Bootstrap Bandwidth Selection

by J. S. Marron

University of North Carolina
Chapel Hill, N.C. 27599–3260

May 13, 1990

Abstract: Various bootstrap methodologies are discussed for the selection of the bandwidth of a kernel density estimator. The smoothed bootstrap is seen to provide new and independent motivation of some previously proposed methods. A curious feature of bootstrapping in this context is that no simulated resampling is required, since the needed functionals of the distribution can be calculated explicitly.

key words: bandwidth selection, bootstrap, density estimation, kernel estimators.

Subject Classification: 62G05.

Research Partially Supported by NSF Grant DMS–8902973

1. Introduction

This is a review of results concerning application of bootstrap ideas to bandwidth or smoothing parameter selection. The main ideas are useful in all types of nonparametric curve estimation settings, including regression, density and hazard estimation, and also apply to a wide variety of estimators, including those based on kernels, splines, orthogonal series, etc. However as much of the work so far has focused on perhaps the simplest of these, kernel density estimation, the discussion here will be given in this context.

The density estimation problem is often mathematically formulated by assuming that observations $X_1,\ldots,X_n$ are a random sample from a probability density $f(x)$, and it is desired to estimate $f(x)$. The kernel estimator of $f(x)$ is defined by

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

where K_h denotes the rescaling $K_h(\cdot) = K(\cdot/h)/h$, of the "kernel function" K (assumed throughout to be a symmetric probability density) by the "bandwidth" h. See Silverman (1986) for early references, good practical motivation, and useful intuitive discussion concerning this estimator.

The most important hurdle in the practical implementation of $\hat{f}(x)$ (and indeed essentially any nonparametric curve estimator), is the choice of the bandwidth h. While many methods have been proposed and studied, see the survey Marron (1988) for example, this remains an important problem in the application of curve estimators in general.

For a mathematical approach to this problem, most workers consider error criteria. Here the focus will be on the expected L^2 norm, or Mean Integrated Squared Error,

$$\mathrm{MISE}(h) = E \int [\hat{f}_h(x) - f(x)]^2 \, dx.$$

Devroye and Györfi (1984) provide an array of arguments in favor of the L^1 norm, but for simplicity of presentation and development of ideas, extensions to this challenging case will only be briefly discussed at the end.

An important advantage of $\mathrm{MISE}(h)$ as an error criterion is that it admits the

simple variance–bias^2 representation

$$\text{MISE}(h) = V(h) + B^2(h)$$

where

$$V(h) = n^{-1}\{h^{-1}\int K^2 + \int (K_h * f)^2\}$$
$$B^2(h) = \int (K_h * f - f)^2$$

using * to denote the convolution. Deep insight into the bandwidth selection problem comes from the asymptotic analysis, as $n \to \infty$, $h \to 0$, with $nh \to \infty$, assuming two uniformly continuous derivatives on f,

$$V(h) = n^{-1} h^{-1} \int K^2 + o(n^{-1}h^{-1}),$$
$$B^2(h) = h^4 \int (f'')^2 (\int x^2 K/2)^2 + o(h^4),$$

see Section 3.3.1 of Silverman (1986) for example. From this one can see why choice of h is not a simple matter, h small gives too much variability to the estimator (expected intuitively since there are not enough points in the effective local average when the window width is too small), h big introduces too much bias (again intuitively clear since a large window width introduces information from too far away).

Section 2 discusses bandwidth selection by minimization of bootstrap estimates of MISE(h). In particular it is seen why the smoothed bootstrap is very important here. An interesting and unusual feature of this case is that the bootstrap expected value can be directly and simply calculated, so the usual simulation step is unnecessary in this case.

Asymptotic analysis and comparison of these methods is described in Section 3. Connection to other methods, including Least Squares Cross Validation, is made in Section 4. Simulation experience and a real data example are given in Section 5.

