Progress in Data-Based Bandwidth Selection for Kernel Density Estimation

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

Date: 23 June 1992

Abstract. We review the extensive recent literature on automatic, data-based selection of a global smoothing parameter in univariate kernel density estimation. Proposals are presented in a unified framework, making considerable reference to their theoretical properties as we go. The results of a major simulation study of the practical performance of many of these methods are summarised. Also, our remarks are further consolidated by describing a small portion of our practical experience on real datasets. Our comparison of methods' practical performance demonstrates that improvements to be gained by using the better methods can be, and often are, considerable. It will be seen that achieving optimal theoretical performance (up to bounds derived by Hall and Marron, 1991) and acceptable practical performance is not accomplished by the same techniques. We put much effort into making good practical choices whenever options arise. We emphasise that arguably the two best known bandwidth selection methods cannot be advocated for general practical use; these are "least squares cross-validation" (which suffers from too much variability) and normal-based "rules-of-thumb" (which are too biased towards oversmoothing). A number of methods that do seem to be worthy of further consideration are listed. We show why our overall current preference is for the method of Sheather and Jones (1991). It is hoped that the lessons learned in this comparatively simple setting will also prove useful in many other smoothing situations.

Key words and phrases: Automatic methods, cross-validation, curve estimation, functional estimation, mean integrated squared error, normal mixture, oversmoothing, rates of convergence, scale estimation, smoothed bootstrapping, smoothing parameter.
1. INTRODUCTION

Automatic selection of a “good” value for the parameter(s) that controls just how much smoothing is done, based only on the data at hand, is an ubiquitous topic in statistical smoothing problems. Considerable progress has recently been made on understanding this question. This is especially so in the particular context of choosing the global smoothing parameter, here called the bandwidth, in kernel estimation of probability density functions. Indeed, developments have happened so fast that the collection of bandwidth selection methods in Silverman's (1986) fine introductory book on nonparametric density estimation and even those in Marron's (1989) recent survey devoted solely to smoothing parameter selection no longer tell anything like a fully up-to-date story. It is one purpose of this article to remedy this situation.

We also provide a summary of a major Monte Carlo evaluation of a number of kernel density estimation bandwidth selectors. These simulations make very clear the relative merits of a number of proposals. Moreover, we also present a comparison of methods' performances on real data examples to demonstrate, convincingly we feel, that what follows is not a purely academic exercise: differences between methods, and improvements to be gained by employing the better ones, can be, and often are, very considerable.

An extended outline of the structure of the paper will prove useful to set the scene. A little background and notation is set up in Section 2, including a brief discussion of the discrepancy measure (between true and estimated densities) that we choose to work with, and which defines, by minimisation, a notion of target "optimal" bandwidth. We review some of the "earlier" methods in Section 3. There, we will concentrate on three methods in particular which have gained a certain popularity in the literature; these procedures will be compared with the newer methods in our simulations and examples. One of these is the usual "default" procedure that alternative proposals are compared with, and which maintains a high profile in practice even though its performance is known to be suspect, namely "least squares cross-validation". Its poor performance, both theoretically and practically, manifests itself as
unacceptably high variability in selected bandwidths. The other two methods in this section are Scott and Terrell's (1987) "biased cross-validation" and the standard "rule-of-thumb" based on an assumed normal distribution. The rule-of-thumb is thought by many to work well enough to be recommended for general use (e.g. Bowman, 1985, Silverman, 1986). We will give evidence that this procedure dramatically oversmooths bimodal densities, and can be clearly improved upon for anything other than normal or near-normal target densities. This selection method produces bandwidth estimates with a low degree of variability, but with centering about the wrong quantity in general.

Sections 4 to 7 review much recent progress in obtaining improved bandwidth selectors. These improvements occur in both theory and (usually) practice. Starting from the cross-validatory methods of Section 3, the main idea resulting in further methods will be seen to be that of improving the quality of estimation of the risk function that we use. An initial, and important, step is the realisation that a more appropriate choice of "secondary" or "pilot" bandwidth involved in risk estimation than that used in Section 3 can be made. This is developed to produce the particular case of the algorithm of Park and Marron (1990) in Section 4.

From this starting point, further developments proceed in two general directions in Sections 5, 6 and 7. First, in Section 5, we modify Park and Marron's (1990) method. Our modification consists in reinstating certain constant terms that had previously been omitted, and using these to reduce bias in the resulting bandwidth (at no expense in variance). The resulting improvement turns out to be remarkably consistent over a wide range of situations and is adopted for all methods from that point on. In particular, this gives the algorithm of Sheather and Jones (1991).

