Bandwidth Selection in Kernel Density Estimation: A Review

Berwin A. Turlach [†]
C.O.R.E. and Institut de Statistique
Université Catholique de Louvain
B-1348 Louvain-la-Neuve, Belgium

Abstract

Allthough nonparametric kernel density estimation is nowadays a standard technique in explorative data–analysis, there is still a big dispute on how to assess the quality of the estimate and which choice of bandwidth is optimal. The main argument is on whether one should use the *Integrated Squared Error* or the *Mean Integrated Squared Error* to define the optimal bandwidth. In the last years a lot of research was done to develop bandwidth selection methods which try to estimate the optimal bandwidth obtained by either of this error criterion. This paper summarizes the most important arguments for each criterion and gives an overview over the existing bandwidth selection methods. We also summarize the small sample behaviour of these methods as assessed in several Monte–Carlo studies. These Monte–Carlo studies are all restricted to very small sample sizes due to the fact that the numerical effort of estimating the optimal bandwidth by any of these bandwidth selection methods is proportional to the square of the sample size. This high computational cost for estimating the optimal bandwidth can be significantly reduced by binning or discretization methods. These methods will be explained and it will be shown how the presented bandwidth selectors can be implemented in a much faster way.

Keywords: Kernel density estimation, error criteria, choice of error criteria, bandwidth selection, automatic methods, cross–validation, plug–in, discretizing data, binning, fast implementations, Monte–Carlo results.

[†] This paper was revised while the author visited the Institut für Statistik and Ökonometrie, Humboldt–Universität zu Berlin

1.) Introduction

How can one estimate a probability density function $f(x)$ given a sequence of independent identically distributed random variables $X_1, \ldots, X_n$ from this density f?

The problem of estimation a probability density function $f(x)$ is interesting for many reasons. Possible applications are in the field of discriminant analysis or the estimation of functions or functionals of the density such as the hazard, or conditional rate of failure, function $f(x)/(1 - F(X))$ or average derivative estimation. By now a rich basket of nonparametric density estimators (kernel, spline, orthogonal series, and histogram) exists. For an easy access to the huge amount of literature on these estimators see the monograph of Tapia and Thompson (1978) and, for example, the overviews by Fryer (1977) and Bean and Tsokos (1980).

This work focuses on kernel density estimators as introduced by Rosenblatt (1956) and Parzen (1962). These estimators $\hat{f}_h(x)$ are defined by

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) \tag{1.1}$$

where h is called the bandwidth and K is a kernel, $K_h(u) = K(u/h)/h$. Usually the following assumptions are imposed on the kernel:

$$\begin{aligned}
&K \text{ is symmetric, i.e., } K(u) = K(-u)\\
&\textstyle\int_{I\!R} K(u)\, du = 1\\
&\textstyle\int_{I\!R} u^j K(u)\, du = 0 \text{ for } j = 1, \ldots, k-1\\
&\textstyle\int_{I\!R} u^k K(u)\, du \neq 0
\end{aligned}$$

In this case K is called kernel of order k. Note that because of the symmetry k is necessarily even and that the second assumption garanties that $\hat{f}_h(x)$ is a denstiy, i.e., $\int_{I\!R} \hat{f}_h(x)\, dx = 1$. For $k = 2$ one may choose K non–negative, i.e., K itself is a probability denstiy. The kernel density estimator $\hat{f}_h(x)$ has then the intuitively motivation that he places at each observation point X_i a probability mass according to K_h and then averages. This is visualized in Figure 1.1.

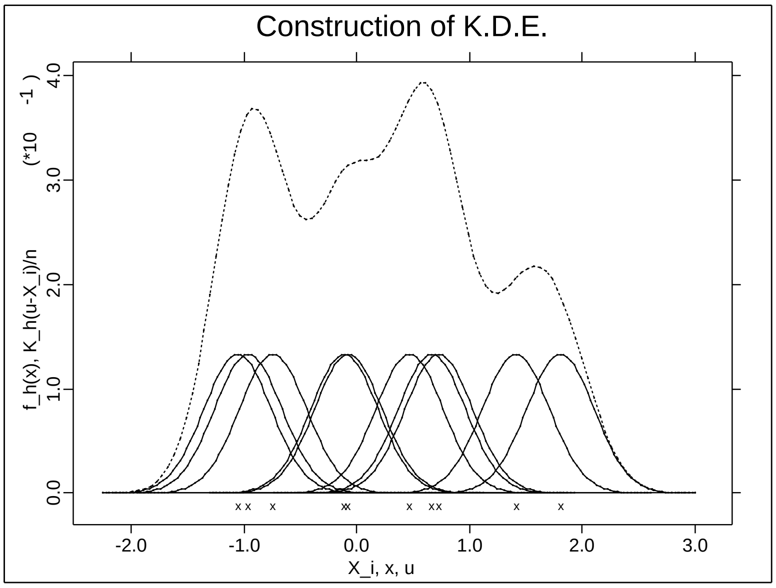

Figure 1.1: *The kernel density estimator $\hat{f}_h(x)$ (dotted line) as average over probability masses K_h (solid lines present $n^{-1}K_h$) centered at the observations X_i (crosses below x–axis).*

For $k > 2$ it is necessary that K takes negative values and thus $\hat{f}_h(x)$ does not have this intuitive motivation — it even may happen that in this case the kernel density estimate becomes negative! Thus normally kernels of order 2 are used. Some of the commonly kernels are summarized in Table 1.1.

Kernel	$K(u)$
Uniform	$\frac{1}{2}I(\lvert u\rvert \le 1)$
Triangle	$(1-\lvert u\rvert)I(\lvert u\rvert \le 1)$
Epanechnikov	$\frac{3}{4}(1-u^2)I(\lvert u\rvert \le 1)$
Quartic	$\frac{15}{16}(1-u^2)^2 I(\lvert u\rvert \le 1)$
Triweight	$\frac{35}{32}(1-u^2)^3 I(\lvert u\rvert \le 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$

Table 1.1: *Common kernels of order 2.*
$I(\bullet)$ denotes the indicator function.

A good introduction to kernel density estimation with an interesting collection of its use in data analysis is given by the monograph of Silverman (1986). It turns out that the choice of h is much more important for the behaviour of $\hat{f}_h(x)$ than the choice of K. Small values of h make the estimate look "wiggly" and show spurious features, whereas to big values of h will lead to an estimate which is too smooth in the sense that it is too biased and may not reveal structural features, like for example bimodality, of the underlying density f. This behaviour is shown in Figure 1.2 and Figure 1.3. They show the underlying bimodal density (a mixture of two normal densities $\frac{1}{2}N\left(-1,\left(\frac{4}{7}\right)^2\right)+\frac{1}{2}N\left(1,\left(\frac{4}{7}\right)^2\right)$) with estimates for different values of h based on a sample of 100 observations.

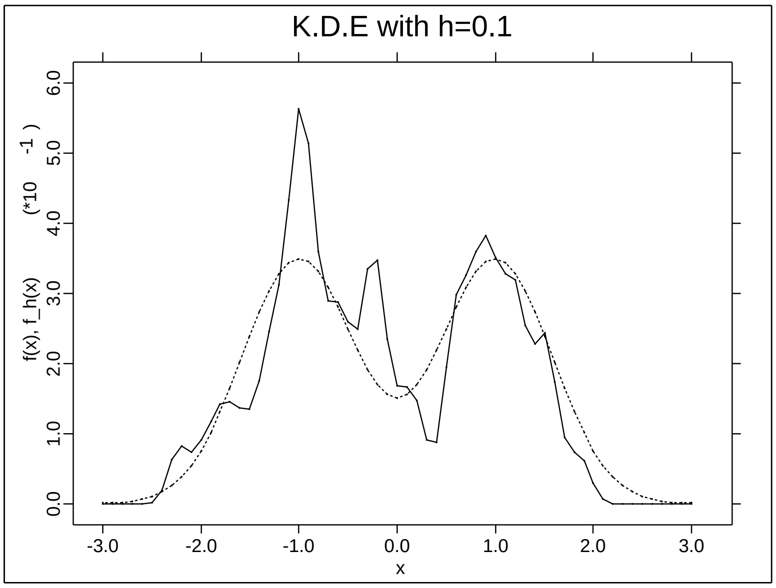

Figure 1.2: *The kernel density estimator (solid line) with the Gaussian kernel and a bandwidth of* h=0.1 *for a sample of 100 observation from the density* $\frac{1}{2}N\left(-1,\left(\frac{4}{7}\right)^2\right)+\frac{1}{2}N\left(1,\left(\frac{4}{7}\right)^2\right)$ *(dotted line)*

Terrell (1990) guesses that "most density estimates are presumably designed on aesthetic grounds: The practitioner picks the smoothing parameter so that the density looks good. Clearly, though, such individualistic approach does not lead to replicable results; nor does it lead reliably to sensible estimates from a novice." Because of such reasons a lot of reasearch was done in the last years to find objective, data–driven bandwidth selection methods. In the following sections an overview over these methods and their motivations is given. Hereby the organization of the paper is as follows. In Section 2 we present several measures which can be used to assess the goodness of the kernel density estimate. These different measures naturally lead to different definitions which bandwidth h is optimal. Section 3 discusses the pro and cons of these different bandwidths. In Section 4 we present different bandwidth selection method which aim to estimate one of the bandwidths presented in Section 2. The performance of these selectors in Monte–Carlo studies is reported in Section 5. Finally in Section 6 we describe fast implementations of these methods by using binning ideas so that even for huge sample sizes the "optimal" bandwidth can be calculated within a reasonable time. For other recent overviews see Marron (1988), Park (1991), Cao, Cuevas, and González–Manteiga (1992) and Jones, Marron, and Sheather (1992).

