A Brief Survey of Bandwidth Selection for Density Estimation

M. C. JONES, J. S. MARRON, and S. J. SHEATHER

There has been major progress in recent years in data-based bandwidth selection for kernel density estimation. Some "second generation" methods, including plug-in and smoothed bootstrap techniques, have been developed that are far superior to well-known "first generation" methods, such as rules of thumb, least squares cross-validation, and biased cross-validation. We recommend a "solve-the-equation" plug-in bandwidth selector as being most reliable in terms of overall performance. This article is intended to provide easy accessibility to the main ideas for nonexperts.

KEY WORDS: Bandwidth selection; Kernel density estimation; Nonparametric curve estimation; Smoothing parameter selection.

1. INTRODUCTION

Smoothing methods provide a powerful methodology for gaining insights into data. Many examples of this may be found in the monographs of Eubank (1988), Härdle (1990), Müller (1988), Scott (1992), Silverman (1986), Wahba (1990), and Wand and Jones (1994). But effective use of these methods requires choice of a smoothing parameter. When insufficient smoothing is done, the resulting density or regression estimate is too rough and contains spurious features that are artifacts of the sampling process. When excessive smoothing is done, important features of the underlying structure are smoothed away.

In the hands of an expert, interactive visual choice of the smoothing parameter is a very powerful way to analyze data. But there are a number of reasons why it is important to be able to choose the amount of smoothing automatically from the data. One is that software packages need a default. This is useful in saving the time of experts through providing a sensible starting point, but it becomes imperative when smoothing is used by nonexperts. Another reason this is important is that in a number of situations many estimates are required, and it can be impractical to manually select smoothing parameters for all (e.g., see the income data in Park and Marron 1990). An extreme case of this comes in the important field of dimensionality reduction. In that context, many methods, such as projection pursuit, additive modeling, ACE, MARS, SIR, and so on, are based on repeated use of smoothers. Manual choice of smoothing parameter at each step is clearly infeasible.

In this article we focus only on the specific case of one-dimensional kernel density estimation. This is done because research in data-based bandwidth selection has progressed much further there than in other contexts. Perhaps this is because of the appealing simplicity of this setting. Of course, more general contexts (e.g., higher dimensions and estimation of other functions, such as regression) are of great interest. We hope that this summary will help stimulate work in those important related areas, and also provide some guidance. (See Gasser, Kneip, and Köhler 1991 and Ruppert, Sheather, and Wand 1995 for first steps in the direction of kernel regression.)

Many proposals for data-based bandwidth selection methods have been made over the years. A few important ones are discussed in Section 2. For simple understanding of the many proposals, we group them into "first generation" and "second generation" methods, because there has been a quantum leap in terms of performance (both theoretical and practical) for a number of more recently developed methods as compared to the earlier ones. Most "first generation" methods were developed before 1990. This decade has seen some major breakthroughs in terms of performance, in several directions, and also a large number of relatively minor variations. We apply the name "second generation" to those with superior performance. Simple access to, and understanding of, these techniques is the main point of this article.

Most of the "first generation methods" have been surveyed by Marron (1989) (see also Scott 1992 and Silverman 1986). An exhaustive treatment of these methods is not a goal of this article (as it would obscure our main points) but has been provided by Jones, Marron, and Sheather (1992). Hence only some of the best known of these—rules of thumb, least squares cross-validation, and biased cross-validation—are explicitly considered here. The motivations for each of these are discussed in Section 2.2.

This article is intended to provide quick access to the main ideas behind "second generation methods." Again for the sake of clarity, because there are many variations that can obscure the main ideas, we focus only on two representative methods: a "solve-the-equation plug-in" method and a "smoothed bootstrap," described in Section 2.3. We avoid a historical treatment, and give only partial references. (For a more complete treatment, from a historical viewpoint, with complete references, and detailed discussion of variations that have been suggested, see Jones et al. 1992.) Quick access to implementation of most of the methods discussed here has been provided by Park and Turlach (1992).

M. C. Jones is Reader, Department of Statistics, The Open University, Milton Keynes MK7 6AA, United Kingdom. J. S. Marron is Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599. S. J. Sheather is Associate Professor, Australian Graduate School of Management, University of New South Wales, Sydney NSW 2052, Australia. Marron's research was supported by National Science Foundation Grant DMS-9203135 and the Australian Graduate School of Management.

Journal of the American Statistical Association
March 1996, Vol. 91, No. 433, Review Paper

Various ways of comparing bandwidth selectors are considered in Section 3. These include real data examples, with a new one shown in Section 3.1; asymptotic analysis, with major ideas surveyed in Section 3.2; and a simulation study, whose main lessons are summarized in Section 3.3.