2. Bootstrap MISE Estimation

The essential idea of bootstrapping in general is to find a useful approximation to the probability structure of the estimation process being studied, denoted in this case by

$\mathcal{L}\{\hat{f}_h(x) - f(x)\}$. The usual simple means of doing this involves thinking about "resampling with replacement". One way of thinking about this probability structure is through random variables $I_1,...,I_n$ which are independent of each other, and of $X_1,...,X_n$, and are uniformly distributed on the integers $\{1,...,n\}$. These new random variables contain the probability structure (conditioned on $X_1,...,X_n$) of the "resample" $X_1^*,...,X_n^*$ defined for $i = 1,...,n$ by

$$X_i^* = X_{I_i}.$$

As a first attempt at using this new conditional probability structure to model the bandwidth trade–off in the density estimation problem, one might define the "bootstrap density estimator"

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

and then hope the approximation

$$\mathcal{L}^*\{\hat{f}_h^*(x) - \hat{f}_h(x) | X_1,...,X_n\} \cong \mathcal{L}\{\hat{f}_h(x) - f(x)\},$$

is useful. Faraway and Jhun (1987) have pointed out that this approximation is in fact not useful for bandwidth selection because

$$E^*\{\hat{f}_h^*(x)\} = E^*\{K_h(x - X_{I_1})\} = \hat{f}_h(x),$$

where E^* is expected value "in the bootstrap world", in other words with respect to the bootstrap distribution $\mathcal{L}^*\{\cdot | X_1,...,X_n\}$, i.e. the average over all possible values of $I_1,...,I_n$. This shows that there is no bias in this bootstrap world, which is disastrous because as shown by the MISE(h) analysis in section 1, bias constitutes one of the two essential quantities to be balanced in bandwidth selection. Actually this is not too surprising because $\hat{f}_h$ is not in fact the density of the $\mathcal{L}^*\{\cdot | X_1,...,X_n\}$ distribution. Note moreover that this philosophical flaw can not be simply fixed because this is a discrete distribution (supported on $\{X_1,...,X_n\}$) and has no density.

This motivates finding another bootstrap probability structure to approximate $\mathcal{L}\{\hat{f}_h(x) - f(x)\}$. A natural candidate, proposed for bandwidth selection by Faraway and

Jhun (1987) and Taylor (1989), which does have a density is the smoothed bootstrap, introduced in Efron (1979). An alternative clever idea based on subsampling is proposed in Hall (1990), but this will not be discussed further here because connections to other methods, and comparison with them have not yet been well understood.

A means of studying the smoothed bootstrap, using notation as above, is to define additional random variables $\epsilon_1,\ldots,\epsilon_n$, independent of each other and of $X_1,\ldots,X_n$ and $I_1,\ldots,I_n$, having probability density $L_g(x)$, where for $g > 0$, L_g denotes the rescaling $L_g(\cdot) = L(\cdot/g)/g$. Now redefine the bootstrap sample by, for $i = 1,\ldots,n$

$$X_i^* = X_{I_i} + \epsilon_i.$$

Observe that the distribution of $X_1^* | X_1,\ldots,X_n$, i.e. the bootstrap distribution $\mathcal{L}^*\{\cdot | X_1,\ldots,X_n\}$, has probability density

$$\hat{f}_g(x) = n^{-1} \sum_{i=1}^{n} L_g(x - X_i).$$

Hence it seems natural to study when the approximation

$$\mathcal{L}^*\{\hat{f}_h^*(x) - \hat{f}_g(x) | X_1,\ldots,X_n\} \cong \mathcal{L}\{\hat{f}_h(x) - f(x)\},$$

is useful. This depends on the choices of L and g which are discussed in the next section.