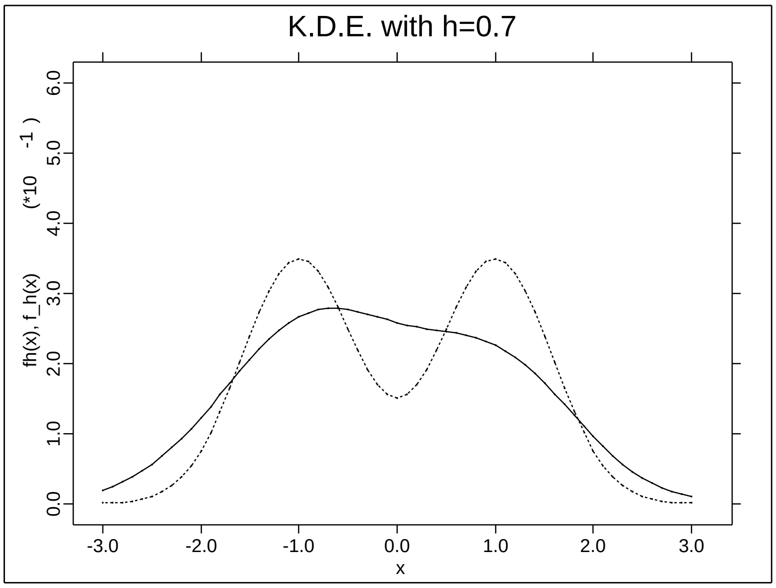

Figure 1.3: *The kernel density estimator (solid line) with the Gaussian kernel and a bandwidth of* h*=0.7 for a sample of 100 observation from the density* $\frac{1}{2}N\left(-1,\left(\frac{4}{7}\right)^2\right)+\frac{1}{2}N\left(1,\left(\frac{4}{7}\right)^2\right)$ *(dotted line)*

2.) How "good" is the estimator?

To evaluate the performance of the kernel density estimator $\hat{f}_h(x)$ it is necessary to choose a measure of distance between the true density f and the estimator $\hat{f}_h(x)$. Since L^2–distances have the advantage that they allow an easier analysis than for example L^1–distances most work was done using the former. Especially common choices are the *Integrated Squared Error (ISE)*

$$ISE(h) = \int \{\hat{f}_h(x) - f(x)\}^2\, dx \tag{2.1}$$

and its expected value, the *Mean Integrated Squared Error (MISE)*,

$$MISE(h) = E[\int \{\hat{f}_h(x) - f(x)\}^2\, dx]. \tag{2.2}$$

Here and in future the integral is to be taken over the whole real line if not indicated otherwise. One could arrive at a big variation of these distances by introducing a weight function but this is normally not done. The overview given here is restricted to work and methods based on one of these two criteria. For a treatment of the analogous L^1–criteria see the monographs of Devroye and Györfi (1984) and Devroye (1987). The question of bandwidth selection in respect to L^1–distances is addressed for example in Hall and Wand (1988a,b) and Devroye (1989).

Note that only the dependence of ISE and $MISE$ on h is reflected in this notation but not the dependence on the chosen kernel K. This is done for notational ease and it is well known that the choice of h is much more important than the choice of K. We will see later some (asymptotic) arguments for this fact.

Denote by $\hat{h}_0$ the minimizer of $ISE(h)$ and by h_0 the minimizer of $MISE(h)$. Note that $\hat{h}_0$ as $ISE(h)$ is a random variable depending on the given sample $X_1, \ldots, X_n$. Clearly, both minimizers depend on the sample size n too, i.e., $\hat{h}_0 = \hat{h}_{0,n}$ and $h_0 = h_{0,n}$ but for notational ease this dependence is surpressed here.

A further well known criterion is derived from an asymptotic analysis of $MISE(h)$. Note that $MISE(h)$ has the presentation:

$$\begin{aligned} MISE(h) &= \int E\{\hat{f}_h(x) - f(x)\}^2 \, dx \\ &= \int E\{\hat{f}_h(x) - E[\hat{f}_h(x)] + E[\hat{f}_h(x)] - f(x)\}^2 \, dx \\ &= \int E\{\hat{f}_h(x) - E[\hat{f}_h(x)]\}^2 + \{E[\hat{f}_h(x)] - f(x)\}^2 \, dx \\ &= \int Var\, \hat{f}_h(x) \, dx + \int bias^2\, \hat{f}_h(x) \, dx \end{aligned}$$

i.e., $MISE(h)$ is the sum of the integrated variance of $\hat{f}_h(x)$, $IV(h) = \int Var\, \hat{f}_h(x)\, dx$, and the integrated squared bias of $\hat{f}_h(x)$, $IB(h) = \int bias^2\, \hat{f}_h(x)\, dx$. Some straightforward analysis shows that

$$\begin{aligned} IV(h) &= \frac{R(K)}{nh} - \frac{1}{n} \int (K_h * f)^2(x) \, dx \\ IB(h) &= \int (K_h * f - f)^2(x) \, dx \\ &= \int (K_h * f)^2(x) \, dx - 2 \int (K_h * f)(x) f(x) \, dx + \int f^2(x) \, dx \end{aligned}$$

Here and in future $R(\bullet)$ denotes for any (square integrable) function L the functional $R(L) = \int L^2(x)\, dx$ and $*$ denotes the convolution of two functions K and L, $K * L(x) = \int K(x-u) L(u)\, du = \int K(u) L(x-u)\, du$.

Assume that f has at least $k+2$ (bounded or square integrable) derivatives and that K is a kernel of order k. Then by change of variables and Taylor expansion of f it is easy to show that

$$\begin{aligned} IV(h) &= \frac{R(K)}{nh} + \frac{1}{n}R(f) + O(n^{-1}h^k) \\ IB(h) &= \frac{h^{2k}}{(k!)^2}\mu_k^2(K)R(f^{(k)}) + O(h^{2k+4}) \end{aligned} \tag{2.3}$$

where $\mu_j(\bullet)$, $j \in \mathbb{N}$, denotes for any function L the functional $\mu_j(L) = \int x^j L(x)\,dx$ and $f^{(j)}$, $j \in \mathbb{N}$, is the j-th derivative of f.

Thus it is clear that for $MISE(h_0) \to 0$ with $n \to \infty$ it is necessary and sufficient that $h_0 \to 0$ such that $nh_0 \to \infty$ with $n \to \infty$. Define the *Asymptotic Mean Squared Error (AMISE)* as

$$AMISE(h) = (nh)^{-1}R(K) + h^{2k}(\mu_k(K)/k!)^2 R(f^{(k)}). \tag{2.4}$$

It follows that for h, such that $h \to 0$ and $nh \to \infty$, we have $MISE(h) = AMISE(h) + o(AMISE(h))$. By calculating $MISE(h)$ for the case where f is a mixture of normal densities and K the Gaussian kernel (i.e., k=2) Marron and Wand (1992) studied how good this approximation of $MISE(h)$ by $AMISE(h)$ is. They found that in some cases this approximation is very poor and it takes sample sizes in the millions to have a good approximation, but that in many cases this approximation is quite good. It turned out that in the cases where $AMISE(h)$ was a poor approximation for $MISE(h)$ the reason was that $IB(h)$ was poorly approximated by $h^4\mu_2^2(K)R(f^{(2)})/4$. They studied the quality of this approximation also for the case $k = 2$ when further terms in the expansion of $IB(h)$ in (2.3) are added. They state that this inclusion of higher terms does not give a big improvement. On the other side $IV(h)$ is generally very well approximated by $R(K)/nh$.

The influence of the choice of h on the density estimator $\hat{f}_h(x)$ is also demonstrated in a nice way by $AMISE(h)$. Small values of h increase the (asymptotic) variance and thus the resulting estimate $\hat{f}_h(x)$ seems "wiggly" with many spurious features if checked graphically. On the other side big values of h reduce the (asymptotic) variance of $\hat{f}_h(x)$ but increase the (asymptotic) bias thus perhaps "smoothing away" some interesting underlying features of the true density f.

Denote by h_∞ the minimizer of $AMISE(h)$. h_∞ is easily obtained from (2.4) by differentiating with respect to h and calculating the root of the derivative. This results in:

$$h_\infty = \left(\frac{R(K)(k!)^2}{2k\mu_k^2(K)R(f^{(k)})} \right)^{\frac{1}{2k+1}} n^{-\frac{1}{2k+1}}. \tag{2.5}$$

For the specially interesting case of $k = 2$ this gives

$$h_\infty = \left(\frac{R(K)}{\mu_2^2(K)R(f^{(2)})} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}. \tag{2.6}$$

Thus it is seen that the optimal rate for h_∞ is $h_\infty \sim n^{-\frac{1}{2k+1}}$ and thus $AMISE(h_\infty) \sim n^{-\frac{2k}{2k+1}}$. This rate gives the best trade off between the (asymptotic) squared bias and the (asymptotic) variance of the estimator. From the asymptotic expansion above it is clear that the same optimal rates hold for h_0 and $MISE(h_0)$.

Any of these three bandwidths $\hat{h}_0$, h_0, and h_∞ could be taken as the "optimal" bandwidth for density estimation. Up to now there is no consensus which bandwidth should be chosen and the next section gives an overview for the different arguments in this dispute. However, even if one of these three bandwidths is chosen it can not be used in practise since all of them depend on the unknown density f. Thus several methods have been developped to estimate the "optimal" bandwidth (one of the three above) from a given data set $X_1, \ldots, X_n$. This estimated bandwidth $\hat{h}$ is then used to calculate the density estimate $\hat{f}_{\hat{h}}(x)$. An overview over several proposed bandwidth selection models and their (theoretical) performance is given in Section 4.