Taken together, these comparisons provide a preponderance of evidence for the three main points of this article.

1. Second generation methods are far superior to the better-known first generation methods.
2. Second generation methods are ready for widespread use as defaults in software packages.
3. The solve-the-equation plug-in method is the best of the second generation methods in terms of overall performance. We suggest that this should become a benchmark for good performance.

2. METHODS AND MOTIVATIONS

Here we discuss the main ideas behind some important bandwidth selection methods. (See Jones et al. 1992 and Marron 1989 for a comprehensive introduction.) Some background material and notation, common to many methods, is discussed first.

2.1 Background

Interesting structure in a set of data, $X_1, \ldots, X_n$, is often revealed through plotting the kernel density estimator,

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

where $K_h(\cdot) = (1/h)K(\cdot/h)$ for a "kernel function" K (often taken to be a symmetric probability density) and a "bandwidth" h (the smoothing parameter). (See Scott 1992 and Silverman 1986 for many interesting examples, and a good introduction to important ideas.)

The behavior of $\hat{f}$ may be useful mathematically analyzed by assuming that the data are independent realizations from a probability density $f(x)$ and by viewing $\hat{f}$ as an estimator of f. A common way of measuring the error in this estimation process is the mean integrated squared error (MISE),

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

where $\int$ denotes definite integration over the real line and dependence on h is made explicit because of the important effect of this smoothing parameter. There has been much discussion of the appropriateness of MISE as a measure of error (see Grund, Hall, and Marron 1994 and Jones et al. 1992, sec. 2, for details). Here we use MISE because it is simple and allows very deep analysis, as was shown by Marron and Wand (1992). Park and Turlach (1992) considered number of modes as a criterion for performance and arrived at similar conclusions to those given here. (See Marron and Tsybakov 1995 for a useful alternative measure of error.)

Asymptotic analysis provides a simple way of quantifying how the bandwidth h works as a smoothing parameter. In particular, under standard technical assumptions (see, for example, Silverman 1986, sec. 3.3), MISE is asymptotically (as $n \to \infty$) approximated by the asymptotic mean integrated squared error (AMISE),

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

where here and in the following the functional notation $R(\varphi) = \int \varphi^2(x)\,dx$ is used and $\int x^2 K = \int x^2 K(x)\,dx$. This quantifies the effect of the smoothing parameter h. In particular, note that the first term (integrated variance) is large when h is too small, and the second term (integrated squared bias) is large when h is too large.

Another useful feature of AMISE(h) is that its minimizer is simply calculated:

$$h_{\text{AMISE}} = \left[\frac{R(K)}{nR(f'')\left(\int x^2 K\right)^2} \right]^{1/5}$$

This provides simple insight into "good" bandwidths. For example, smaller bandwidths are better for larger n (sensible, because the estimator should be "more local" when more information is present) and when the density is rougher (because the bias effect is stronger). In many circumstances h_{AMISE} is a good approximation to h_{MISE}, the minimizer of MISE (with ties broken arbitrarily in the case of multiple minimizers), but sometimes it is not, as indicated by Marron and Wand (1992).

Note that the estimator discussed here uses the same amount of smoothing at all locations. As noted by Scott (1992, sec. 6.6) and Silverman (1986, sec. 5.1), in some contexts large improvements can be made with local bandwidth methods. But these require an even more difficult bandwidth choice (a whole function, not just a number) and hence are not discussed in this article.

2.2 First Generation Methods

First generation methods for bandwidth selection were mostly proposed before 1990. Three of the best known of these are discussed here (see Scott 1992 and Silverman 1986 for detailed discussion of some of their properties). Many others have been surveyed by Jones, Marron, and Sheather (1992) and by Marron (1989).

2.2.1 Rules of Thumb. This idea goes back at least to Deheuvels (1977) and was popularized by Silverman (1986). It involves replacing the unknown part of h_{AMISE}, $R(f'')$, by an estimated value based on a parametric family. Because scale is very important for bandwidth choice but location is not, a natural choice for parametric family is $N(0, \sigma^2)$. (See Janssen, Marron, Veraverbeke, and Sarle 1995 and Silverman 1986, sec. 3.4.2, for discussion of scale estimates.). In this article we use h_{ROT} to denote this bandwidth based on standard deviation, as in (3.28) of Silverman (1986). In our simulations this was somewhat worse than the best scale measure of Janssen et al. (1995) but was close enough to be reasonably representative.