This motivates the use of the bandwidth h^* which minimizes (break ties arbitrarily) the MISE in the bootstrap world,

$$\begin{aligned} \mathrm{MISE}^*(h) &= E^* \int \{\hat{f}_h^*(x) - \hat{f}_g(x)\}^2 \, dx, \\ &= V^*(h) + B^{*2}(h), \end{aligned}$$

where $\hat{f}_h$ is defined as above but using the smoothed bootstrap data, and where

$$\begin{aligned} V^*(h) &= n^{-1}\{h^{-1}\int K^2 + \int (K_h * \hat{f}_g)^2\}, \\ B^{*2}(h) &= \int (K_h * \hat{f}_g - \hat{f}_g)^2. \end{aligned}$$

An interesting fact about this setup is that the desired functionals of the bootstrap distribution can be simply calculated, so the usual computationally expensive simulation step is unnecessary. This is usually the case only for very simple examples, see section Chapter 5 of Efron (1982), although an interesting exception is in quantile estimation, see

Sheather and Marron (1990).

It is straightforward to compute $MISE^*(h)$ because V^* and B^{*2} admit the simple representations

$$V^*(h) = n^{-1}[h^{-1}\int K^2 + n^{-2}\sum_i\sum_j\{K_h{*}K_h{*}K_g{*}K_g\}(X_i - X_j)],$$

$$B^{*2}(h) = n^{-2}\sum_i\sum_j\{K_h{*}K_h{*}K_g{*}K_g - 2K_h{*}K_g{*}K_g + K_g{*}K_g\}(X_i - X_j).$$

Hence calculation of h^* requires about the same computational effort as the least squares cross–validated bandwidth (discussed in section 3.4.3 of section Silverman 1986).

For a completely different approach to bootstrapping in density estimation, see Hall (1990).

3. Asymptotics

In this section, choice of g and L is considered. A sensible first attempt, see Taylor (1989), would be to try $g = h$ and $K = L$. This can be easily analyzed using the assumptions and asymptotics at the end of section 1, with the important part being

$$B^{*2}(h) \cong h^4 \int(\hat{f}_g'')^2 (\int x^2K/2)^2.$$

This presents a problem, because for $g \sim n^{-1/5}$, which is the reasonable range for h see section 3.3.2 of Silverman (1986) for example, $\hat{f}_g''(x)$ does not even converge to $f''(x)$ (because the variance does not tend to 0). For this reason, Faraway and Jhun (1987) propose using $g > h$. However observe that $f''(x)$ is not what is needed here, instead we need the functional $\int(f'')^2$ which is a different problem. Indeed for $g \sim n^{-1/5}$, Hall and Marron (1987) show that

$$\int(\hat{f}_g'')^2 \to \int(f'')^2,$$

although this choice of the bandwidth g is quite inefficient in the sense that it gives a slower than necessary rate of convergence.

A means of quantifying this inefficiency, which is relevant to bandwidth selection, is

to study its effect on the relative rate of convergence. In remark 3.6 of Hall, Marron and Park (1990), it is shown that

$$(h^*/h_0) - 1 \sim n^{-1/10},$$

when $g = h$, where h_0 denotes the minimizer of MISE(h). This very slow rate of convergence is the same as that well known to obtain for least squares cross–validation, and for the biased cross–validation method of Scott and Terrell (1987) (which uses $g = h$ in a slightly different way). For this reason, as well as the fact that the appropriate bandwidth for estimating $\int (f'')^2$ is different from that for estimation of $f(x)$, the choice $g = h$ does not seem appropriate.