Certain choices will have been made for a variety of options in arriving at the particular algorithms of Sections 4 and 5. These will be discussed, and further methods based on variations on these choices will be introduced, in Section 6. An asymptotic benefit of our small sample improvement of Section 5 exists, but it is not as large as is theoretically possible. In Section 7, we take the improved risk estimation idea further to provide a variety of methods which achieve theoretical bounds on the performance of any bandwidth selection procedure; this section overlaps with Marron (1991). For the most part, the excellent theoretical behaviour of these procedures is not entirely reflected in a similar quality of finite sample performance. A recent contribution by Chiu (1992) may be an exception to this statement.

Section 8 summarises some of the results of our extensive simulation study of the performance of many of the bandwidth selectors described in the earlier sections. Of course, no one method is universally best, but we are able to reach some tentative conclusions, summarised in our brief closing Section 11.
In Section 9 we make a practical reinforcement of some of the points made earlier by application to two real datasets selected from the larger collection of examples discussed in Sheather (1992). Then, in Section 10, we make brief mention of many related bandwidth selection problems to which the ideas in this paper have some relevance.

For another comparison of many of the methods discussed in this paper (and some that aren't) see also Cao, Cuevas and González-Manteiga (1992).

2. PRELIMINARIES

Given \( X_1, \ldots, X_n \), a sample of \( n \) independent observations from a common distribution with density \( f \), say, the kernel estimate of \( f \) is

\[
\hat{f}(x; h) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i).
\]

Here \( K \) is the kernel function which we take to be a symmetric probability density function and \( K_h(y) = h^{-1} K(h^{-1}y) \). Also, \( h \) is the bandwidth of the kernel density estimate that we wish to develop automatic selection procedures for.

Throughout, we use the mean integrated squared error (MISE) as our measure of discrepancy between \( f \) and \( \hat{f}(.; h) \):

\[
\text{MISE}(h) = E \int \{ f(x) - \hat{f}(x; h) \}^2 dx.
\]

(1)

Each of the letters M, I and S in MISE invites controversy! The I simply reflects our desire to estimate \( f \) globally, and hence to do a decent estimation job everywhere (although one might argue for including some weighting function in the integrand), rather than locally, concentrating on \( f \) at a point.

The latter is approachable by appropriate modifications to what we do here, and some related references are given in Section 10.

There are arguments for and against including the M at all, and these are summarised in Hall and Marron (1991a) and Jones (1991a); their recommendations to prefer the risk function to the loss function in practical terms are followed here. Very briefly, although there are appealing conceptual arguments for working with ISE, a purely data-based bandwidth selection procedure does not have the information to be expected to behave very well in those terms, while good performance on the average, and hence with respect to measures more like MISE, is a much more realistic target.
The $S$ for squared error, particularly in conjunction with the $I$ for integration, is also controvertible, and with good reason. Indeed, in 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". A competitor with certain appealing properties is the "$L_1$ error" propounded by Devroye and Györfi (1985), but although this eases some of the problems associated with squared error, it still falls short of providing entirely appropriate loss and risk functions for curve estimation. Furthermore, for most target densities, the lessons concerning comparisons of the bandwidth selectors considered here are much the same when MISE is replace by a suitable $L_1$ analogue. (For work on $L_1$ based bandwidth selection, see Hall and Wand, 1988a,b, and Devroye, 1989). Provision of measures properly reflecting perceptions of what constitutes a good density estimate - while preferably affording tractable calculations - remains an important open question. For now we continue with squared error loss, and in particular MISE, because for all the above it is not entirely unreasonable, and it does possess the great merit of tractability.

The (global) minimiser of MISE will be denoted by $h_0$, and it will be the "target" of any (MISE-based) bandwidth selection scheme in that it represents the bandwidth of the best possible kernel estimator in these terms. $h_0$ is not, of course, a practically available proposition itself. All the data-based bandwidth procedures in this paper can be thought of as proceeding by estimating the minimum of the MISE curve by (at least approximately) minimising an estimate of MISE, giving a data-based choice $\hat{h}$, say. Often here, we assess the worth of proposed $\hat{h}$'s by examining properties of their relative error, $\hat{h}/h_0 - 1$. As noted in Remark 3.7 of Park and Marron (1990), this concentration on bandwidth behaviour is enough to tell us about the performance of the associated density estimates as estimators of $f$ relative to that of $\hat{f}$ using $h_0$. Hall and Marron (1991a) greatly encourage pursuit of good $\hat{h}$'s in these terms by showing that the best possible rate of convergence of any data-based procedure to $h_0$ is of the very respectable order $n^{-1/2}$, sufficient smoothness of $f$ permitting. (Methods for achieving this optimal rate are discussed in Section 7).