An interesting variation of this idea is "oversmoothing," proposed by Terrell and Scott (1985) and Terrell (1990). They solved the variational problem of minimizing $R(f'')$ subject to various types of scale constraints. This solution, together with a scale estimate, results in a "maximal possible amount of smoothing," and an "oversmoothed bandwidth" that comes from using this in h_{AMISE}. Upper bounds on a "reasonable amount of smoothing" are quite useful in certain contexts; for example, determination of a grid for searching by more sophisticated methods. But we do not view the oversmoothed bandwidth as suitable for general use, because it is larger than h_{ROT}, which already suffers from being unacceptably large in that it often smooths out important features such as major modes.

2.2.2 Least Squares Cross-Validation. This idea was first published by Bowman (1984) and Rudemo (1982). A simple motivation comes from representing the integrated squared error (ISE) as

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

Note that the minimizer of the ISE is the same as the minimizer of the first two terms of the final form. The first term is entirely known, and the second term can essentially be estimated by the method of moments. For technical reasons not discussed here, the second term is estimated by $-2n^{-1}\sum_{i=1}^{n} \hat{f}_i(X_i)$, where $\hat{f}_i$ is the "leave-one-out" kernel density estimator defined using the data with X_i removed. The largest local minimizer (which gives better empirical performance than the global minimizer) is denoted by h_{CV}. The function being minimized has fairly frequent local minima, as quantified by Hall and Marron (1991a).

2.2.3 Biased Cross-Validation. The biased cross-validation (BCV) method, proposed by Scott and Terrell (1987), attempts to directly minimize the AMISE. This requires estimation of the unknown $R(f'')$, which requires selecting another bandwidth. This difficulty is addressed by taking the bandwidth to be the dummy variable of minimization. The smallest local minimizer (which gives better empirical performance than the global minimizer) of

$$\mathrm{BCV}(h) = n^{-1}h^{-1}R(K) + h^4 \times \left[R(\hat{f}_a'') - \frac{R(K'')}{mh}\right]\left(\int x^2 K/2\right)^2,$$

is denoted by h_{BCV}.

2.3 Second Generation Methods.

None of the second generation bandwidth selection methods have been carefully assessed in monographs other than that by Wand and Jones (1994). Many reasonably effective selectors have been proposed. Here we represent the collection by two different approaches.

2.3.1 Solve-the-Equation Plug-In Approach. Many authors have written on plug in ideas, going back at least to Woodroofe (1970). The main idea is to plug an estimate of the unknown $R(f'')$ into the representation for h_{AMISE}. A practical difficulty is the choice of the bandwidth of the pilot estimate. The "solve-the-equation" approach to this was proposed by Hall (1980) and Sheather (1983, 1986) and refined in a series of papers, with the version presented here developed by Sheather and Jones (1991). The idea is to take h_{SJPI} to be the solution of the fixed-point equation

$$h = \left[\frac{R(K)}{nR(\hat{f}''_{g(h)})\left(\int x^2 K\right)^2}\right]^{1/5}$$

In addition to the different form of the quantity being optimized, a very important difference between this approach and BCV is that the pilot bandwidth here is written in the form $g(h)$. This is because bandwidths that are appropriate for curve estimation are quite different from those that are right for estimation of $R(f'')$. In fact when using the optimal bandwidth for estimation of f, the estimate $\hat{f}''$ is asymptotically inconsistent for f'', and $R(\hat{f}'')$ is only barely consistent for $R(f'')$, with much better performance available from better bandwidths (see Hall and Marron 1987a and Jones and Sheather 1991). We believe this is the main reason that h_{BCV} gives performance only in the "first generation" class.

A drawback to using a better bandwidth for estimation of $R(f'')$ is that then this better bandwidth must be chosen. This is done by finding an analog of h_{AMISE} for the problem of estimating $R(f'')$ by $R(\hat{f}''_g)$. In particular, and minimizer of the asymptotic mean squared error (AMSE) for this problem has the form

$$g_{\mathrm{AMSE}} = C_1\{R(f''')\}C_2(K)n^{-1/7}$$

for suitable functionals C_1 and C_2. An expression for "g in terms of h" comes from solving the representation of h_{AMISE} for n and substituting to get

$$g(h) = C_3\{R(f''), R(f''')\}C_4(K)h^{5/7}$$

for appropriate functionals C_3 and C_4. The unknowns $R(f'')$ and $R(f''')$ are estimated by $R(\hat{f}'')$ and $R(\hat{f}''')$, with bandwidths chosen by reference to a parametric family, as for h_{ROT}.

