A brief survey of bandwidth selection for density estimation

M. C. Jones, J. S. Marron, S. J. Sheather

April 3, 1994

Abstract

There has been major progress made 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, which 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 paper is intended to provide easy accessibility to the main ideas for non experts.

KEY WORDS: Bandwidth selection; Kernel density estimation; Non-

parametric curve estimation; Smoothing parameter selection.

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 27599-3260 USA (research supported by NSF Grant DMS-9203135 and the Australian Graduate School of Management). S. J. Sheather is Associate Professor, Australian Graduate School of Management, University of New South Wales, Sydney NSW 2052, Australia.

Short Title: Survey of Bandwidth selection

1. Introduction

Smoothing methods provide a powerful methodology for gaining insights into data. Many examples of this may found in the monographs of Silverman (1986), Eubank (1988), Müller (1988), Härdle (1990), Wahba (1990), Scott (1992) and Wand and Jones (1994). But effective use of these methods requires choice of a smoothing parameter. When not enough smoothing is done, the resulting density or regression estimate is too rough, and contains spurious features, which are artifacts of the sampling process. When too much 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 becomes imperative when smoothing is used by non-experts. 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, etc. are based on repeated use of smoothers. Manual choice of smoothing parameter at each step is clearly infeasible.

In this paper we focus only on the specific case of one dimensional kernel density estimation. This is done since 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 the present summary will help to stimulate work in those important related areas, and provide some guidance. See Gasser, Kneip and Köhler (1991) and Ruppert, Sheather, and Wand (1995) for first steps in the direction of kernel regression.

There have been many proposals for data based bandwidth selection methods made over the years. A few important ones are discussed in section **??**. For simple understanding of the many proposals, we group them into "first generation" and "second generation" methods, because these has been a quantum leap in terms of performance (both theoretical and practical) for a number of more recently developed methods, compared to 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 paper.

Most of the "first generation methods" are surveyed in Marron (1989). See also Silverman (1986) and Scott (1992). An exhaustive treatment of these methods is not a goal of this paper (as it would obscure our main points), but may be found in

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 paper is intended to provide quick access to the main ideas behind "second generation methods". Again for the sake of clarity, since 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, Marron and Sheather (1992). Quick access to implementation of most of the methods discussed here is provided in Park and Turlach (1992).

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 paper:

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 Marron (1989) and Jones, Marron and Sheather (1992) 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, $X_1, ..., X_n$, is often revealed through plotting the kernel density estimator,

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

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

The behavior of $\widehat{f}$ may be usefully mathematically analyzed by assuming the data are independent realizations from a probability density $f(x)$, and by viewing $\widehat{f}$ as an estimator of f. A common way of measuring the error in this estimation process is the Mean Integrated Squared Error,

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

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 section 2 of Jones, Marron and Sheather (1992), and Grund, Hall and Marron (1994) for details. Here we use $MISE$ because it is simple, and allows very deep analysis, as shown in 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 e.g. section 3.3 of Silverman (1986)), $MISE$ is asymptotically (as $n \to \infty$) approximated by the Asymptotic Mean Integrated Squared Error,

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

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:

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

This provides simple insight into "good" bandwidths. For example, smaller bandwidths are better for larger n (sensible since the estimator should be "more local" when there is more information present) and when the density is rougher (since the bias effect is stronger). Under 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 not, as indicated in Marron and Wand (1992).

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

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 Silverman (1986) and Scott (1992) for detailed discussion of some of the properties of these. Many others are surveyed in Marron (1989) and Jones, Marron and Sheather (1992).

2.2.1. Rules of Thumb

This idea goes back at least to Deheuvels (1977), and was popularized by Silverman (1986). It is to replace the unknown part of h_{AMISE}, $R(f'')$, by an estimated value based on a parametric family. Since scale is very important for bandwidth choice, but location is not, a natural choice for parametric family is $N(0, \sigma^2)$. See section 3.4.2 of Silverman (1986) and Janssen, Marron, Veraverbeke and Sarle (1995) for discussion of scale estimates. In this paper 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, Marron, Veraverbeke and Sarle (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 solve 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" which comes from using this in h_{AMISE}. Upper bounds on "a reasonable amount of smoothing" are quite useful in certain contexts, e.g. determination of a grid for searching by more sophisticated methods. However 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 Rudemo (1982) and Bowman (1984). A simple motivation comes from representing the Integrated Squared Error as

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

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 $-2n^{-1} \sum_{i=1}^{n} \widehat{f}_i(X_i)$, where $\widehat{f}_i$

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 h_{CV}. The function being minimized has fairly frequent local minima, as quantified in Hall and Marron (1991a).

