290)	AMISE	(0.01013,0.01300)
	CV	(0.10971,0.14076)	CV	(0.07775,0.09975)
	AK	(0.29710,0.38118)	K	(0.09480,0.12163)
	K	(0.34257,0.43952)	AK	(0.09814,0.12592)
n = 1600				
	AMISE	(0.00078,0.00100)	AMISE	(0.00326,0.00418)
	CV	(0.08341,0.10701)	CV	(0.05869,0.07531)
	AK	(0.51087,0.65545)	AK	(0.23043,0.29564)
	K	(0.54836,0.70355)	K	(0.23374,0.29989)

Table 3 cont.

		Variance Mixture 1		Variance Mixture 2
n = 25				
	AK	(0.01212,0.01555)	AK	(0.00528,0.00677)
	AMISE	(0.05855,0.07512)	ISE	(0.03285,0.04215)
	ISE	(0.06637,0.08516)	AMISE	(0.06000,0.07698)
	K	(0.13350,0.17128)	K	(0.23967,0.30750)
	CV	(0.32394,0.41562)	CV	(0.25418,0.32611)
n = 100				
	AMISE	(0.02042,0.02620)	AMISE	(0.02093,0.02686)
	ISE	(0.05186,0.06653)	ISE	(0.02950,0.03785)
	AK	(0.05572,0.07149)	AK	(0.03096,0.03972)
	K	(0.09306,0.11940)	K	(0.05277,0.06771)
	CV	(0.16944,0.21739)	CV	(0.10303,0.13219)
n = 1600				
	AMISE	(0.00237,0.00304)	AMISE	(0.00248,0.00318)
	CV	(0.03481,0.04466)	CV	(0.06231,0.07995)
	K	(0.32047,0.41117)	K	(0.16077,0.20627)
	AK	(0.32063,0.41137)	AK	(0.25205,0.32338)

Caption for Figure

Figure 1: Overlay of MISE(h) and AMISE(h) curves, with kernel density estimates of distributions of $\hat{h}_K$ (smaller variance) and $\hat{h}_{CV}$ (larger variance) distributions. Vertical lines show locations of h_{AK}, h_{MISE}, and h_{AMISE}. Underlying density is .5N(-1.5)+.5N(1.5,1). Bandwidths were selected from 500 Monte Carlo data sets of size: (a) n = 25, (b) n = 100, (c) n = 400, (d) n = 1600.