As a general notation define the functional $R(x) \equiv \int x^2(x)dx$. Also, define $\sigma_K^2 = \int x^2K(x)dx$. Then a very useful, and quite standard, approximation of MISE (call it AMISE) is

$$AMISE = \frac{1}{4}\sigma_K^4h^4R(f'') + \frac{R(K)}{nh}$$

(e.g. Silverman, 1986, Section 3.3). AMISE is asymptotically the same
as MISE as $n \to \infty$ and $h = h(n) \to 0$ in such a way that $nh(n) \to \infty$. AMISE is easily explicitly minimised to give an equally well-known asymptotic representation for $h_0$. Call it $h_\ast$; it is

$$h_\ast = \left( \frac{R(K)}{\sigma_k^2 R(f'')} \right)^{1/5} n^{-1/5}.$$  \hspace{1cm} (3)

(The corresponding optimal AMISE and MISE therefore tend to zero at the rate $n^{-4/5}$). These approximations are extremely helpful in assisting an intuitive appreciation of the factors affecting smoothing performance (Silverman, 1986), and in particular for our purposes in making even more explicit the inevitability of the best possible choice of bandwidth depending on the true $f$. Asymptotically, we see that this dependence is via the functional $R(f'')$. One more thing to note for future reference is that

$$h_\ast/h_0 = 1 + O(n^{-2/5}).$$

We have no wish to obscure our development by explicitly stating all of the technical conditions required for the results quoted; for these, reference can be made to the cited source material. It can be assumed that conditions on $K$ are satisfied by the popular choices of ("second order" i.e. nonnegative) kernel, such as the many polynomial (mostly Beta) densities on finite support, and the Gaussian density. Higher order kernel versions of $f$ are not considered here because we do not view them as practically important, for the reasons given in Marron and Wand (1992). But we must say a few words about smoothness conditions on $f$. Expansion (2) assumes two bounded continuous derivatives of $f$ (plus $R(f'') < \infty$). This will be taken as a minimum requirement throughout (although one could also look at the effect on these bandwidth selectors of $f$ possessing less smoothness than this, as done, for example, for cross-validatory choice by van Es, 1992). Moreover, convergence rates quoted in what follows (often with associated constants) are those that hold in a "maximal smoothness" situation i.e. when as much smoothness as is necessary is assumed. For least squares cross-validation in Section 3, the requirement remains at $p = 2$ derivatives of $f$. Alternative methods need more. Loosely summarising further requirements, other methods in Section 3 possessing the same convergence rate as least squares cross-validation need $2^{1\frac{1}{4}}$ derivatives, where we mean that $|f''(x) - f''(y)| \leq M|x - y|^{1/4}$ for some $M > 0$ and all $x$ and $y$. The techniques of Sections 4 and 5 require 3 and $3^{1\frac{1}{4}}$ derivatives, respectively, while to achieve the $n^{-1/2}$ convergence of Section 7, $p = 4^{1\frac{1}{4}}$ is necessary. Convergence rates fall continuously as $p$ decreases from these levels towards 2, and there is consequently some reordering of performance as $f$ gets less smooth (but not, note, immediately the given $p$ fails to
hold). It is important to keep in mind that number of derivatives is only a very crude asymptotic approximation for "difficulty of estimation". In Section 8, we discuss estimation settings where bandwidth selection and density estimation are both "very difficult" although $f$ is infinitely differentiable.

We use the following method of establishing names for those bandwidth selection methods that figure in our reported simulation results and/or data examples. If the bandwidth estimator is based directly on minimising an estimated MISE function, we refer to that estimated function by an appropriate acronym in large capitals (e.g. BCV), and we denote corresponding bandwidths by the same acronym in small capitals (such as BCV). Some other bandwidth selections, which do not arise so directly from minimisation of estimated MISE, nonetheless acquire a similar acronym (without a corresponding estimated risk function). Whatever is the case, there will be an associated phrase printed in boldface which explains the genesis of each acronym.