Again, many variations have been proposed and studied. One of these is to try to reduce the influence of the normal parametric family even further by using pilot kernel estimates instead of the normal reference (with the higher-stage pilot bandwidths chosen by the normal reference method). The obvious hierarchy of such methods has been considered by Park and Marron (1992), who showed that there are gains, in terms of asymptotic rate of convergence, up to the point described explicitly earlier, but no gains for higher-order versions. Simulations confirm this and show that higher-order pilot estimation entails some cost in terms of more variability. Jones, Marron and Sheather (1992) have discussed some other variations, and Chiu (1992) and Engel, Herrmann, and Gasser (1994) have given other plug-in ideas that are very successful.

2.3.2 Smoothed Bootstrap. One approach to this method is to consider the bandwidth that is a minimizer

of a smoothed bootstrap approximation to the MISE. Early versions of this were proposed by Faraway and Jhun (1990) and by Taylor (1989). An interesting feature of this approach is that unlike most bootstrap applications, the MISE in the "bootstrap world" can be calculated exactly, instead of requiring the usual simulation step, which makes it as computationally fast as other methods discussed here. In particular, following straightforward calculations (e.g., as in Marron 1992), the smoothed bootstrap estimate of the MISE (BMISE) has the form

$$\text{BMISE}(h) = n^{-1}\{h^{-1}R(K) + r(K_h * \hat{f}_g)\} + R(K_h * \hat{f}_g - \hat{f}_g),$$

where * denotes convolution. This is seen as a simple and appealing estimate of the MISE by rewriting the latter in the form

$$\text{MISE}(h) = n^{-1}\{h^{-1}R(K) + R(K_h * f)\} + R(K_h * f - f).$$

An attractive feature of this approach is that it does not work through the asymptotic AMISE but rather more directly targets MISE itself. For choice of the pilot bandwidth g, a number of approaches have been proposed, most involving stages of pilot estimation and use of reference distributions at some point, in the same spirit as in the previous sections (see Jones et al. 1992 for details and references).

With some algebra (see Marron 1992), the smoothed bootstrap approach can be seen to be nearly equivalent to some other approaches that have different motivations. These include the "double smoothing" idea that goes back at least to Müller (1985), and also "smoothed cross-validation" as proposed by Hall, Marron, and Park (1992). Successful variations of this idea were discussed by Chiu (1992).

3. COMPARISONS

In this section we compare the various bandwidth selectors in three different ways: through applying them to real data sets, through asymptotic analysis, and through simulation. Each of these has its obvious drawbacks and limitations, which is why it is imperative to study all three.

3.1 Real Data Examples

An important measure of the performance of any statistical method is how well it performs in practice. A number of interesting examples were presented by Sheather (1992). The main lessons from these examples are reasonably well represented in a new example presented here. Figure 1 shows kernel density estimates constructed using several methods discussed earlier, for the variable "Lean Body Mass" in the Australian Institute of Sports data set in exercise 2.4 of Cook and Weisberg (1994).

The density estimate using h_{ROT} is somewhat oversmoothed. There is some suggestion of bimodality, but it is weakened by this bandwidth being too large. This problem is common and is often much worse than this. The density estimate based on the bandwidth h_{LSCV} is severely undersmoothed. There are many spurious bumps, which make it hard to understand the structure of the data. The bandwidth h_{BCV} is more oversmoothed than h_{ROT}, and the suggestion of bimodality is even weaker. The density estimate based on h_{SJPI} shows a stronger indication of bimodal structure.

An important issue is whether or not two modes are "really present," or are they just artifacts of undersmoothing