Good insight for the problem of how to choose g has been provided by the main results of Hall, Marron and Park (1990). The minor modification of these results presented here is explicitly given in Jones, Marron and Park (1990), where it is seen that if f has slightly more than four derivatives, and L is a probability density, for C_1, C_2 and C_3 constants depending on f, K and L,

$$(h^*/h_0) - 1 \xrightarrow{d} C_1 n^{-1} g^{-9/2} Z + (C_2 g^2 + C_3 n^{-1} g^{-5}),$$

where Z is a standard normal random variable. Note $C_3 = 0$ and $g \sim n^{-1/5}$ gives the slow $n^{-1/10}$ rate in the above paragraph. This expansion is important because it quantifies the trade–off involved in the choice of g. In particular there is too much "variance" present if $g \to 0$ too rapidly, and a "bias" term that penalizes $g \to 0$ too slowly. This variance and bias can be combined them into an "asymptotic mean squared error" which can then be optimized over g to see that the best g has the form

$$g \sim C_4(f,K,L)\, n^{-1/7},$$

which gives

$$(h^*/h_0) - 1 \sim n^{-5/14}.$$

Data based methods for estimating C_4 are given in Jones, Marron and Park (1990). Note that this rate of convergence is much faster than $n^{-1/10}$.

A natural question at this point is: can the rate $n^{-5/14}$ be improved? As noted in

remark 3.3 of Hall, Marron and Jones (1990), by taking L to be a higher order kernel, this rate can be improved all the way up to the parametric $n^{-1/2}$ (L needs to be of "order 6" for this fast rate). This rate has been shown to be best possible by Hall and Marron (1990). However there is a distinct intuitive drawback to this in that when L is a higher order kernel, it can no longer be thought of as a probability density, so h^* is no longer a bootstrap bandwidth, at least in the usual sense of the word.

A more intuitive way of achieving root n convergence is given in Jones, Marron and Park (1990), who consider factorizations of g, in terms of h, of the form

$$g = C n^p h^m.$$

In particular for $m = -2$ and suitable p they obtain

$$(h^*/h_0) - 1 \sim n^{-1/2},$$

in the much more natural case $K = L$.

4. Connection to Other Methods

Bandwidths that are essentially the same as h^* have been derived from considerations much different from bootstrapping. In particular note that the dominant part of the representation of $V(h)$ at the end of section 1 does not depend on f, so it is natural to estimate $V(h)$ by $n^{-1}h^{-1}\int K^2$, which is asymptotically equivalent to $V^*(h)$. The fact that $B^{*2}(h)$ provides a natural estimate of $B^2(h)$ can also be derived in a natural way by thinking about replacing the unknown f in B^2 by the pilot kernel estimator $\hat{f}_g$. Such nonbootstrap motivations for a bandwidth selector very close to h^* were developed independently in an unpublished paper by Ramlau–Hansen and in the related regression setting by Müller (1985). See Chiu (1990) for related ideas from the Fourier Transform point of view.

Hall, Marron and Park (1990) motivate a very similar bandwidth selector, but by a different method. They propose decreasing the variability of the least squares

cross–validated bandwidth through a "pre–smoothing" of the pairwise differences of the data. Note that, using $\hat{f}_{hj}$ to denote the kernel estimator based on the sample with X_j excluded, the least squares cross–validation criterion can be written in the form

$$CV(h) = \int \hat{f}_h^2 - 2\, n^{-1} \sum_{j=1}^{n} \hat{f}_{hj}(X_j)$$
$$\cong n^{-1}h^{-1}\int K^2 + n^{-1}(n-1)^{-1} \sum_{i \neq j} \sum \{K_h * K_h - 2K_h\}(X_i - X_j),$$

where the approximation comes from replacing a factor of n^{-1} by $(n-1)^{-1}$. Note that the first term provides the same sensible estimate of $V(h)$ discussed in the paragraph above, while the second term has features reminiscent of the representation of B^{*2} given at the end of Section 2. To make this connection more precise, note that when there are no replications among $X_1,\ldots,X_n$, the second term is the limit as $g \to 0$ of

$$\hat{B}^2(h) = n^{-1}(n-1)^{-1} \sum_{i \neq j} \sum \{K_h * K_h * K_g * K_g - 2K_h * K_g * K_g + K_g * K_g\}(X_i - X_j).$$

Note also that by the associative law for convolutions, one may view this as first plugging the differences into $K_g * K_g$, and then putting the result into the bias part of CV, which is why this idea was called smoothed cross–validation by Hall, Park and Marron (1990).