Figure 1a

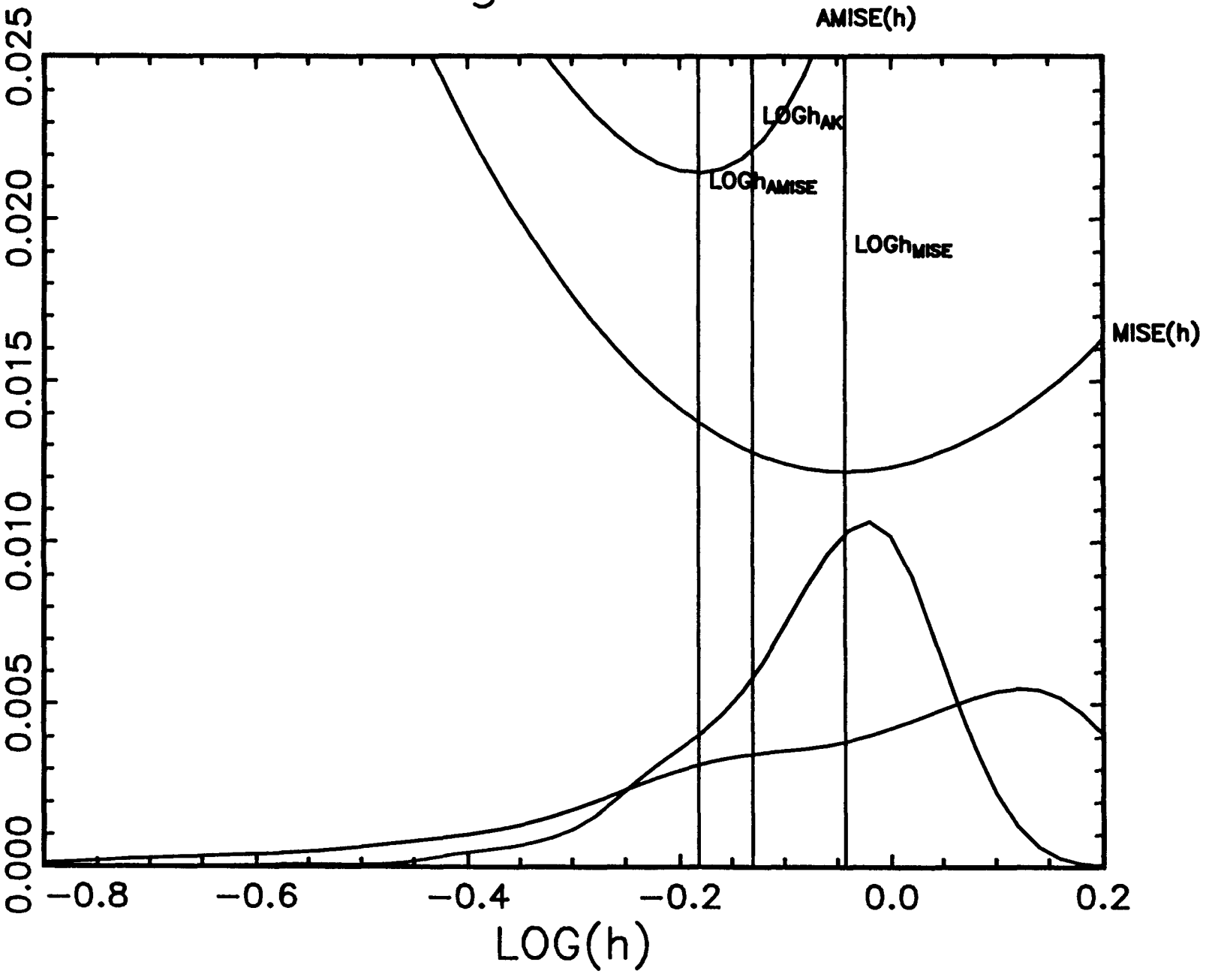

AMISE(h)
LOGh_AK
LOGh_AMISE
LOGh_MISE
MISE(h)
0.025
0.020
0.015
0.010
0.005
0.000
−0.8
−0.6
−0.4
−0.2
0.0
0.2
LOG(h)