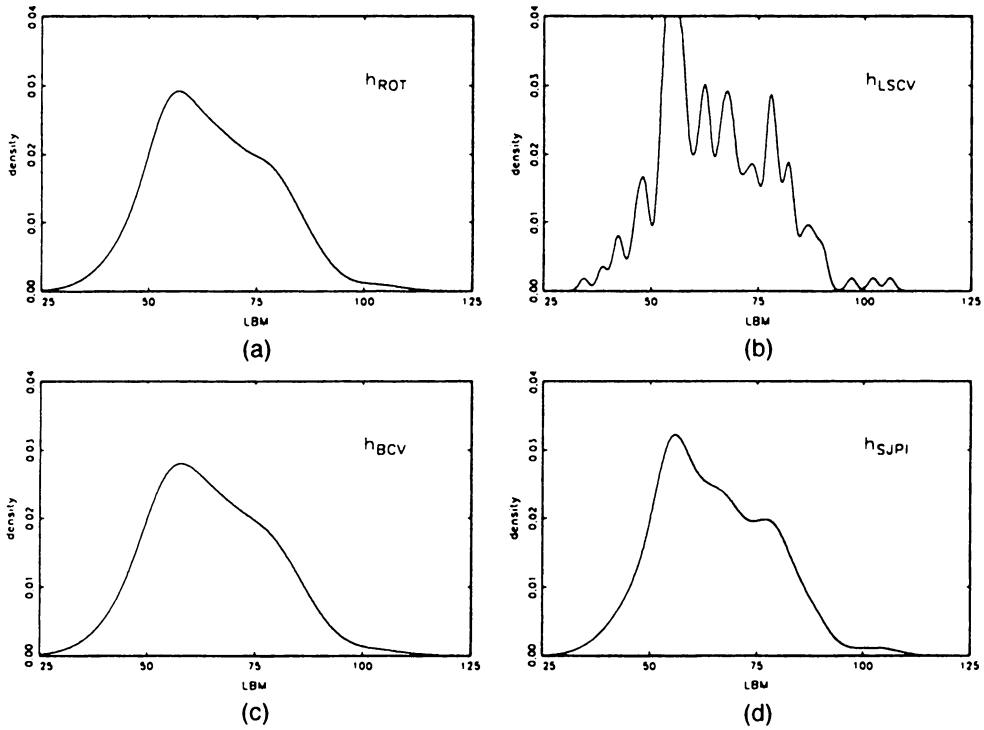

Figure 1. Gaussian Kernel Density Estimates Showing the Distribution of n = 202 Measurements of Lean Body Mass From the Australian Institute of Sport Data, Using Various Data-Driven Bandwidths: (a) h_{ROT}; (b) h_{LSCV}; (c) h_{BCV}; (d) h_{SJPI}. The strongest suggestion of bimodality is given by h_{SJPI}.

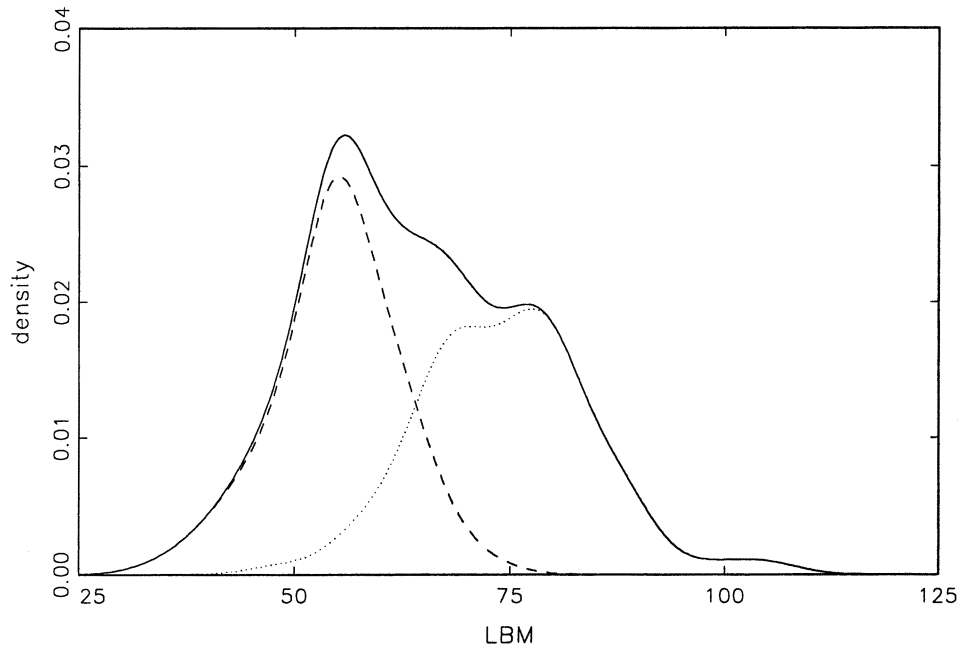

Figure 2. Density Estimates, Using h_{SJPI} (for Full Population), Based on Subpopulations of n = 102 Females (Dashed Line) and n = 100 Males (Dotted Line). The solid line represents the combined data set. This verifies that the bimodal structure in Figure 1(d) was an important feature of the data worth deeper investigation.

by h_{SJPI}. An answer is available in the present case by stratifying the population according to sex. Figure 2 shows kernel density estimation for the subpopulations (vertically scaled according to sample size, so the sum of the two is the density estimate for the whole population), using the same bandwidth. This makes it clear that the two modes are important features of the population.