The important difference between $\hat{B}^2(h)$ and $B^{*2}(h)$ is whether or not the "terms on the diagonal" are included in the double sum. At first glance one may feel uncomfortable about these terms because they do not depend on the data. At second glance it is not clear that they will have a very large effect, unless $g <= h$, when they contribute a term of order $n^{-1}h^{-1}$. For this reason Taylor(1989) deleted these terms in his $g = h$ implementation of the smoothed bootstrap. However more careful analysis, in the case $g >> h$, shows a rather slight theoretical superiority of B^{*2} over $\hat{B}^2$ has been demonstrated by Jones, Marron and Park (1990) in terms of the relative rate of convergence. Simulation work has also indicated usually small superiority of the diagonals in approach, although the improvement is sometimes much larger because the diagonals out version is less stable. One possible explanation as to why this happens is that B^{*2} is the smoothed bootstrap estimate of B^2, while $\hat{B}^2$ does not seem to have any such

representation.

Faraway and Jhun (1987) have pointed out that the bootstrap approximation can be used to understand the bandwidth trade–off entailed by other means of assessing the error in $\hat{f}_h$. For example one could replace the L^2 based MISE with the expected L^1 norm. A major drawback to this is that it seems that an exact calculation of the bootstrap expected value E^* is no longer realistically available. Hence this expected value will need to be evaluated by simulation, which will be far more (perhaps prohibitively?) expensive from the computational viewpoint. Another example is the replacement of MISE by the pointwise Mean Squared Error, where one focuses on estimation of f at one fixed location x. Here the exact calculation of E^* can be done, however this has not been explored carefully yet, mostly because it seems sensible to postpone investigation of this more challenging pointwise case until more is understood about the global MISE problem.

5. Simulations and an Application

To see how these methods worked in a simulation context, various versions of the bootstrap bandwidth selectors were tested. Several methods of choosing the pilot bandwidth g, as discussed in Jones, Park and Marron (1990), including immediate use of a $N(0,\hat{\sigma}^2)$ reference distribution and also estimation of the unknown functionals as suggested in Jones and Sheather (1990), were tried. The results were usually better when estimates were used, so one step estimators of this type were used for the following discussion. To speed the computations, a binned implementation of the type described in Scott and Härdle (1990) was employed.

For this, 500 pseudosamples of size $n = 100$ were generated from the normal mean mixture density described in Park and Marron (1990). The results are visually summarized in Figure 1, which is very similar to Figure 3b in that paper. The bandwidth selectors CV, BCV and OS there are not shown here, because as one would expect from

the results of that paper they were inferior to these newer ones. PI is the main bandwidth selector discussed by Park and Marron. Note that Taylor's $g = h$ method performed quite poorly in comparison to the others, with a strong bias towards oversmoothing. This poor performance is not surprising in view of the theoretical results described above. The simple bootstrap, denoted BSS, which uses a data based g chosen independently of h, gave performance roughly comparable to the Park and Marron PI. It is not straightforward to compare these, because there is slightly more bias, but slightly less variability. However the bandwidth factorized bootstrap, i.e. the $n^{-1/2}$ method described at the end of Section 3, denoted BSF, gave much better performance, having less variability and also less bias than the others.

[put Figure 1 about here]

These selectors have also been tried for other sample sizes and other densities as well. For those densities not too far from the Normal in shape, the asymptotics describe the situation well, with larger sample sizes giving more rapid improvements in BSF than the others (as expected from its faster rate of convergence). For the N(0,1) BSF gave really superlative performance, in fact even beating out the Normal reference distribution bandwidth given at (3.28) of Silverman (1986). For densities which are still unimodal, but depart strongly from the normal in directions of strong skewness or kurtosis, the performance was not so good (in fact CV is typically the best in terms of MISE), but can be improved a lot by using a scale estimate which is more reasonable than the sample standard deviation in such situations, such as the interquartile range. On the other hand when f is far from normal in the direction of heavy multimodality, again most of these newer bandwidth selectors were inferior to CV in the MISE sense, but the sample standard deviation was a more reasonable scale estimate than the IQR. A way to view both of the above situations, is that they are cases where it takes very large sample sizes before the effects described by the asymptotics take over. There is still work to be done in finding a bandwidth selector which works acceptably well in all situations.