Figure 1b

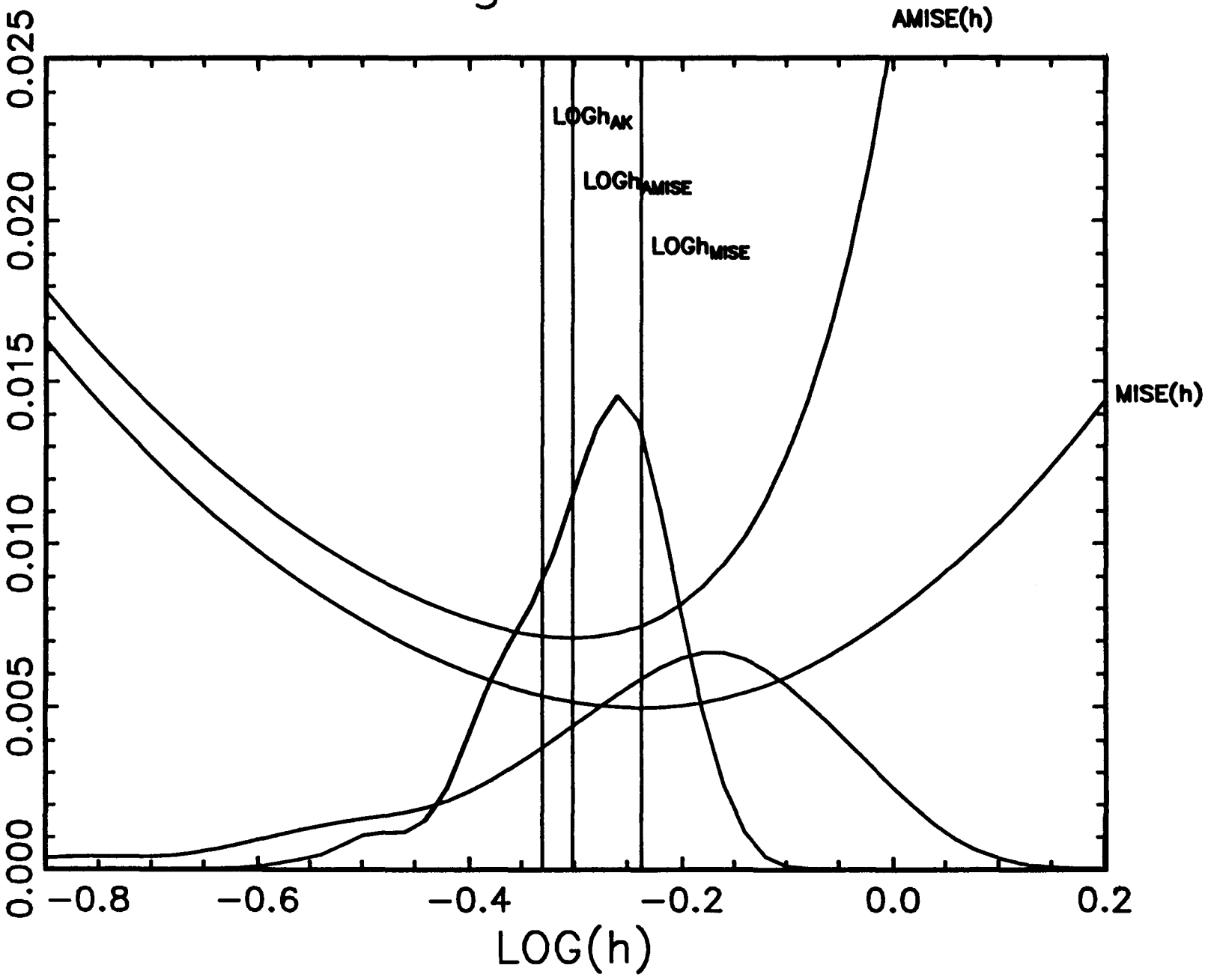

AMISE(h)
MISE(h)
LOGh_AK
LOGh_AMISE
LOGh_MISE
0.025
0.020
0.015
0.010
0.005
0.000
−0.8
−0.6
−0.4
−0.2
0.0
0.2
LOG(h)