Finally, plugging h_∞ back into $AMISE(h)$ shows that

$$\begin{aligned} AMISE(h_\infty) &= \left(\frac{2k+1}{2k} \right) \left(\frac{2k\mu_k^2(K)R(f^{(k)})R(K)^{2k}}{(k!)^2} \right)^{\frac{1}{2k+1}} n^{-\frac{2k}{2k+1}} \\ &= C_1(K)C_2(f)n^{-\frac{2k}{2k+1}} \end{aligned} \tag{2.7}$$

Thus the "optimal" Kernel K_0 which minimizes the mean integrated squared error is the one who minimizes $C_1(K)$. However, define the *efficiency* of K by

$$eff(K) = \left(\frac{C_1(K_0)}{C_1(K)} \right)^{\frac{2k+1}{k}}.$$

Intuitively this means that for large n the mean integrated squared error will be the same whether we use n observations and the kernel K or $n\ eff(K)$ observations and the optimal kernel K_0. It turns out that for $k = 2$ most commonly used kernel have an efficiency of nearly 1 (Silverman, 1986, Chapter 3.3.2). However, when estimating a density derivative $f^{(p)}$ by $\hat{f}_h^{(p)}$ the choice of the kernel plays an important role (Härdle, Marron, and Wand, 1990).

Thus for practical reasons the choice of K is not as important as the choice of h. However, the same value of h gives for different kernel K a different amount of smoothing, i.e., the picture of $\hat{f}_h(x)$ for the same h but different K will differ a lot. One method to make different kernels comparable was proposed by Marron and Nolan (1989). Essentially they advice to use instead of K the rescaled version K_δ which has the same influence on the asymptotic variance as on the asymptotic squared bias, i.e., for which $R(K_\delta) = \mu_2^2(K_\delta)$.

3.) Which bandwidth is optimal?

In the preceding section three bandwidths $\hat{h}_0$, h_0, and h_∞ were presented as possible optimal choices for density estimation. However, in practice none of them is known since they depend on the unknown density f. So the question is which of these bandwidths should be target, i.e., which one should one try to estimate given a sample $X_1, \ldots, X_n$. Such an estimate $\hat{h}$ could be used to replace the "optimal" density estimate $\hat{f}_{h_*}(x)$, $h_* \in \{\hat{h}_0, h_0, h_\infty\}$, by $\hat{f}_{\hat{h}}(x)$.

Since $\hat{h}_0$ is the minimizer of $ISE(h)$ we have for $h_* \in \{h_0, h_\infty\}$:

$$ISE(\hat{h}_0) \leq ISE(h_*) \quad \Longrightarrow \quad E[ISE(\hat{h}_0)] \leq E[ISE(h_*)] = MISE(h_*)$$

Since $ISE(h)$ measures the distance of the estimate $\hat{f}_h(x)$ from the true density f, i.e., is a "loss" function, this means that the optimal bandwidth should be $\hat{h}_0$ if the aim is the best estimation of f.

Mammen (1990) points out that whereas $ISE(h)$ has this interpretation as a loss function in a classical decision–theoretic sense, $MISE(h)$ has no such interpretation. For any bandwidth selector, i.e., a procedure which returns a bandwidth $\hat{h}$ depending only on a given sample $X_1, \ldots, X_n$, the corresponding "risk" function is $E[ISE(\hat{h})]$ which is not equal to $MISE(\hat{h})$, the latter being a random variable. However, Grund, Hall, and Marron (1992) showed that you can put $MISE(h)$ into a decision theoretical framework.

Since $AMISE(h)$ strongly depends on asymptotic arguments it has even less interpretability than $MISE(h)$. The advantage of $AMISE(h)$, as given in (2.4) is that it nicely reflects the behaviour of $\hat{f}_h(x)$ for too small values of h (too much variance, "wiggly") and for h too large (too much bias, "oversmoothing"). Nevertheless, some of the data–driven bandwidth selectors which are presented in the next section are motivated from the asymptotic representation (2.4). This is a result of thinking that h_∞ is close to h_0, but Marron and Wand (1992) showed that this is sometimes only the case for sample sizes close to one million. Thus h_∞ is rarely considered as the optimal bandwidth and the main dispute is between choosing $\hat{h}_0$ or h_0.

But since $\hat{h}_0$ is a random variable it is harder to estimate $\hat{h}_0$ than h_0. This was quantified by Hall and Marron (1991a). They proved that any data–driven bandwidth selector $\hat{h}$ has in the best case (even with f arbitrarily smooth) a relative rate of convergence to $\hat{h}_0$ of $n^{-1/10}$. On the other side the relative rate of convergence to h_0 is of (the extremely fast) order $n^{-1/2}$, i.e., in the best case we have

$$\frac{\hat{h}}{\hat{h}_0} = 1 + O_p(n^{-1/10}) \quad \text{but} \quad \frac{\hat{h}}{h_0} = 1 + O_p(n^{-1/2}).$$

Intuitively this means that a bandwidth selection method $\hat{h}$ which aims for $\hat{h}_0$ will suffer a lot from sampling variation and will have a big variance itself whereas methods aiming for h_0 are less influenced by sampling fluctuations. Therefore Hall and Marron (1991a) argue that h_0 should be used as optimal bandwidth despite the above mentioned philosophical arguments against it.

Jones (1991) favours h_0 too as the optimal bandwidth. His main arguments are:

- "Through *EISE* and other aspects of its distribution, *ISE* is much to be preferred to *MISE* as a tool for assessing any $\hat{f}_{\hat{h}}(x)$ as an estimate of $f(x)$. Relatedly, the *ISE*–optimal bandwidth $\hat{h}_0$ is a much more appropriate target that is the *MISE*–optimal bandwidth h_0. However, practical procedures based on *MISE* remain one particularly sensible way to go about choosing $\hat{h}$, even with the *ISE* target in mind."
- "The bandwidth $\hat{h}_0$ is an unrealistic target because it takes the true f too much into account. It thus seems that we can only hope that $\hat{h}$ and $\hat{h}_0$ match well when the sample at hand is 'typical' (i.e. reflects the structure) of the distribution from which it was drawn. ... It would be nice if further theory based on these 'typicality' arguments could be developed. Instead we turn this around to state that it is only reasonable to measure $\hat{h}$'s performance in terms of estimating f in an average sense."

For these reasons there is a tendancy to accept h_0 as the optimal bandwidth. But beside this controversy about $\hat{h}_0$ and h_0 the question is still open whether either of them is really a worthwile target. Jones, Marron, and Sheather (1992) state that "observing the outcomes of our practical experimentation, we have become more and more convinced of the inadequacy of the 'L_2 error', in the sense that it does not very well reflect human perceptions of when f and $\hat{f}$ are 'close'." Indeed, if the aim is to recuperate as much as possible of the structure of the underlying densities the bandwidth selectors based on the L^2–distance behave poor, probably since a global bandwidth is chosen which is inappropriate for a density with a lot of features, i.e., with many modes. A specially striking example of the inadequacy of the L^2–distance is given in Kooperberg and Stone (1991). They consider a bimodal density function and construct two different estimates (not kernel estimates) for it, one of which is unimodal and the other one is bimodal and resembles the curvature of the true density. In this example the unimodal estimate has half the integrated squared error than the bimodal one which would be preferred by most peoples. However, until some new error criteria are developed which reflect the human perception we will have to stick to $ISE(h)$, $MISE(h)$, $AMISE(h)$, and their minimizers.

4.) Bandwidth Selection Methods

The previous section summarized the advantages and disadvantages of three bandwidths which naturally arise as possible choices for an optimal bandwidth in kernel density estimation. Unfortunately, none of this bandwidths is available in practice since all of them depend on the unknown density function f. This section provides an overview over several methods which have been proposed to estimate one of these three bandwidths from a sample $X_1, \ldots, X_n$. For simplicity it is assumed that the kernel K used in the density estimator is of order $k = 2$.

a.) "Quick and Dirty" Methods

The two methods falling into this category are the *Rule of thumb* and the *Maximal Smoothing Principle*. Both are based on $AMISE(h)$, the asymptotic mean squared error (see (2.4))

$$AMISE(h) = (nh)^{-1}R(K) + h^4\mu_2^2(K)R(f^{(2)})/4$$

and the optimal bandwidth h_∞ (see (2.6)) derived from this:

$$h_\infty = \left(\frac{R(K)}{\mu_2^2(K)R(f^{(2)})}\right)^{\frac{1}{5}} n^{-\frac{1}{5}}.$$

Note that only the term $R(f^{(2)})$ is unknown in this expression. The *Rule of thumb* replaces the unknown density function f in this functional by a reference distribution function. The reference distribution function is rescaled to have variance equal to the sample variance. This method seems to date back to Deheuvels (1977), who proposed it for histograms. If, e.g., we take K as the Gaussian kernel, the standard normal distribution as reference distribution the *Rule of thumb* yields the estimate

$$\hat{h}_{rot} = 1.06\hat{\sigma}n^{-\frac{1}{5}}$$

where $\hat{\sigma}^2$ is the sample variance. A version which is more robust against outliers in the sample can be constructed if the interquartile range R is used as a measure of spread instead of the variance. This modified estimator is

$$\hat{h}_{rot} = 1.06 \min\left(\hat{\sigma}, \frac{\hat{R}}{1.34}\right) n^{-\frac{1}{5}}$$

The details for this calculus are given for example in Silverman (1986, Section 3.4.2) and in Härdle (1991, Section 4.1).