To see how well these methods work on a real data set, they were tried on the income data shown in Figure 2 of Park and Marron (1990). The data and importance of that type of display are discussed there. Several of the bootstrap bandwidth selectors considered in this paper were tried on this data set. The best result was for SBF with the $N(0,\hat{\sigma}^2)$ reference distribution used immediately. Figure 2 here, which compares nicely to Figure 2 in Park and Marron shows the result. The other variants, involving estimation steps in the choice of g, tended to give smaller bandwidths, which are probably closer to the MISE value, but gave estimates that are too rough for effective presentation of this type.

[put Figure 2 about here]

REFERENCES

Chiu, S. T. (1990) Bandwidth selection for kernel density estimation, unpublished manuscript.

Devroye, L. and Györfi, L. (1984), Nonparametric density estimation: the L_1 view. Wiley, New York.

Efron, B. (1979) Bootstrap methods: another look at the jackknife, Annals of Statistics, 7, 1–26.

Efron, B. (1982) The jackknife, the bootstrap and other resampling plans, CBMS Regional Conference series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia.

Faraway, J. J. and Jhun, M. (1987) Bootstrap choice of bandwidth for density estimation, unpublished manuscript.

Hall, P. (1990) Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems, to appear in Journal of Multivariate Analysis.

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

Hall, P. and Marron, J. S. (1990) Lower bounds for bandwidth selection in density estimation, to appear in Probability Theory and Related Fields.

Hall, P., Marron, J. S. and Park, B. U. (1990) Smoothed cross–validation, unpublished manuscript.

Jones, M. C. and Sheather, S. J. (1990) Using nonstochastic terms to advantage in kernel–based estimation of integrated squared density derivatives, unpublished manuscript.

Jones, M. C., Marron, J. S. and Park, B. U. (1990) A simple root n bandwidth selector, unpublished manuscript.

Marron, J. S. (1988) Automatic smoothing parameter selection: a survey, Empirical Economics, 13, 187–208.

Müller, H.–G. (1985) Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators, Statistics and Decisions, Supplement no. 2, 193–206.

Scott, D. W. and Härdle, W. (1990) Weighted averaging using rounded points, to appear in Journal of the Royal Statistical Society, Series B.

Scott, D. W. and Terrell, G. R. (1987) Biased and unbiased cross–validation in density estimation, Journal of the American Statistical Association, 82, 1131–1146.

Sheather, S. J. and Marron, J. S. (1990) Kernel quatile estimation, to appear in Journal of the American Statistical Association.

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

Taylor, C. C. (1989) Bootstrap choice of the smoothing parameter in kernel density estimation, Biometrika 76, 705–712.

Captions

Figure 1: MISE and kernel density estimates of the distributions of the automatically selected bandwidths (on $\log_3$ scale). Taken from 500 Monte Carlo replications of samples of size 100 from $.5N(-1,4/9) + .5N(1,4/9)$.

Figure 2: Expanded respresentation of 16 density estimates for incomes in the United Kingdom, 1968–1983. Bandwidths chosen by bandwidth factorized smoothed bootstrap.