2.2.3. Biased Cross-Validation

This method, proposed by Scott and Terrell (1987), attempts to directly minimize the $AMISE$. This requires estimation of the unknown $R(f'')$ which requires another bandwidth be selected. 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

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

is denoted h_{BCV}.

2.3. Second Generation Methods

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

2.3.1. Solve the Equation Plug In

There have been many papers 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 in Hall (1980) and Sheather (1983, 1986), and refined in a series of papers, with the version presented here developed in 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 is here 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, the minimizer of the Asymptotic Mean Squared Error for this problem has the form,

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

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 there are many variations that 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 show that there are gains, in terms of asymptotic rate of convergence, up to the point described explicitly above, but none 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) discuss some other variations. See Chiu (1992) and Engel, Herrmann and Gasser (1994) for other plug in ideas which are very successful.

2.3.2. Smoothed Bootstrap

One approach to this method is to consider the bandwidth which is a minimizer of a smoothed bootstrap approximation to the $MISE$. Early versions of this were proposed by Taylor (1989) and Faraway and Jhun (1990). 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 computationally as fast as other methods discussed here. In particular, following straightforward calculations, e.g. as in Marron (1992), the smoothed bootstrap estimate of the $MISE$ has the form

$$BMISE(h) = n^{-1}\left\{h^{-1}R(K) + R(K_h * \widehat{f}_g)\right\} + R(K_h * \widehat{f}_g - \widehat{f}_g)$$

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

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

An attractive feature of this approach is that it does not work through the asymptotic $AMISE$, but instead more directly targets $MISE$ itself. For choice of the pilot bandwidth g, there are a number of approaches that 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, Marron and Sheather (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, which 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 can be found in Chiu (1992).

3. Comparisons

In this section, we compare the various bandwidth selectors in three different ways. One is through applying them to real data sets. The second is through asymptotic analysis. The third is 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 are presented in 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 above, for the variable "Lean Body Mass" in the Australian Institute of Sports data set in exercise 2.4 of Cook and Weisberg (1994).

[put figure 1 about here]

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. The strongest suggestion of bimodality is given by h_{SJPI}.*

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: are two modes "really present"? Or are they just artifacts of undersmoothing 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.

[put figure 2 about here]

FIGURE 2: *Density estimates, using h_{SJPI} (for full population), based on subpopulations of $n = 102$ females and $n = 100$ males. This verifies that the bimodal structure in Figure 1d was an important feature of the data worth deeper investigation.*

The examples in Sheather (1992) and Jones, Marron, 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), i.e. 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 below 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 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.

However 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 $\widehat{h}$, it is often useful to study the asymptotic behavior of the random variable

$$\frac{\widehat{h} - h_{MISE}}{h_{MISE}},$$

(relative error is appropriate, since reasonable bandwidths tend to 0 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 in Park and Marron (1990), this measure of performance is the driving force behind more interesting measures of error, such as

$$MISE(\widehat{h}) - MISE(h_{MISE}).$$

Another alternative would be to study asymptotics based on the $ISE = \int(\widehat{f}_h - f)^2$, but as noted in Grund, Hall and Marron (1994), there is not an important practical difference between this and $MISE$ for assessing the performance of bandwidth selectors.

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

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

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 the oversmoothed bandwidths.

The main distinction between first and second generation bandwidths is in the rate of convergence. For most methods, see Jones, Marron and Sheather (1992) for references and detailed discussion,

$$\frac{\widehat{h} - h_{MISE}}{h_{MISE}} \sim n^{-p},$$

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 $n^{-1/10} = 0.1$: Note 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 for analogous versions of the smoothed bootstrap, $p = 5/14 \approx 0.36$.

A variety of 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 too. Interested readers are referred to Jones, Marron and Sheather (1992), but these points are not discussed in detail here, since 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 $n = 100$, 1000. Again to avoid confusing the main issues, we give here only a summary of the results, and the interested reader is referred to Jones, Marron and Sheather (1992).

An important concept to our results is that some of the target densities are "easy to estimate", meaning that there is enough information present in the data to recover most of the features of the target. Others in this set are "hard to estimate" in the sense that the density contains features (e.g. thin spikes) that can not 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 these:

1. The distribution for h_{ROT} had a mean that was usually unacceptably large (since this method is not unlike the oversmoother, which is usually too large). However 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 $n = 1000$, 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}. However, 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 non-asymptotic 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 Park and Turlach (1992) and Cao, Cuevas and Gonzales-Mantiega (1994), although those studies only consider "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 a theory may be available to the effect that for a given amount of bias, there is a minimal variance 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.

References

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

[2] 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.