The *Maximal Smoothing Principle* for density estimation was proposed by Terrell (1990) and for histograms and frequency polygons by Terrell and Scott (1985). Terrell (1990) gives an lower bound for the functional $R(f^{(2)})$ and thus an upper bound for h_∞. He proposes to use this upper bound as estimate. In the case where the sample variance is used as measure of spread this leads to

$$\hat{h}_{MSP} = 3\,(35)^{-1/5}\,\hat{\sigma}\,\left(R(K)/\mu_2^2(K)\right)^{1/5} n^{-\frac{1}{5}}.$$

For unimodal densities these two methods seem to work fairly well. However, for multimodal densities they tend to oversmooth the data and hide the features of the underlying density which can be viewed as a drawback. But Terrell (1992) advices the use of these two methods "because they start with a sort of null hypothesis that there is no structure of interest, and let the data force us to conclude otherwise. These are my favorites: they reflect traditional statistical conservatism, ...".

b.) Cross–Validation Methods

Pseudo Likelihood Cross–Validation

This method was proposed by Habbema, Hermans, and van den Broeck (1974) and by Duin (1976). They proposed to choose h so that the pseudo–likelihood $\prod_{i=1}^{n} \hat{f}_h(X_i)$ is maximized. However this has a trivial maximum at $h = 0$, so the cross–validation principle is invoked by replacing $\hat{f}_h(x)$ in the pseudo–likelihood by the *leave–one–out* version $\hat{f}_{h,i}(x)$, where

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

It is known that the bandwidth selected by this method minimizes the Kullback–Leibler distance between $\hat{f}_h(x)$ and $f(x)$. Moreover it has some nice L^1–properties. The biggest drawback of this method is, that it yields inconsistent estimates for heavy–tailed densities, e.g., the family of student's t–distributions. For furter results and references see Marron (1985) and Cao, Cuevas, and González–Manteiga (1992).

Least Squares Cross–Validation

This method, which was proposed by Rudemo (1982) and by Bowman (1984), is probably the most popular and best studied one. It aims to estimate $\hat{h}_0$, the minimizer of $ISE(h)$. Expanding $ISE(h)$ results in

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

Since the first term is known and the last term is independent of h only the term in the middle has to be estimated. Using a method of moments estimate for this term results in the *Least Squares Cross–Validation* function

$$LSCV(h) = R(\hat{f}_h) - 2 \sum_{i=1}^{n} \hat{f}_{h,i}(X_i)$$

which is an estimator for $ISE(h) - R(f)$. The minimizer $\hat{h}_{LSCV}$ of $LSCV(h)$ is then taken as an estimate for $\hat{h}_0$. Scott and Terrell (1987) called this method *Unbiased Cross–Validation* since for fixed h, i.e., non random bandwidth, it is easy to check that $E[LSCV(h)] = E[ISE(h)] - R(f) = MISE(h) - R(f)$. Thus the *Least Squares Cross–Validation* function is an unbiased estimator for $ISE(h) - R(f)$.

The popularity of this method is due to this intuitive motivation and the fact that it is asymptotically optimal under very weak conditions (Hall, 1983 and Stone, 1984).

The relative rate of convergence of $\hat{h}_{LSCV}$ to $\hat{h}_0$ or h_0 is of order $O_p(n^{-1/10})$ (Hall and Marron, 1987a and Scott and Terrell, 1987), which is extremely slow but the best possible rate when aiming for $\hat{h}_0$ (Hall and Marron, 1991a). As one may expect from this result, *Least Squares Cross–Validation* suffers a lot under sample variation, i.e., for different samples from the same distribution the estimated bandwidths have a big variance. Moreover, Rudemo (1982) observed that $\hat{h}_{LSCV}$ and $\hat{h}_0$ seem to be negatively correlated, a fact which was quantified by Hall and Marron (1987a). Another drawback of *Least Squares Cross–Validation* is that $LSCV(h)$ often has several minima (this feature inhibits the use of Newton–Raphson like methods to search for $\hat{h}_{LSCV}$ and makes a grid search necessary), with some spurious ones often quite far over on the side of undersmoothing (Hall and Marron, 1991b). Simulation studies have shown that this problem can be fairly fixed by selecting the **largest** value of h for which a local minimum occurs. This rule is especially useful since $LSCV(h) \to -\infty$ with $h \to 0$ if the data is discretized and has several replications (see Silverman, 1986, pp. 51–52 and Chiu, 1991a). On the other hand sometimes the problem occurs that no minimum exists at all, see Sheather (1992) for occasions where this happens with real data.

Biased Cross–Validation

Biased Cross–Validatation was proposed by Scott and Terrell (1987). Consider again the asymptotic mean squared error

$$AMISE(h) = (nh)^{-1}R(K) + h^4(\mu_2(K)/2)^2 R(f^{(2)}).$$

As the *Rule of thumb* and the *Maximal Smoothing Principle* they substitute $R(f^{(2)})$ by an estimate. Instead of using a reference distribution they estimate the functional (essentially) by $R(\hat{f}_h^{(2)})$ to derive a score function *BCV(h)* which is minimized with respect to h.

Following the usual convention that rescaling is done before differentiating, i.e., $K_h^{(p)}(u) = h^{-p-1}K^{(p)}(u/h)$, straighforward manipulations show that

$$\begin{aligned} R(\hat{f}_h^{(2)}) &= \frac{1}{n}K_h^{(2)} * K_h^{(2)}(0) + \frac{1}{n^2}\sum\sum_{i\neq j} K_h^{(2)} * K_h^{(2)}(X_i - X_j) \\ &= \frac{1}{nh^5}K^{(2)} * K^{(2)}(0) + \frac{1}{n^2h^5}\sum\sum_{i\neq j} K^{(2)} * K^{(2)}(X_i - X_j) \\ &= \frac{1}{nh^5}R(K^{(2)}) + \frac{1}{n^2h^5}\sum\sum_{i\neq j} K^{(2)} * K^{(2)}(X_i - X_j) \end{aligned}$$

However, if h is of the optimal order $n^{-4/5}$ Scott and Terrell (1987, Lemma 3.2) showed that $R(\hat{f}_h^{(2)})$ is a biased estimate for $R(f^{(2)})$ since $E[R(\hat{f}_h^{(2)})] = R(f^{(2)}) + n^{-1}h^{-5}R(K^{(2)}) + O(h^2)$. Therefore they proposed to estimate $R(f^{(2)})$ by $\hat{R}(f^{(2)}) = R(\hat{f}_h^{(2)}) - n^{-1}h^{-5}R(K^{(2)})$ which leads to the score function

$$BCV(h) = \frac{R(K)}{nh} + h^4 \frac{\mu_2^2(K)}{4n^2} \sum\sum_{i \neq j} K_h^{(2)} * K_h^{(2)}(X_i - X_j).$$

Scott and Terrell (1987) proposed to use the minimizer $\hat{h}_{BCV}$ of $BCV(h)$ as bandwidth. They showed that $\hat{h}_{BCV}$ has the same relative rate of convergence to $\hat{h}_0$ as $\hat{h}_{LSCV}$ but that the constant is often much smaller. The problem of multiple minima is also less often observed for $BCV(h)$ than for $LSCV(h)$. However, in the case of multiple minima simulation studies have shown that the best performance is obtained by choosing the **smallest** value of h for which a local minimum occurs. Cao, Cuevas, and González–Manteiga (1992) point out, that with commonly used kernels $\lim_{h \to 0+} BCV(h) = \infty$, $\lim_{h \to \infty} BCV(h) = 0$ and in case that $4R(K) - \mu_2^2(K)K(0) > 0$ (which is the case for the Gaussian kernel) $BCV(h) > 0$ for all $h > 0$. Thus in this case no global minimum exists!

Smoothed Cross–Validation

Smoothed Cross–Validation as proposed by Hall, Marron, and Park (1992) uses the representation $MISE(h) = IV(h) + IB(h)$ from Section 2. Marron and Wand (1992) showed that $(nh)^{-1}R(K)$ is a good estimator for $IV(h)$ and Hall, Marron, and Park (1992) use this term as $\widehat{IV}(h)$. To estimate $IB(h) = \int (K_h * f - f)^2(x)\, dx$ they propose to estimate f in $IB(h)$ by $\hat{f}_g(x)$, where $\hat{f}_g(x)$ is a second density estimator with a possibly different bandwidth g and kernel L. Straightforward calculation shows that

$$\widehat{IB}(h) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (K_h * K_h - 2K_h + K_0) * L_g * L_g(X_i - X_j)$$

where K_0 denotes the Dirac function. Deleting the diagonal terms $i = j$ as in BCV and using the approximation $n \approx n - 1$ they derive the score function

$$SCV(h) = \frac{R(K)}{nh} + \frac{1}{n(n-1)} \sum\sum_{i \neq j} (K_h * K_h - 2K_h + K_0) * L_g * L_g(X_i - X_j).$$

The name *Smoothed Cross–Validation* is motivated form the fact that $LSCV(h)$ can be written as (again using $n \approx n-1$)

$$LSCV(h) = \frac{R(K)}{nh} + \frac{1}{n(n-1)} \sum\sum_{i \neq j} (K_h * K_h - 2K_h)(X_i - X_j)$$
$$= \frac{R(K)}{nh} + \frac{1}{n(n-1)} \sum\sum_{i \neq j} (K_h * K_h - 2K_h + K_0)(X_i - X_j)$$

where the last equality holds if there are no duplication in the sample (which happens with probability one for continuous data). Thus $SCV(h)$ can be viewed as a version of $LSCV(h)$ where the differences $X_i - X_j$ are presmoothed. $SCV(h)$ can also be motivated by bootstrapping ideas (see Hall, Marron, and Park, 1992 and Cao, Cuevas, and González–Manteiga, 1992).

Hall, Marron, and Park (1992) proposed to use $\hat{h}_{SCV} = \hat{h}_{SCV}(g)$ the minimizer of $SCV(h)$. They showed that under proper choices of g and L the relative rate of convergence of $\hat{h}_{SCV}$ to h_0 is of order $O_p(n^{-1/2})$. This is the best rate achievable as Hall and Marron (1991a) have shown. To achieve this rate L must be of order 6 at least.