3. SOME ESTABLISHED METHODS

Take $R(f)$ away from the exact MISE given by (1), leaving only the quantities that depend on $h$, namely

$$E\{R(\hat{f}(.; h)) - 2 \int \hat{f}(x; h)f(x)dx\}.$$ 

Minimising an unbiased estimate of this, $\text{MISE}(h) - R(f)$, gives rise to the first of the well established methods that we wish to consider in this paper. This is **Least Squares Cross-Validation**, due to Rudemo (1982) and Bowman (1984). The LSCV objective function is

$$\text{LSCV}(h) = R(\hat{f}(.; h)) - 2n^{-1} \sum_{i=1}^{n} \hat{f}^{(i)}(X_i)$$ (4)

and, following our convention, the bandwidth that minimises LSCV is denoted LSCV. In (4), $\hat{f}^{(i)}$ means the kernel estimate based on the $n - 1$ datapoints other than $X_i$, in the usual cross-validatory manner. It is a simple matter to check that LSCV($h$) is indeed unbiased for $\text{MISE} - R(f)$, **Least squares** cross-validation was pre-dated, unsurprisingly, by **likelihood cross-validation**, but the latter will not be considered here since it has been shown to have some undesirable properties in a global density estimation context (which is not to say that it might not have value elsewhere, based as it is on the Kullback-Liebler loss function) (e.g. Silverman, 1986, Section 3.4.4, Hall.)

However, LSCV turns out to be a rather disappointing method (despite its practical popularity). Unbiasedness of objective function estimation is one thing, but it comes at the expense of excessive variability — and these properties transfer directly to LSCV itself. Indeed, as Hall and Marron (1987a) and Scott and Terrell (1987) first showed, it is the case that

\[ \text{LSCV}/h_0 = 1 + O_p(n^{-1/10}). \]

So, although LSCV shares with the vast majority of methods considered herein the property of consistency, it converges to the target value \( h_0 \) at a (relative) rate of \( O(n^{-1/10}) \) which is usually described as being extremely slow. The relative MSE of \( \text{LSCV} \) (of \( O(n^{-1/5}) \)) is dominated by its variance term. (It may be of interest to point out that if \( h_0 \) were replaced by \( \hat{h}_0 \) defined to be the minimiser of ISE, this same order \( n^{-1/10} \) convergence rate would hold for LSCV, Hall and Marron, 1987a, plus all the other consistent methods in this paper, and no method can do better in these terms, Hall and Marron, 1987b, 1991a; see Hall and Johnstone, 1992, for the latest on targeting \( \hat{h}_0 \).

In order to improve on LSCV, Scott and Terrell (1987) proposed an alternative scheme based on the asymptotic representation (2) instead of the exact MISE (1). Only \( R(f'') \) in (2) is unknown, and Scott and Terrell based their estimate of it on the kernel based estimate \( R(\hat{f}'(;h)) \), noticing that this gives rise to a positive constant bias term \( R(K'')/(nh^5) \) which could be subtracted away. The result they called Biased Cross-Validation:

\[ \text{BCV}(h) = \frac{1}{4}\sigma_K^2 h^4 S(\hat{f}'(;h)) + \frac{R(K)}{nh}, \]

(5)

where

\[ S(\hat{f}'(;h)) \equiv R(\hat{f}'(;h)) - \frac{R(K'')}{nh^5}. \]

Scott and Terrell (1987) compared BCV with LSCV. Theoretically, they showed that 1) BCV shares with LSCV the rate property that

\[ \text{BCV}/h_0 = 1 + O_p(n^{-1/10}), \]

but that 2) the constant multiplier of the leading term in \( E((\text{BCV}/h_0 - 1)^2) \) can be very much decreased compared with that of LSCV (Scott and Terrell quote a ratio of 24.8 in relative MSE terms using a “triweight” kernel, and Jones and Kappenman, 1991, obtain 15.7 for a normal kernel). Expressions for the constants involved in the limiting distributions of these bandwidths can be found for LSCV in Hall and Marron (1987a) and for both the above in

It turns out that many other bandwidth selection methods also possess order \( n^{-1/10} \) relative error rate of convergence to \( h_0 \), and Jones and Kappenman (1991) have investigated a number of them in a unified framework. One more of these will figure in our simulation investigations later; it was proposed by Taylor (1989), derived from a smoothed bootstrapping point of view, so we will refer to it as Taylor's Smoothed Bootstrap method. It will be more conveniently precisely defined in Section 6. This TSB turned out to beat even BCV in terms of its theoretical constant multiplier of \( n^{-1/5} \) in MISE, but see the simulation results of Section 8. Also included by Jones and Kappenman were versions of some of the earliest "plug-in" ideas in the literature (which, loosely, correspond to popping estimates of unknown quantities into (3) in much the same way as BCV does with (2)) plus one or two novel suggestions. Faraway and Jhun's (1990) method is closely related to Taylor's. Stute's (1992) method is also in this class.