The examples of Jones et al. (1992) and Sheather (1992) similarly reveal that h_{ROT} is too often seriously oversmoothed (missing important features), h_{LSCV} is too variable (especially in the direction of undersmoothing) and hence unreliable, and h_{BCV} also has a tendency to oversmooth (and has some instability problems as well). These examples make it clear that first generation methods are not appropriate for widespread use. On the other hand, h_{SJPI} is a consistent and stable performer that is ready for use as the default in software packages. (The smoothed bootstrap is not considered here because in the simulations that follow it had similar, but slightly worse, performance than h_{SJPI}.)

3.2 Asymptotic Analysis

Asymptotic analysis, as the sample size $n \to \infty$, has proved to be a useful tool in understanding the performance of data-based bandwidth selectors. Often it is useful in distinguishing between "first generation" and "second generation" bandwidth selectors. This is not always true, and indeed some selectors (e.g., those proposed in Hall, Sheather, Jones and Marron 1991 and in Park, Kim, and Marron 1994) have excellent asymptotics but very poor performance for real data sets and in simulation studies. The problem appears to be that for some methods, the asymptotics do not "come into effect" for reasonable sample sizes. This is why it is so important to look more deeply in comparing bandwidth selectors.

Nonetheless, for many bandwidth selectors of both generations, the asymptotic lessons do kick in at moderate sample sizes and complement the lessons learned from other viewpoints quite well. In particular, for a data-driven bandwidth selector $\hat{h}$, it is often useful to study the asymptotic behavior of the random variable

$$\frac{\hat{h} - h_{\mathrm{MISE}}}{h_{\mathrm{MISE}}}$$

(relative error is appropriate, because reasonable bandwidths tend to zero as the sample size grows) under appropriate technical assumptions. This quantity is not of a priori interest for bandwidth selection, but as discussed, for example, by Park and Marron (1990), this measure of performance is the driving force behind more interesting measures of error, such as

$$\mathrm{MISE}(\hat{h}) - \mathrm{MISE}(h_{\mathrm{MISE}}).$$

Another alternative would be to study asymptotics based on the ISE $= \int (\hat{f}_h - f)^2$, but as noted by Grund et al. (1994), there is no important practical difference between this and MISE in assessing the performance of bandwidth selectors.

Following the classical notion of "consistency," one might first ask that data-driven bandwidth satisfies

$$\frac{\hat{h} - h_{\mathrm{MISE}}}{h_{\mathrm{MISE}}} \to 0.$$

This is true for most bandwidth selectors of both generations, with the important exception that it does not hold (in general) for h_{ROT} or for the oversmoothed bandwidths.

The main distinction between first generation and second generation bandwidths is in the rate of convergence.

For most methods (see Jones et al. 1992 for references and detailed discussion),

$$\frac{\hat{h} - h_{\mathrm{MISE}}}{h_{\mathrm{MISE}}} \sim n^{-p}$$

for some power p, in the sense that when multiplied by n^p, the ratio has a limiting Normal distribution. Most first generation methods, including h_{LSCV} and h_{BCV}, suffer from the excruciatingly slow rate of $p = \frac{1}{10}$. To dramatize how slow this is, it is interesting to ask how large n should be to give $n^{-1/10} = .1$; note that this requires $n = 10{,}000{,}000{,}000$. On the other hand, second generation methods enjoy much faster rates of convergence. For example, for both h_{SJPI} and analogous versions of the smoothed bootstrap, $p = \frac{5}{14} \approx .36$.

Various other rates are available for various second generation bandwidth selectors, with some as fast as $p = \frac{1}{2}$. This rate is known to be the best possible, and even the best constant is known as well. Interested readers are referred to work of Jones et al. (1992), but these points are not discussed in detail here, because many of these asymptotics do not have an important practical effect in the simulations for reasonable sample sizes.

3.3 Simulations

A major simulation study has been performed, using the 15 normal mixture densities of Marron and Wand (1992), for sample sizes $n = 100, 1{,}000$. Again, to avoid confusing the main issues, we give here only a summary of the results, and the interested reader is referred to work of Jones et al. (1992).

A concept important to our results is that some of the target densities are "easy to estimate," meaning that enough information is present in the data to recover most of the features of the target. Other densities in this set are "hard to estimate" in the sense that they contain features (e.g., thin spikes) that cannot be recovered from the sample sizes considered.