Figure 1

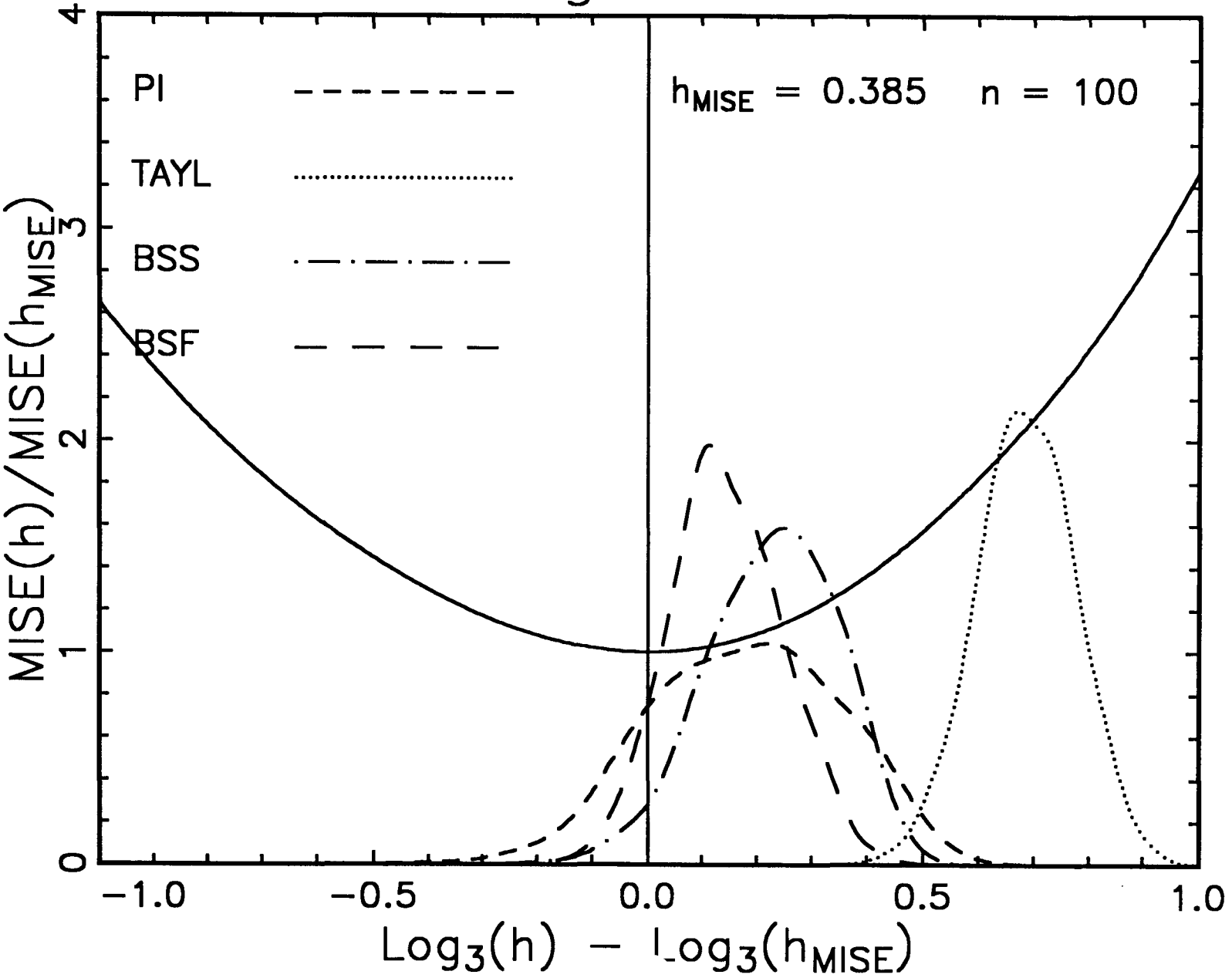
PI
TAYL
BSS
BSF
h_{MISE} = 0.385 n = 100
MISE(h)/MISE(h_{MISE})
0
1
2
3
4
−1.0
−0.5
0.0
0.5
1.0
$Log_3(h) - Log_3(h_{MISE})$

Figure 2