Bandwidth Factorized Smoothed Cross–Validation

In their proposal for *Smoothed Cross–Validation* Hall, Marron, and Park (1992) deleted the diagonal terms in $\widehat{IB}(h)$. Their choice for g was $\hat{C}n^{-\alpha}$ where $\hat{C}$ is an estimate of C and the asymptotically optimal choice for g is $Cn^{-\alpha}$. Thus their choice of g was independent of h (they only considered the case $g = h$). Jones, Marron, and Park (1991) use the same score function $SCV(h)$ but investigated what happens if the diagonal terms are reintroduced and g is allowed to depend on h, i.e., they choose $g = Cn^p h^m$ for several combination of m and p. Let

$$JMP(h) = \frac{R(K)}{nh} + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (K_h * K_h - 2K_h + K_0) * L_g * L_g(X_i - X_j)$$

denote the score function and $\hat{h}_{JMP}$ its minimizer. The most important result of their study was that with the choice $g \sim n^{-23/45} h^{-2}$ it is possible to make $\hat{h}_{JMP}$ relative root n convergent even if both K and L are of order 2. Until then several relative root n convergent method were already known, however, all of them used at some point a kernel of higher order (or Fourier transformations which were essentially equivalent to the use of higher order kernels). This is important, since the performance of higher order kernel in practice, as already mentioned above and shown by Marron and Jones (1992), is poor in small sample situations even though their asymptotic properties make them by far superior to second order kernels.

c.) Plug–in Methods

Early versions

The idea of Plug–in methods goes back to Woodroofe (1970). They are based on the asymptotically best choice of h given in (2.6). As mentioned above the only unknown quantity is the functional $R(f^{(2)})$. Woodroofe (1979) proposed to use a bandwidth h_1 to calculate $\hat{f}_{h_1}(x)$, take this estimate to calculate $\hat{R}(f^{(2)}) = R(\hat{f}_{h_1}(x))$, and to plug $\hat{R}(f^{(2)})$ into (2.6) to obtain h_2, the final bandwidth.

Scott, Tapia, and Thompson (1977) proposed to iterate this process, i.e., calculate $\hat{R}(f^{(2)}) = R(\hat{f}_{h_2}(x))$, plug $\hat{R}(f^{(2)})$ into (2.6) to obtain h_3 etc., until h_i converges. Note the difference of this method to *Biased Cross–Validation*, the latter method uses the estimate $\hat{R}(f^{(2)})$ in $AMISE(h)$ given by (2.4) and than minimizes the resulting function $BCV(h)$ over h. The method of Scott, Tapia, and Thompson (1977) (and the other methods described below) first take the analytic minimizer h_∞ of $AMISE(h)$ as if $\hat{R}(f^{(2)})$ **would not** depend on h. Then they plug in $\hat{R}(f^{(2)})$ so that the right side depends on h and search for a fix point. A minor difference is that *Biased Cross–Validation* uses the diagonal–out version $\hat{R}(f^{(2)})$ described above whereas Scott, Tapia, and Thompson (1977) uses the diagonal–in version $R(\hat{f}_h^{(2)})$ to estimate $R(f^{(2)})$. The small sample behaviour of this method is described in Scott and Factor (1981).

Park and Marron Plug–In

Hall and Marron (1987b) studied the possibility to use $R(\hat{f}_h^{(p)})$ as estimator for $R(f^{(p)})$. But using $R(\hat{f}_h^{(p)})$ as an estimator yields a constant term $n^{-1}K_h^{(p)} * K_h^{(p)}(0) = n^{-1}h^{-2p-1}R(K^{(p)})$ which does not depend on the sample. Assuming that this term adds to the bias of the estimate they proposed to estimate $R(f^{(p)})$ by

$$\hat{R}(f^{(p)}) = R(\hat{f}_h^{(p)}) - \frac{R(K^{(p)})}{nh^{2p+1}}.$$

One important result of Hall and Marron (1987b) was that the optimal rate of h for estimating $R(f^{(p)})$ is different than the optimal rate for estimating $f(x)$.

Park and Marron (1990) used these results to modify the plug–idea. They estimated $R(f^{(2)})$ by $\hat{R}(f^{(2)}) = R(\hat{f}_g^{(p)}) - \frac{R(K^{(p)})}{ng^{2p+1}}$ with g having the optimal rate given in

Hall and Marron (1987b). Combining the formula for the optimal rate of h and of g they obtained the relation $g = C(f,K)h^{10/13}$. To estimate the unknown constant $C(f,K)$ they proposed to use the normal distribution as reference distribution. Using the variance as measure of spread this yields $g = \hat{C}(f,K)h^{10/13} = C(\varphi_{\hat{\sigma}^2},K)h^{10/13} = g(\hat{\sigma}^2,h)$ where $\hat{\sigma}^2$ denotes the sample variance and φ the density of the standard normal distribution. Thus they proposed to use the bandwidth $\hat{h}_{PM}$ obtained as solution of

$$h = \left(\frac{R(K)}{\mu_2^2(K)\hat{R}(\hat{f}^{(2)}_{g(\hat{\sigma}^2,h)})}\right)^{1/5} n^{-1/5} \tag{4.1}$$

where $\hat{R}(\hat{f}^{(2)}_{g(\hat{\sigma}^2,h)}) = R(\hat{f}^{(2)}_{g(\hat{\sigma}^2,h)}) - \frac{R(K^{(p)})}{ng(\hat{\sigma}^2,h)^{2p+1}}$. Park and Marron (1990) showed that $\hat{h}_{PM}$ has a relative rate of convergence to h_0 of order $O_p(n^{-4/13})$. This was until then the fastest rate achieved by a bandwidth selection method. Beside this the performance of $\hat{h}_{PM}$ in simulation studies was very good, too.

Write (4.1) in the form $PI(h) = 0$. Then $\hat{h}_{PI}$ is the root of $PI(h)$ and can be found by a root finding algorithm. However, due to the deletion of the diagonal terms the estimator $\hat{R}(f^{(2)})$ may become negative, especially for small bandwidths. Thus it may happen that $PI(h)$ is negative for small h, makes a jump from $-\infty$ to ∞ at a point h_* and is decreasing afterwards, having a root $h_{**} > h_*$. An example for this is given in Sheather and Jones (1991). This behaviour may be misleading for some root–finding algorithm which take h_* as the root and thus as $\hat{h}_{PI}$.

Sheather and Jones Plug–In

Jones and Sheather (1991) reconsidered the problem of estimating the functionals by $R(f^{(p)})$. They demonstrate that the constant term due to the diagonal terms has the opposite sign than the leading bias term of the estimator proposed by Hall and Marron (1987b). Thus by reintroducing this term and estimating $R(f^{(p)})$ by $R(\hat{f}_h^{(p)})$ they improved the mean squared error of this estimation step. A further advantage of taking $\hat{R}(f^{(2)}) = R(\hat{f}_h^{(p)})$ is that the estimate is always positive, which is natural since a positive quantity is estimated.

Sheather and Jones (1991) applied this idea to bandwidth selection in density estimation. Using the same approach as Park and Marron (1990) but replacing the diagonal–out estimator of $R(f^{(2)})$ by the diagonal–in version they find that the optimal choice for g is $g = C(K,L)(R(f^{(2)})/R(f^{(3)}))^{1/7}h^{5/7}$. Here L is the kernel used to estimate the term $R(f^{(2)})$ in (2.6) which may differ from K. However, it turns out that using a reference distribution at this step is not sufficient to achieve the cancellation effect wished by reintroducing the diagonal terms. Thus $R(f^{(2)})$ and $R(f^{(3)})$ are estimated via $R(\hat{f}_a^{(2)})$ and $R(\hat{f}_b^{(3)})$, where a and b are chosen according to the asymptotic optimal value (see Jones and Sheather, 1991) and only at this second step f is replaced by the normal distribution as reference distribution. Denote by h_{SJ} the minimizer of (4.1) with the above modifications. Sheather and Jones (1991) showed that this selector has a relative order of convergence to h_0 of $O_p(n^{-5/14})$ thus still improving on h_{PM}.

Note that the implementation of h_{SJ} given in Section 5 of Sheather and Jones (1991) is only valid if K and L are Gaussian kernels. In the general case the constant in front of $\hat{\alpha}_2(h)$ should be $D_1(L)(\sigma_K^4/R(K))^{1/7}$ (notation as in original paper).

Hall, Sheather, Jones, and Marron Plug-In

This bandwidth selection method proposed by Hall, Sheather, Jones, and Marron (1991) is a relative root$-n$ convergent method. The main idea is the use of a kernel K with order 2 but to take one further term in the asymptotic expansion of the integrated squared bias which yields

$$AMISE_2(h) = (nh)^{-1}R(K) + h^4\mu_2^2(K)R(f^{(2)})/4 - h^6\mu_2(K)\mu_4(K)R(f^{(3)})/24.$$

Unlike the minimizer of $AMISE(h)$ which is given by (2.5) respectively (2.6), the minimzer of $AMISE_2(h)$ is not easily calculated analytically since this involves the finding of a root of a polynomial of degree 7. Hall, Sheather, Jones, and Marron (1991) proposed to use instead the bandwidth given by

$$h_{HSJM} = (\hat{J}_1/n)^{1/5} + \hat{J}_2(\hat{J}_1/n)^{3/5}$$

where $\hat{J}_1 = R(K)/(\mu_2^2(K)\hat{R}(f^{(2)}))$ and $\hat{J}_2 = \mu_4(K)\hat{R}(f^{(3)})/(20\mu_2(K)\hat{R}(f^{(2)}))$. This bandwidth is asymptotically equivalent to a minimizer of $AMISE_2(h)$. To achieve the relative root n convergence higher order kernels and the ideas of Jones and Sheather (1991) are used to estimate the functionals $R(f^{(2)})$ and $R(f^{(3)})$. For details see Hall, Sheather, Jones, and Marron (1991).