Figure 1c

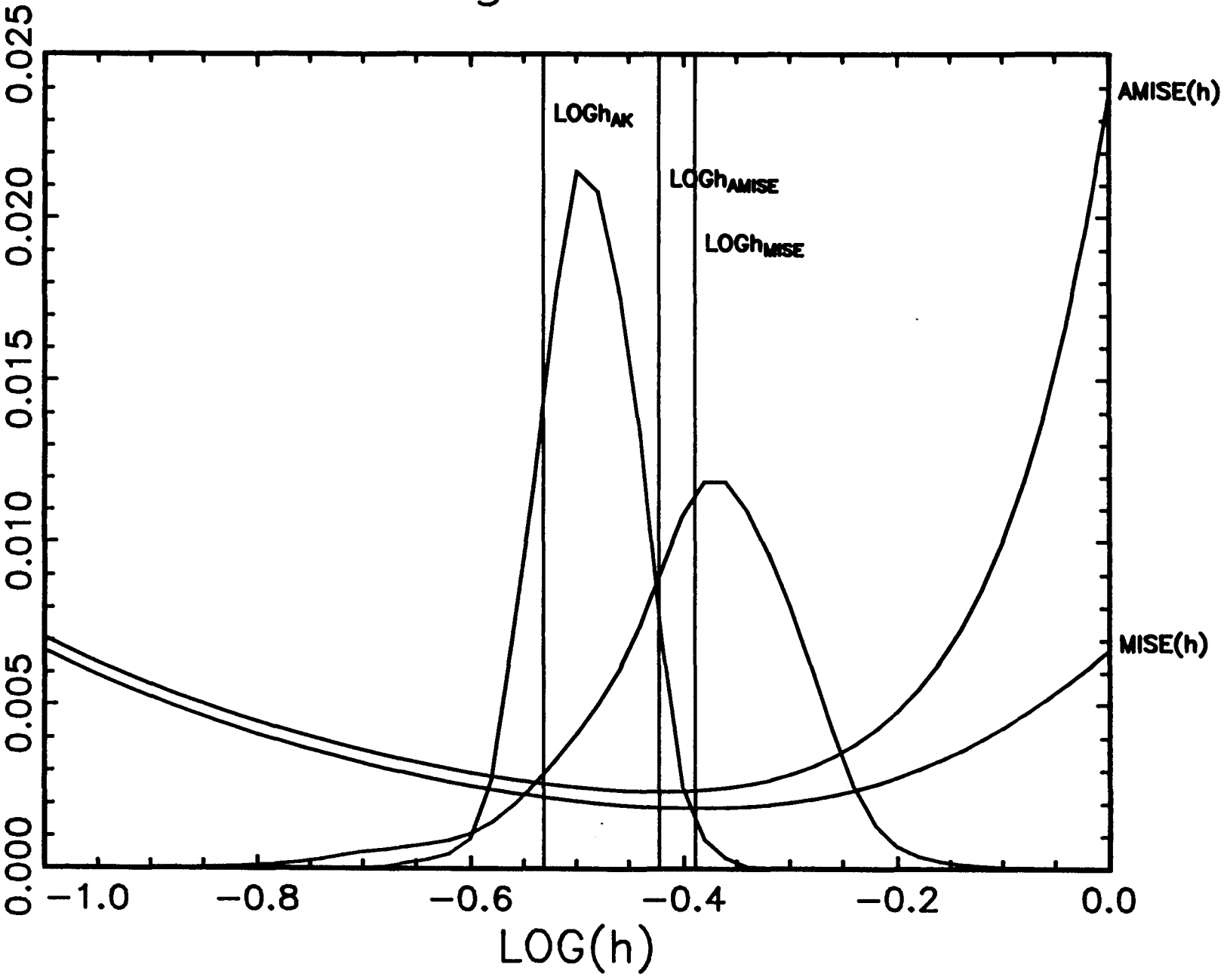
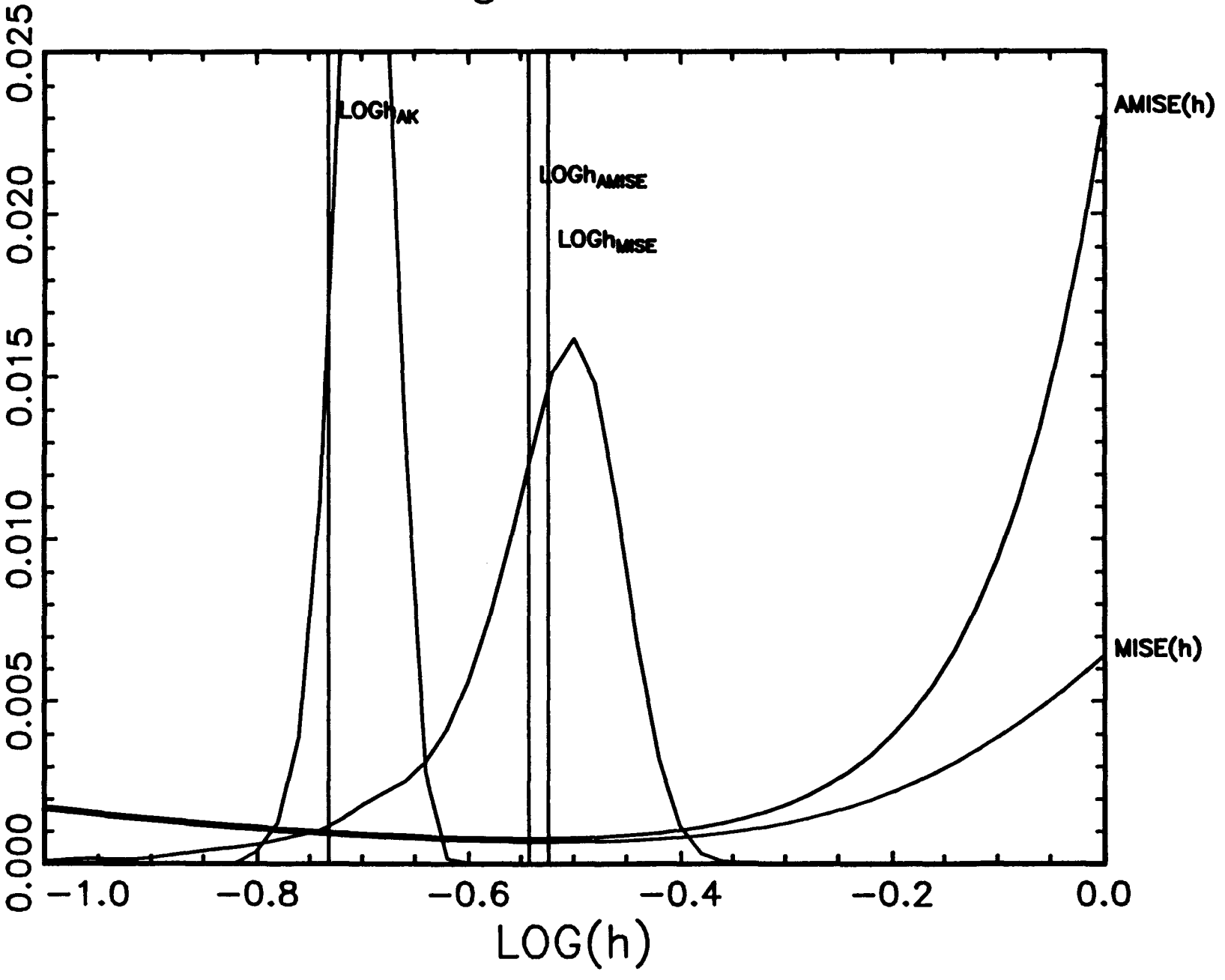

0.025
0.020
0.015
0.010
0.005
0.000
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
LOG(h)
LOGh_AK
LOGh_AMISE
LOGh_MISE
AMISE(h)
MISE(h)

Figure 1d


0.025
0.020
0.015
0.010
0.005
0.000
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
LOG(h)
LOGh_AK
LOGh_AMISE
LOGh_MISE
AMISE(h)
MISE(h)