The shape of the distribution of the random bandwidths was important in determining performance—in particular, the mean and the variance. The main results were as follows.

1. The distribution for h_{ROT} had a mean that was usually unacceptably large (because this method is not unlike the oversmoother, which is usually too large). But its variance was usually much smaller than for the other methods (not surprising, because its randomness comes only from the scale estimate).

2. The distribution for h_{LSCV} was "centered correctly" (i.e., had a mean near h_{MISE}) but was unacceptably spread out (i.e., had too large a variance relative to other methods). This spreading was particularly bad in the direction of undersmoothing.

3. The distribution for h_{BCV} is harder to characterize because of erratic performance. It generally suffers from being quite variable, although usually not so variable as h_{LSCV}. For $n = 100$, its mean is consistently unacceptably large. For $n = 1{,}000$, its mean behavior is less predictable, sometimes being way too large and sometimes being quite close to h_{MISE}.

4. The behavior of h_{SJPI} can be viewed as a useful compromise between that of h_{ROT} and h_{LSCV}. For easy-to-estimate densities, its distribution tends to be centered near h_{MISE}, but it has much less spread than h_{LSCV}. For harder-to-estimate densities, the h_{SJPI} distribution is centered at larger values but still much smaller than h_{ROT}. But despite this "bias," it is still usually superior to h_{LSCV} because its distribution is much less variable.

5. The behavior of the smoothed bootstrap bandwidths was typically fairly close to that of h_{SJPI}. A consistent difference was that the smoothed bootstrap values tended to be consistently slightly bigger. This indicates that while the nonasymptotic goal implicit in the smoothed bootstrap approach may be more appealing, better practical performance seems to come from attempting to estimate the AMISE than the MISE.

These same general lessons agree with the simulation studies of Cao, Cuevas, and Gonzáles-Mantiega (1994) and Park and Turlach (1992) although those studies consider only "easy-to-estimate" densities.

The trade-off between "bias" and "variance" in the bandwidth distributions seems to be an intrinsic part of the performance of data-based bandwidth selectors. Less bias seems to entail more variance (h_{LSCV} is an extreme case of this), and at some cost in bias, much less variance can be obtained (h_{ROT} is the extreme here). We believe that a theory may be available to the effect that for a given amount of bias, a minimal variance is possible. Most of the methods discussed here are likely to be close to optimal in their ranges (and a number of methods not discussed here are clearly not). What seems to make the second generation bandwidth selectors so effective is that they provide a sensible trade off of this bias and variance.

[Received June 1992. Revised March 1995.]

REFERENCES

Bowman, A. W. (1984), "An Alternative Method of Cross-Validation for the Smoothing of Density Estimates," *Biometrika*, 71, 353–360.

Cao, R., Cuevas, A., and González-Manteiga, W. (1994), "A Comparative Study of Several Smoothing Methods in Density Estimation," *Computational Statistics and Data Analysis*, 17, 153–176.

Chiu, S.-T. (1992), "An Automatic Bandwidth Selector for Kernel Density Estimate," *Biometrika*, 79, 177–182.

Cook, R. D., and Weissberg, S. (1994), *An Introduction to Regression Graphics*, New York: John Wiley.

Deheuvels, P. (1977), "Estimation Nonparamétrique de la Densité par Histogrammes Généralisés," *Revue Statistique Applique*, 25, 5–42.

Engel, J., Herrmann, E., and Gasser, T. (1994), "An Iterative Bandwidth Selector for Kernel Estimation of Densities and Their Derivatives," *Journal of Nonparametric Statistics*, 4, 21–34.

Eubank, R. L. (1988), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker.

Faraway, J. J., and Jhun, M. (1990), "Bootstrap Choice of Bandwidth for Density Estimation," *Journal of the American Statistical Association*, 85, 1119–1122.

Gasser, T., Kneip, A., and Köhler, W. (1991), "A Fast and Flexible Method for Automatic Smoothing," *Journal of the American Statistical Association*, 86, 643–652.

Grund, B., Hall, P., and Marron, J. S. (1994), "Loss and Risk in Smoothing Parameter Selection," *Journal of Nonparametric Statistics*, 4, 107–132.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge, U.K.: Cambridge University Press.

Hall, P. (1980), "Objective Methods for the Estimation of Window Size in the Nonparametric Estimation of a Density," unpublished manuscript.

Hall, P., and Marron, J. S. (1987a), "Estimation of Integrated Squared Density Derivatives," *Statistics and Probability Letters*, 6, 109–115.