To see that h_{HSJM} is asymptotically equivalent to a minimizer of $AMISE_2(h)$

observe that

$$\begin{aligned}
\frac{\partial}{\partial h} AMISE_2(h) &= -\frac{1}{nh^2} R(K) + h^3 \mu_2(K) R(f^{(2)}) - h^5 \mu_2(K) \mu_4(K) R(f^{(3)})/4 \\
h^3_{HSJM} &\approx \hat{J}_1^{3/5} n^{-3/5} + 3\hat{J}_2 \hat{J}_1 n^{-1} \\
h^5_{HSJM} &\approx \hat{J}_1 n^{-1} \\
n^{-1} h^{-2}_{HSJM} &\approx \hat{J}_1^{-2/5} n^{-3/5} - 2\hat{J}_2 n^{-1}
\end{aligned}$$

Thus $\frac{\partial}{\partial h} AMISE_2(h)\Big|_{h_{HSJM}} \approx 0.$

d.) Other methods and variations

A further interesting and according to some simulation studies promising method was proposed by Chiu (1991b) and modified by Chiu (1992). Chiu (1991b) approaches the problem of bandwidth selection via the Fourier transformation of the sample characteristic function, i.e., by a frequency analysis. To achieve a better bandwidth selection rule he proposed to modify the sample characteristic function beyond some cut–off frequency. Chiu (1992) modifies this by choosing the cut–off frequency by a crossvalidation criteria and thus giving a fully automatic bandwidth selection method. Taylor (1989) and Faraway and Jhun (1990) proposed bootstrap method to determine the optimal bandwidth. But these ideas are closely related to *Smoothed Cross–Validation.*

Apart from these methods many variations of the methods described above are possible. Sheather and Jones (1991) for example report that they used the diagonal–in version of the estimator for $R(f^{(2)})$ also in $AMISE(h)$ as given in (2.4) and then minimized with respect ot h, thus actually a variation of *Biased Cross–Validation.* A second variation they considered was to take g only depending on n and not on h. However this two methods turned out to be inferior to the one presented above. Nevertheless, the question how g should depend on h in general remains an open question. An account on possible choices is given in Marron (1991).

Another source of variation is the number of pilot estimation steps used, e.g., for $\hat{h}_{PM}$ one pilot estimation was necessary, namely $\hat{R}(\hat{f}_g^{(2)})$. The dependence of g on f was resolved by referring to the normal distribution. However for h_{SJ} this was no longer sufficient, here the functionals of f on which g is depending are estimated via kernel density estimators too, only referring in this second step to the normal distribution. This is necessary since this method uses the diagonal–in terms to cancel some terms in the asymptotic bias and thus needs estimates of the constants which have a certain convergence rate (this is necessary for all methods using diagonal–in ideas). But nothing forbids us to iterate this process arbitrarily often and using a reference distribution only in the fourth, fifth, or any higher step. Park and Marron (1992) give a deeper analysis of this option for $JMP(h)$ and $SCV(h)$. They show that there is a certain trade–off between better bias behaviour and worser variance behaviour of the bandwith selectors when using an increasing number of pilot estimation steps. But they did not find a concise answer how many steps should be used. This is still an open question.

e.) The "optimal" method

Fan and Marron (1992) derived a "Fischer" like lower bound of the relative errors of bandwidth selectors which is given by

$$\sigma^2(f) = \frac{4}{25}\left(\frac{\int f^{(4)}(x)^2 f(x)\,dx}{R(f^{(2)})} - 1\right)$$

Using this and the relative order of convergence to h_0 as criteria, the best a bandwidth selector $\hat{h}$ can do is

$$n^{1/2}(\hat{h}/h_0 - 1) \quad \xrightarrow{\mathcal{L}} \quad N(0, \sigma^2(f)).$$

The selectors proposed by Chiu (1991b) and Hall, Sheather, Jones, and Marron (1991) have this asymptotic property. However, they have the drawback that they use unappealing higher order kernels respectively Fourier transformations equivalent to this use. The only root–n convergent methods which uses only second order kernels until now is *Bandwidth Factorized Smoothed Cross–Validation*, however, this method has a larger asymptotic variance. Park, Kim, and Marron (1991) propose a method which combines both, using only second order kernels and achieving the optimal rate of convergence and the optimal asymptotic variance. The method is based on a exact expression for $MISE(h)$ in which an additional function (namely $f(-x)$) is introduced. However, Park, Kim, and Marron (1991) report poor behaviour in simulation studies, sometimes as bad as *Least Squares Cross–Validation*. Thus this method is mainly a theoretical toy, demonstrating that best rate of convergence and best asymptotic variance are achievable with only second order kernels. For a method which achieves this and has good small sample behaviour the search is still going on.

5.) Simulation Studies

The previous section described several bandwidth selection methods and their theoretical behaviour. This theoretical results are, however, of asymptotic nature and it remains the question how these methods perform in small samples. One way to access the small sample behaviour is through a simulation study. Since in the last years computer power became cheaper at an enormous rate it is nowadays feasible to perform such studies. This section gives an overview of such studies and the lessons learned by them. Another method is to compare the behaviour of the bandwidth selection methods on real data sets. This approach has the immediate drawback that the real density function f is not known, which makes it difficult to choose criteria for the performance of the bandwidth selection methods. A recent study taking this approach using many of the bandwidth selection methods discussed before and several data sets was done by Sheather (1992).

Practically each paper which proposes a new bandwidth selection method contains or reports of a small simulation study which compares the performance of the new method with that of one or two other methods on two to four densities. Thus many papers mentioned in the section before could be cited again. It seems that until now no simulation study has been performed which compares all of the bandwidth selection methods presented in Section 4 on a broad variety of densities. The most extensive study in this direction is done by J.S. Marron, but unfortunately most of his results are yet unpublished. Some results can be found in Marron (1989) and Jones, Marron, Sheather (1992). S.T. Chiu made some extensive simulations too which showed the good behaviour of the method proposed in Chiu (1991b) and Chiu (1992). The complete results of this study are unpublished too. Further recent simulation studies were done by Cao, Cuevas and González–Manteiga (1992) and by Park and Turlach (1992). Neither of these compares all the methods presented in Section 4, but Cao, Cuevas and González–Manteiga (1992) included some methods which have not been discussed here, as for example the double kernel method proposed by Devroye (1989) for bandwidth selection optimal with respect to a L^1–distance. Earlier studies were performed by Scott and Factor (1981) and Bowman (1985).

However, neither of these studies give a clear answer which bandwidth selection method is the best. Regardless of the questions how representative the densities used in the study are or how sensible the criteria for assessing the performance of the bandwidth selection methods were chosen, it is seen that no bandwidth selection method outperforms the other equally well over all densities considered. The recently proposed bandwidth selection methods, such as $\hat{h}_{SCV}$ or $\hat{h}_{JMP}$, show a better variance behaviour than $\hat{h}_{LSCV}$ for example, as predicted by theory through their fast relative rate of convergence. But they seem to pay for this by a bias behaviour which is worse and makes them in some cases unacceptable. For these reasons Jones, Marron, and Sheather (1992) choose $\hat{h}_{SJ}$ as their favorite bandwidth selection method which is not a relative root n convergent method, whereas Marron (1992) favours a version of $\hat{h}_{JMP}$ which is relative root n convergent. On the other side, Terrell (1992) argues to use a kernel of order 4 together with maximal smoothing which would lead to an integrated squared error of the order $n^{-8/9}$ (see (2.7)). His conclusion is based on the (philosophical) point that for the other methods it is not clear "what question these methods are in answer to" (see also Section 3, $MISE(h)$ $vs.$ $ISE(h)$) and that there is no clear winner in the simulation studies. But as pointed out in Section 1 kernels of order 4 take necessarily negative values and thus it is not clear what they are doing. Furthermore, Marron and Wand (1992) showed that it is not worthwile to use a higher order kernel since it requires big sample sizes before the asymptotic superiority of higher order kernels are noticeable.

6.) Improving the Computational Speed

In Section 4 several data–driven methods to choose the bandwidth h have been described. Most of these methods involve the evaluation of the estimator for several bandwidths $h_1, \ldots, h_k$. This leads to $O(n^2k)$ evaluations of the kernel. It is clear that this number grows very fast. The simulation for the six density in Park and Turlach (1992) using direct implementation of the bandwidth selectors took nearly four days for a sample size of $n = 100$ and 500 samples per density.

Schmitz (1989) used Least–Squares Cross–Validation and Biased Cross–Validation to analyze the distribution of net income in Great Britain for the years 1968–1983. For each year the sample size is roughly 7000 data points. In his case the calculations for the Least–Squares Cross–Validation function with 100 bandwidths and the data of one year took nearly seven hours on a mainframe computer of the late 80's. Schmitz (1989) reports a drastic reduction of computation time by using discretization methods.

The idea of discretizing the data goes back to Silverman (1982) who calculated the density estimate via a Fourier transformation, thus implicitly discretizing the data. An intuitively easier approach to the idea of discretization is given by the Averaging of Shifted Histograms (ASH, see Scott, 1985). Härdle and Scott (1992) proposed Weighted Averaging of Rounded Points (WARPing) to make calculations in nonparametric density and regression estimation faster. Fan and Marron (1993) investigate in an extensive Monte–Carlo study the benefits of using such ideas in nonparametric curve estimation.

How can these ideas be used to speed up the calculations? Define the discretization or prebinning by

The bins:	$B_z = B_z(x_0, \delta) = [x_0 + (2z-1)\frac{\delta}{2}, x_0 + (2z+1)\frac{\delta}{2})$	$z \in \mathbb{Z}$
The bincenters:	$b_z = x_0 + z\delta$	$z \in \mathbb{Z}$
The bincounts:	$n_z = \#\{i : X_i \in B_z\} = \sum_{i=1}^n I_{B_z}(X_i)$	

where $I_A(\bullet)$ denotes the indicator function for the set A. Essentially this operation means that we prebin the observation in a histrogram (centered at x_0) with binwidth δ. This binwidth δ should be chosen sufficiently small, especially smaller than the smallest bandwidth h which is considered. Another way to create the bincounts was proposed by Jones and Lotwick (1984) which is called "linear binning":

$$n_z = \sum_{i=1}^{n} \left\{ I_{(b_{z-1}, b_z]}(X_i) \frac{X_i - b_{z-1}}{\delta} + I_{(b_z, b_{z+1}]}(X_i) \frac{b_{z+1} - X_i}{\delta} \right\}.$$

Note that now n_z is no longer integer valued, but still $\sum_{z \in \mathbb{Z}} n_z = n$ holds. Jones (1989) showed that it is preferable to use this kind of prebinning. Still other binning schemes are proposed by Hall and Wand (1993) who also show their benefits in density estimation.