[3] Chiu, S.-T. (1992), "An automatic bandwidth selector for kernel density estimate," *Biometrika,* 79, 177-182.

[4] Cook, R. D. and Weissberg, S. (1994), *An introduction to regression graphics,* New York: Wiley.

[5] Deheuvels, P. (1977), "Estimation non paramétrique de la densité par histogrammes généralisés," *Revue Statistique Applique,* 25, 45-42.

[6] 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.

[7] Eubank, R. L. (1988), *Spline smoothing and nonparametric regression*, New York: Dekker.

[8] Faraway, J.J. and Jhun, M. (1990), "Bootstrap choice of bandwidth for density estimation," *Journal of the American Statistical Association,* 85, 1119-1122.

[9] 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.

[10] Grund, B., Hall, P. and Marron, J. S. (1994), "Loss and risk in smoothing parameter selection," *Journal of Nonparametric Statistics,* 4, 107-132.

[11] Härdle, W. (1990), *Applied nonparametric regression*, Cambridge: Cambridge University Press.

[12] Hall, P. (1980), "Objective methods for the estimation of window size in the nonparametric estimation of a density," Unpublished manuscript.

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

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

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

[16] Hall, P. and Marron, J.S. (1991a), "Local minima in cross-validation functions," *Journal of the Royal Statistical Society, Series B,* 53, 245-252.

[17] Hall, P. and Marron, J.S. (1991b), "Lower bounds for bandwidth selection in density estimation," *Probability Theory Related Fields,* 90, 149-173.

[18] Hall, P., Marron, J.S. and Park, B.-U. (1992), "Smoothed cross-validation," *Probability Theory and Related Fields,* 92, 1-20.

[19] 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.

[20] Janssen, P., Marron, J.S., Veraverbeke, N. and Sarle, W. (1995), "Scale measures for bandwidth selection," to appear in *Journal of Nonparametric Statistics.*

[21] Jones, M. C., Marron, J. S. and Sheather, S. J. (1992), "Progress in data based bandwidth selection for kernel density estimation," Under consideration.

[22] Jones, M.C. and Sheather, S.J. (1991), "Using non-stochastic terms to advantage in estimating integrated squared density derivatives," *Statistics and Probability Letters,* 11, 511-514.

[23] Marron, J.S. (1989), "Automatic smoothing parameter selection: a survey," *Empirical Economics,* 13, 187-208.

[24] Marron, J.S. (1992), "Bootstrap bandwidth selection," In *Exploring the limits of bootstrap* (R. LePage and L. Billard, eds) 249-262, New York: Wiley.

[25] Marron, J. S. and Tsybakov, A. B. (1995), "Visual error criteria for qualitative smoothing," *Journal of the American Statistical Association* to appear.

[26] Marron, J.S. and Wand, M.P. (1992), "Exact mean integrated squared error," *Annals of Statistics,* 20, 712-736.

[27] Müller, H. G. (1985), "Empirical bandwidths choice for nonparametric kernel regression by means of pilot estimators," *Statistics and Decisions*, Supp. Issue 2, 193-206.

[28] Müller, H. G. (1988), *Nonparametric regression analysis of longitudinal data*, Berlin: Springer.

[29] 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.

[30] Park, B.-U. and Marron, J.S. (1990), "Comparison of data-driven bandwidth selectors," *Journal of the American Statistical Association,* 85, 66-72.

[31] Park, B.-U. and Turlach, B.A. (1992), "Practical performance of several data driven bandwidth selectors (with discussion)," *Computational Statistics,* 7, 251-285.

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

[33] 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.

[34] Scott, D. W. (1992), *Multivariate density estimation; theory, practice, and visualization*, New York: Wiley.

[35] Scott, D.W. and Terrell, G.R. (1987), "Biased and unbiased cross-validation in density estimation," *Journal of the American Statististical Association,* 82, 1131-1146.

[36] 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.

[37] Sheather, S.J. (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.

[38] Sheather, S.J. (1992), "The performance of six popular bandwidth selection methods on some real data sets (with discussion)," *Computational Statistics,* 7, 225-250.

[39] 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, Series B* 53, 683-690.

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

[41] Taylor, C.C. (1989), "Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika,* 76, 705-712.

[42] Terrell, G.R. (1990), "The maximal smoothing principle in density estimation. *Journal of the American Statistical Association,* 85, 470-477.

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

[44] Wahba, G. (1990), *Spline models for observational data*, Philadelphia: SIAM.

[45] Wand, M. P. and Jones, M. C. (1994), *Kernel smoothing*, London: Chapman and Hall.

[46] Woodroofe, M. (1970), "On choosing a delta-sequence," *Annals of Mathematical Statistics,* 41, 1665-1671.