——— (1987b), "Extent to Which Least-Squares Cross-Validation Minimises Integrated Square Error in Nonparametric Density Estimation," *Probability Theory and Related Fields*, 74, 567–581.

——— (1987c), "On the Amount of Noise Inherent in Band-Width Selection for a Kernel Density Estimator," *The Annals of Statistics*, 15, 163–181.

——— (1991a), "Local Minima in Cross-Validation Functions," *Journal of the Royal Statistical Society*, Ser. B, 53, 245–252.

——— (1991b), "Lower Bounds for Bandwidth Selection in Density Estimation," *Probability Theory and Related Fields*, 90, 149–173.

Hall, P., Marron, J. S., and Park, B.-U. (1992), "Smoothed Cross-Validation," *Probability Theory and Related Fields*, 92, 1–20.

Hall, P., Sheather, S. J., Jones, M. C., and Marron, J. S. (1991), "On Optimal Data-Based Bandwidth Selection in Kernel Density Estimation," *Biometrika*, 78, 263–269.

Janssen, P., Marron, J. S., Veraverbeke, N., and Sarle, W. (1995), "Scale Measures for Bandwidth Selection," to appear in *Journal of Nonparametric Statistics*.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1992), "Progress in Data-Based Bandwidth Selection for Kernel Density Estimation," unpublished manuscript.

Jones, M. C., and Sheather, S. J. (1991), "Using Nonstochastic Terms to Advantage in Estimating Integrated Squared Density Derivatives," *Statistics and Probability Letters*, 11, 511–514.

Marron, J. S. (1989), "Automatic Smoothing Parameter Selection: A Survey," *Empirical Economics*, 13, 187–208.

Marron, J. S. (1992), "Bootstrap Bandwidth Selection," in *Exploring the Limits of Bootstrap*, eds. R. LePage and L. Billard, New York: John Wiley, pp. 249–262.

Marron, J. S., and Tsybakov, A. B. (1995), "Visual Error Criteria for Qualitative Smoothing," *Journal of the American Statistical Association*, 90, 499–507.

Marron, J. S., and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics*, 20, 712–736.

Müller, H. G. (1985), "Empirical Bandwidth Choice for Nonparametric Kernel Regression by Means of Pilot Estimators," *Statistics and Decisions*, Supp. Issue 2, 193–206.

——— (1988), *Nonparametric Regression Analysis of Longitudinal Data*, Berlin: Springer.

Park, B. U., Kim, W. C., and Marron, J. S. (1994), "Asymptotically Best Bandwidth Selectors in Kernel Density Estimation," *Statistics and Probability Letters*, 19, 119–127.

Park, B.-U., and Marron, J. S. (1990), "Comparison of Data-Driven Bandwidth Selectors," *Journal of the American Statistical Association*, 85, 66–72.

Park, B.-U., and Turlach, B. A. (1992), "Practical Performance of Several Data-Driven Bandwidth Selectors" (with discussion), *Computational Statistics*, 7, 251–285.

Rudemo, M. (1982), "Empirical Choice of Histograms and Kernel Density Estimators," *Scandanavian Journal of Statistics*, 9, 65–78.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, to appear.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: John Wiley.

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. (1983), "A Data-Based Algorithm for Choosing the Window Width When Estimating the Density at a Point," *Computational Statistics and Data Analysis*, 1, 229–238.

——— (1986), "An Improved Data-Based Algorithm for Choosing the Window Width When Estimating the Density at a Point," *Computational Statistics and Data Analysis*, 4, 61–65.

——— (1992), "The Performance of Six Popular Bandwidth Selection Methods on Some Real Data Sets" (with discussion), *Computational Statistics*, 7, 225–250.

Sheather, S. J., and Jones, M. C. (1991), "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society*, Ser. B, 53, 683–690.

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

Taylor, C. C. (1989), "Bootstrap Choice of the Smoothing Parameter in Kernel Density Estimation," *Biometrika*, 76, 705–712.

Terrell, G. R. (1990), "The Maximal Smoothing Principle in Density Estimation," *Journal of the American Statistical Association*, 85, 470–477.

Terrell, G. R., and Scott, D. W. (1985), "Oversmoothed Nonparametric Density Estimates," *Journal of the American Statistical Association*, 80, 209–214.

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: SIAM.

Wand, M. P., and Jones, M. C. (1994), *Kernel Smoothing*, London: Chapman and Hall.

Woodroofe, M. (1970), "On Choosing a Delta-Sequence," *Annals of Mathematical Statistics*, 41, 1665–1671.