To calculate the bandwidth selectors described in Section 4 it is necessary to estimate $R(f^{(p)})$ for some p, e.g., for $BCV(h)$ we need $R(f^{(2)})$ and $\hat{h}_{SJ}$ requires $R(f^{(2)})$ and $R(f^{(3)})$. (actually $\hat{h}_{SJ}$ needs **two** estimates of $R(f^{(2)})$). Thus it is necessary to calculate these quantities fast. One way would be to estimate $f^{(p)}(x)$ by $\hat{f}_h^{(p)}(x)$ which is in turn approximated by a binned version $\hat{f}^{(p)}$ calculated from the discretized data as described in Turlach (1993). A numerical integration of the square $\hat{f}^{(p)}(x)^2$ would then yield an estimate for $R(f^{(p)})$.

Another method to estimate $R(f^{(p)})$ was pointed out by J.S. Marron. As proposed by Hall and Marron (1987b) and by Jones and Sheather (1991), a natural estimator for $R(f^{(p)})$ is $R(\hat{f}_g^{(p)})$ (here L denotes the kernel used in $\hat{f}_g^{(p)}$). Their proposals only differ in whether the diagonal term should be used or not, but this results in a difference of $\Delta = n^{-1}g^{-2}L_g^{(p)} * L_g^{(p)}(0) = n^{-1}g^{-2p-1}L^{(p)} * L^{(p)}(0)$. Observe now that

$$\begin{aligned} R(\hat{f}_g^{(p)}) &= \frac{(-1)^p}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} L_g^{(p)} * L_g^{(p)} (X_i - X_j) \\ &= \frac{(-1)^p}{n^2 g^{2p+1}} \sum_{i=1}^{n} \sum_{j=1}^{n} L^{(p)} * L^{(p)} \left(\frac{X_i - X_j}{g} \right) \end{aligned}$$

The idea of discretizing is now to replace the original observation X_i by the bincenter b_z of the bin into which X_i falls. Thus it is no longer necessary to evaluate all $\frac{n(n-1)}{2}$ differences $X_i - X_j$, and the difference between bincenters is always an integer multiple of δ. It is thus only necessary to evaluate $L_g^{(p)} * L_g^{(p)}(u)$ on a grid $k\delta$, $k = 0, \pm 1, \pm 2, \ldots, \pm l$. Define $L_{i-j} = L_g^{(p)} * L_g^{(p)}((i-j)\delta)$ and let i' and j' denote the indices of the bins into which X_i respectively X_j fall. In this case $L_g^{(p)} * L_g^{(p)} (X_i - X_j)$ in the above summation is replaced by $L_{i'-j'}$. On the other side, for given bin indices i' and j' it is clear that $n_{i'} n_{j'}$ combinations of observations X_i and X_j (i not necessarily different from j) exist such that $L_g^{(p)} * L_g^{(p)} (X_i - X_j)$ is replaced by $L_{i'-j'}$ in the above summation. Thus we get the binned estimator for $R(f^{(p)})$:

$$\begin{aligned} \hat{R}(\hat{f}_g^{(p)}) &= \frac{(-1)^p}{n^2} \sum_{i' \in \mathbb{Z}} \sum_{j' \in \mathbb{Z}} n_{i'} n_{j'} L_{i'-j'} \\ &= \frac{(-1)^p}{n^2} \sum_{i' \in \mathbb{Z}} n_{i'} \sum_{j' \in \mathbb{Z}} n_{j'} L_{i'-j'}. \end{aligned}$$

Clearly the sum over j' is a convolution. Thus in higher level statistical language packages, as for example XploRe and GAUSS, this sum is easily calculated with the build–in convolution command. The sum over i' is now only the pointwise multiplication of the vector containing the bincounts with the vector which contains the result of the convolution and the sum over all these multiplications. This operation is also easy to implement in the matrix–oriented languages mentioned above.

Most of the bandwidth selection methods described in Section 4 were used in the simulation study by Park and Turlach (1992). Below the formula for the version used in their simulation study is given. Here φ denotes the Gaussian kernel which is used because of the properties he has with respect to convolution and derivation (see Aldershof, Marron, Park, and Wand, 1990). The required formulæ for evaluating the derivatives of φ used below are:

$$\varphi_g^{(p)}(x) = g^{-(p+1)}\varphi^{(p)}(x/g)$$

$$\varphi^{(p)}(x) = (-1)^p H_p(x)\varphi(x)$$

and

$$H_p(x) = xH_{p-1}(x) - (p-1)H_{p-2}(x) \qquad p = 2, 3, \ldots$$

where $H_0(x) \equiv 1$ and $H_1(x) \equiv x$ ($H_p(\bullet)$ is the pth Hermite polynomial).

With the help of these relations it is clear how to evaluate the double sums in the formulæ of the bandwidth selectors using the binning ideas presented above. Thus it is very easy to program fast implementations of these bandwidth selectors. These discretized methods have been implemented in XploRe and GAUSS, an implementation for Splus is still in work. The GAUSS code is on request available from the author. Each algorithm can be used to compute the values of the corresponding function on a grid of h–values using an interpolating algorithm to find the local minima respectively roots of the objective function.

Least Squares Cross-Validation

$$LSCV(h) = \frac{1}{2\sqrt{\pi}nh} + \frac{1}{n(n-1)}\sum\sum_{i\neq j}\left\{\varphi_{\sqrt{2}h}(x_i - x_j) - 2\varphi_h(x_i - x_j)\right\}$$

Find:

$$\hat{h} = \arg\min_h CV(h)$$

Biased Cross-Validation

$$BCV(h) = \frac{1}{2\sqrt{\pi}nh} + \frac{1}{4}h^4 n^{-2}\sum\sum_{i\neq j}\varphi^{(4)}_{\sqrt{2}h}(x_i - x_j)$$

Find:

$$\hat{h} = \arg\min_h BCV(h)$$

Smoothed Cross-Validation

$$SCV(h) = \frac{1}{2\sqrt{\pi}nh} + \frac{1}{n(n-1)}\sum\sum_{i\neq j}\left\{(\varphi_{\sqrt{2h^2+2g^2}} - 2\varphi_{\sqrt{h^2+2g^2}} + \varphi_{\sqrt{2}g}\right\}(x_i - x_j)$$

with $g = \lambda\left(\frac{21}{40\sqrt{2}}\right)^{1/13} n^{-2/13}$ and λ the standard deviation of the sample. Now find:

$$\hat{h} = \arg\min_h SCV(h)$$

Bandwidth Factorized Smoothed Cross-Validation

$$\begin{aligned} JMP(h) = {} & \frac{1}{2\sqrt{\pi}nh} + \frac{1}{n(n-1)}\sum\sum_{i,j}\varphi_{\sqrt{2h^2+2g(h)^2}}(x_i - x_j) \\ & + \frac{1}{n(n-1)}\sum\sum_{i,j}\left\{-2\varphi_{\sqrt{h^2+2g(h)^2}} + \varphi_{\sqrt{2}g(h)}\right\}(x_i - x_j) \end{aligned}$$

with $g(h) = \hat{C}n^{-23/45}h^{-2}$. Where $\hat{C}$ is an estimate for the constant C which depends on f, K and L. We use the $N(0, \hat{\sigma}^2)$ reference for the unknown f in C and $\hat{\sigma}$ is the sample standard deviation. Thus calculate $\hat{C}$ by:

$$\begin{aligned} a &= \hat{\sigma}\left(\frac{128}{9\sqrt{2}n}\right)^{\frac{1}{11}} \\ b &= \hat{\sigma}\left(\frac{32}{5\sqrt{2}n}\right)^{\frac{1}{7}} \\ \hat{R}_a(f^{(4)}) &= \frac{1}{n^2}\sum\sum_{i,j}\varphi_a^{(8)}(x_i - x_j) \\ \hat{R}_b(f^{(2)}) &= \frac{1}{n^2}\sum\sum_{i,j}\varphi_b^{(4)}(x_i - x_j) \\ \hat{C} &= \left(\frac{21}{8\sqrt{\pi}\hat{R}_a(f^{(4)})}\right)^{\frac{1}{9}}\left(\frac{1}{2\sqrt{\pi}\hat{R}_b(f^{(2)})}\right)^{\frac{2}{5}} \end{aligned}$$

Find:

$$\hat{h} = \arg\min_h JMP(h)$$

Park and Marron Plug-In

A detailed description of this selector can be found in Park and Marron (1991), but note that in the line after equation (2.10) in Park and Marron (1991) it should read $C_3(K) = \left\{18R(K * K^{(4)})\sigma_K^8/\sigma_{K*K}^4 R(K)^2\right\}^{1/13}$.

With λ the interquartile range calculate:

$$
\begin{aligned}
a(h) &= \frac{189}{640\sqrt{2}}\lambda^{3/13}h^{10/13} \\
\hat{R}_{a(h)}(f^{(2)}) &= n^{-2}\sum_{i \neq j}\sum \varphi^{(4)}_{\sqrt{2}a(h)}(x_i - x_j) \\
PM(h) &= \left(\frac{1}{2\sqrt{\pi}}\right)^{1/5} \hat{R}_{a(h)}(f^{(2)})^{-1/5} n^{-1/5} - h
\end{aligned}
$$

Now find:

$$\hat{h} \quad \text{with:} \quad PM(\hat{h}) = 0$$

.

Sheather and Jones Plug-In

With λ the interquartile range and L the gaussian kernel calculate:

$$
\begin{aligned}
a &= 0.920\lambda n^{-1/7} \\
b &= 0.912\lambda n^{-1/9} \\
\hat{R}_a(f^{(2)}) &= (n(n-1))^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} \varphi^{(4)}_a(x_i - x_j) \\
\hat{R}_b(f^{(3)}) &= -(n(n-1))^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} \varphi^{(6)}_b(x_i - x_j) \\
\alpha(h) &= (6\sqrt{2})^{1/7}\left(\frac{\hat{R}_a(f^{(2)})}{\hat{R}_b(f^{(3)})}\right)^{1/7} h^{5/7} \\
\hat{R}_{\alpha(h)}(f^{(2)}) &= (n(n-1))^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} \varphi^{(4)}_{\alpha(h)}(x_i - x_j) \\
SJ(h) &= \left(\frac{1}{2\sqrt{\pi}}\right)^{1/5} \hat{R}_{\alpha(h)}(f^{(2)})^{-1/5} n^{-1/5} - h
\end{aligned}
$$

Now find:

$$\hat{h} \quad \text{with:} \quad SJ(\hat{h}) = 0.$$

References

Aldershof, B., Marron, J.S., Park, B.U., and Wand, M.P., Facts about the Gaussian Probability Density Function, Mimeo. (Department of Statistic, University of Chapel Hill, North Carolina, 1990).

Bean, S.J. and Tsokos, C.P., Developments in Nonparametric Density Estimation, *International Statistical Review,* **48** (1980) 267–287.

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

Bowman, A., A Comparative Study of Some Kernel-Based Nonparametric Density Estimators, *Journal of Statistical Computation and Simulation,* **21** (1985) 313–327.

Cao, R., Cuevas, A., and González–Manteiga, W., A comparative study of several smoothing methods in density estimation, under consideration (1992) .

Chiu, S.T., The effect of discretization error on bandwidth estimates, *Biometrika,* **78** (1991a) 436–441.

Chiu, S.T., Bandwidth Selection for Kernel Density Estimation, *Annals of Statistics,* **19** (1991b) 1883–1905.

Chiu, S.T., An Automatic Bandwidth Selector for Kernel Density Estimation, *Biometrika,* **79** (1992) 771–782.

Deheuvels, P., Estimation non paramétrique de la densité par histogrammes généralisés, *Revue de Statistique Applique,* **25** (1977) 5–42.

Devroye, L. , *A course in density estimation* (Birkhäuser, Boston, 1987).

Devroye, L. , The double kernel method in density estimation, *Ann. Inst. Henri Poincaré,* **25** (1989) 533–580.

Devroye, L. and Györfi, L., *Nonparametric density estimation: The L^1-view* (Wiley, New York, 1984).

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

Fan, J. and Marron, J.S., Best possible constant for bandwidth selection, *Annals of Statistics,* **20** (1992) 2057–2078.

Fan, J. and Marron, J.S., Fast implementation of nonparametric curve estimators, Mimeo. (Department of Statistic, University of Chapel Hill, North Carolina, 1993).

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

Fryer, M.J., A review of Some Non-parametric Methods of Density Estimation, *Journal of the Institute of Mathematics and its Applications,* **20** (1977) 335–354.

GAUSS, A matrix computing language, (Aptech Systems Inc., 26250 196th Place South East, Kent Washington 98042, U.S.A.).

Grund, B., Hall, P., and Marron, J.S., 1992, Mimeo. (Loss and Risk in Smoothing Parameter Selection).School of Statistics, University of Minnesota, 1992

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

Härdle, W., *Smoothing Techniques, With Implementations in S* (Springer, New York, 1991).

Härdle, W., Marron, J.S., and Wand, M.P., Bandwidth Choice for Density Derivatives, *Journal of the Royal Statistical Society, Series B,* **52** (1990) 223–232.

Härdle, W. and Scott, D.W., Smoothing by Weighted Averaging of Rounded Points, *Computational Statistics,* **7** (1992) 97–128.

Hall, P., Large Sample Optimality of Least Squares Cross-Validation in Density Estimation, *Annals of Statistics,* **11** (1983) 1156–1174.

Hall, P. and Marron, J.S., Extent to which Least-Squares Cross-Validation Minimises Integrated Square Error in Nonparametric Density Estimation, *Probability Theory and Related Fields,* **74** (1987a) 567–581.

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

Hall, P. and Marron, J.S., Lower bounds for bandwidth selection in density estimation, *Probability Theory and Related Fields,* **90** (1991a) 149–173.

Hall, P. and Marron, J.S., Local Minima in Cross-validation Functions, *Journal of the Royal Statistical Society, Series B,* **53** (1991b) 245–252.

Hall, P., Marron, J.S., and Park, B.U., Smoothed Cross-Validation, *Probability Theory and Related Fields,* to appear (1992) .

Hall, P., Sheather, S.J., Jones, M.C., and Marron, J.S., On optimal data-based bandwidth selection in kernel density estimation, *Biometrika,* **78** (1991) 263–269.

Hall, P. and Wand, M.P., Minimizing L_1 distance in nonparametric density estimation, *Journal of Multivariate Analysis,* **26** (1988a) 59–88.

Hall, P. and Wand, M.P., On the minimization of absolute distance in kernel density estimation, *Statistics & Probability Letters,* **6** (1988b) 311–314.

Hall, P. and Wand, M.P., On the Accuracy of Binned Kernel Density estimators, unpublished manuscript (1993) .

Jones, M.C., Discretized and Interpolated Kernel Density Estimates, *Journal of the American Statistical Association,* **84** (1989) 733–741.

Jones, M.C., The roles of ISE and MISE in density estimation, *Statistics & Probability Letters,* **12** (1991) 51–56.

Jones, M.C. and Lotwick, H.W., A Remark on Algorithm AS176: Kernel Density Estimation Using the Fast Fourier Transform (Remark ASR50), *Applied Statistics,* **33**

(1984) 120–122.

Jones, M.C., Marron, J.S., and Park, B.U., A simple root n bandwidth selector, *Annals of Statistics,* **19** (1991) 1919–1932.

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

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

Kooperberg, C. and Stone, C.J., A study of logspline density estimation, *Computational Statistics & Data Analysis,* **12** (1991) 327–347.

Mammen, E., A short note on optimal bandwidth selection for kernel estimators, *Statistics & Probability Letters,* **9** (1990) 23–25.

Marron, J.S., Automatic Smoothing Parameter Selection: A Survey, *Empirical Economics,* **13** (1988) 187–208.

Marron, J.S., Comments on a data based bandwidth selector, *Computational Statistics & Data Analysis,* **8** (1989) 155–170.

Marron, J.S., Bias in Bandwidth Selection, unpublished manuscript (1991) .

Marron, J.S. (1992), Discussion of: "The Performance of Six Popular Bandwidth Selection Methods on some Real Data Sets" by Sheather, *Computational Statistics,* to appear.

Marron, J.S. and Nolan, D., Canonical kernels for density estimation, *Statistics & Probability Letters,* **7** (1989) 195–199.

Marron, J.S. and Wand, M.P., Exact mean integrated squared errors, *Annals of Statistics,* **20** (1992) 712–736.

Park, B.U., Advances in Data-Driven Bandwidth Selection, *Journal of the Korean Statistical Society,* **20** (1991) 1–16.

Park, B.U., Kim, W.C., and Marron, J.S., Asymptotically Best Bandwidth Selectors in Kernel Density Estimation, *Core Discussion Paper $N°$9154* (1991) .

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

Park, B.U. and Marron, J.S., On the Use of Pilot Estimators in Bandwidth Selection, *Nonparametric Statistics,* **1** (1992) 231–240.

Park, B.U. and Turlach, B.A., Practical Performance of Several Data Driven Bandwidth Selectors, *Computational Statistics,* **7** (1992) 251–270.

Parzen, E., On estimation of a probability density and mode, *Annals of Mathematical Statistics,* **33** (1962) 1065–1076.

Rosenblatt, M., Remarks on some non-parametric estimates of a density function, *An-*

nals of Mathematical Statistics, **27** (1956) 642–669.

Rudemo, M., Empirical choice of histograms and kernel density estimators, *Scandanavian Journal of Statistics,* **9** (1982) P65–78.

Schmitz, H.-P., *Die zeitliche Invarianz von Einkommensverteilungen* (PhD Thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, 1989).

Scott, D.W., Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions, *Annals of Statistics,* **13** (1985) 1024–1040.

Scott, D.W. and Factor, L.E., Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimators, *Journal of the American Statistical Association,* **76** (1981) 9–15.

Scott, D.W., Tapia, R.A., and Thompson, J.R., Kernel density estimation revisited, *Nonlinear Analysis, Theory, Methods and Applications,* **1** (1977) 339–372.

Scott, D.W. and Terrell, G.R., Biased and Unbiased Cross-Validation in Density Estimation, *Journal of the American Statistical Association,* **82** (1987) 1131–1146.

Sheather, S.J., The Performance of Six Popular Bandwidth Selection Methods on some Real Data Sets, *Computational Statistics,* **7** (1992) .

Sheather, S.J. and Jones, M.C., A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society, Series B,* **53** (1991) 683–690.

Silverman, B.W., Kernel density estimation using the fast Fourier transform. Statistical Algorithm AS 176, *Applied Statistics,* **31** (1982) 93–97.

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

Stone, C.J., An Asymptotically optimal Window Selection Rule for Kernel Density Estimates, *Annals of Statistics,* **12** (1984) 1285–1297.

Tapia, R.A. and Thompson, J.R., *Nonparametric Density Estimation* (John Hopkins University Press, Baltimore, Maryland, 1978).

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

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

Terrell, G.R. (1992), Comment (on Sheather 1992 and Park and Turlach 1992), *Computational Statistics,* **7**.

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

Turlach, B.A., Discretization Methods for Average Derivative Estimation, under consideration (1993) .

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

XploRe, A computing environment for e**Xplo**ratory **Re**gression and data analysis, (Available from XploRe Systems, Institut für Statistik und Ökonometrie, Humboldt–Universität zu Berlin, Germany).