We take a slightly different tack to produce the final well-established method that we include in our simulations. Rather than continuing with the nonparametric estimation of unknown quantities in MISE-related expressions, a parametric approach at this stage has oft been recommended. Replacing \( f \) in (3) by a normal density with unknown scale parameter, \( \lambda \) say, (e.g. Deheuvels, 1977) leads to a Rule Of Thumb such as

\[
\text{ROT} = 0.9 \min(\hat{\lambda}_1/1.34, \hat{\lambda}_2)n^{-1/5}.
\]

This version of this idea is (3.30) of Silverman (1986); \( \hat{\lambda}_1 \) is the sample interquartile range, \( \hat{\lambda}_2 \) is the sample standard deviation, and a Gaussian kernel is in operation. While ROT undoubtedly has a most appealing simplicity, it remains only, as its name suggests, a "quick and dirty" method, erring (as we shall see later, and as its tailoring to the normal distribution suggests) on the side of oversmoothing (Bowman, 1985, demonstrates its generally good performance for unimodal densities). We note that if \( f \) is not a normal density, then ROT is not even consistent for \( h_0 \), and that even in the normal \( f \) case, \( \text{ROT}/h_0 = 1 + O_p(n^{-2/5}) \) because of the discrepancy between \( h_0 \) and \( h_* \).

Talking of oversmoothing, Terrell and Scott (1985) and Terrell (1990) derive upper bounds on \( h_* \) which, when estimated, give selection formulae closely related to the normal-based rule of thumb. Note, however, that these methods smooth slightly more than does ROT, with obvious consequences for performance relative to that of ROT demonstrated later. As upper bounds per se, limiting other bandwidth selection procedures from occasional overly high smoothing, these ideas may have a useful role. For interesting related
ideas concerning "one-sided inference about functionals of a density", see Donoho (1988).

4. PARK AND MARRON'S METHOD

Estimating the functional of \( f \) that is the MISE of \( \hat{f}(.; h) \) is a different problem from estimating the whole of \( f \) itself. This is particularly clear from the asymptotic representation of MISE given at (2) where it is seen that dependence of AMISE on \( f \) is through the particular functional \( R(f'') \). Now, in kernel-based estimation of such a functional, the value of the bandwidth, \( g \) say, most appropriate to estimating the functional will differ from \( h \), the bandwidth intended to be suitable for estimating \( f \) globally. Moreover, it is clear from formula (3) that best estimating \( R(f'') \) is equivalent to best estimating \( h \).

When \( S(\hat{f}''; g) \), defined under (5), is used to estimate \( R(f'') \), Hall and Marron (1987c) show that the optimal \( g, g_0 \) say, is of the form

\[
g_0 = C_1(K)D_1(f)n^{-2/13} \tag{6}
\]

for some coefficients \( C_1 \) depending on \( K \), and hence known, and \( D_1 \) a function of \( f \) (these are explicitly given in the cited work). Comparing (6) with (3), we immediately see that the optimal choice of \( g \) is an order of magnitude bigger than that of \( h \) (which is used, implicitly or explicitly, in place of \( g \) in the \( O(n^{-1/10}) \) methods above). Park and Marron (1990) utilise this in their successful bandwidth selector which takes the solution in \( H \) of the equation

\[
H = \left\{ \frac{R(K)}{\sigma_K^4 S(\hat{f}''; g_1(H)))} \right\}^{1/5} n^{-1/5} \tag{7}
\]

as what we shall call PM. This is, clearly, based on (3) with the estimate \( S(\hat{f}''; g_1(H))) \) given beneath (5) replacing \( R(f'') \). \( g_1(H) \) arises from noting that

\[
g_0 = C_2(K)D_2(f)h_*^{10/13} \tag{8}
\]

for appropriate \( C_2 \) and \( D_2 \), so that \( g_1 \) is defined to be the same function as in (8) only with \( H \) in place of \( h_* \) (as suggested by Hall, 1980, and Sheather, 1986). To complete implementation details, Park and Marron (1990) chose to estimate \( D_2(f) \) — which involves \( R(f), R(f'') \) and \( R(f''') \) — by taking \( f \) at this pilot stage to be a normal density with (unknown) scale \( \lambda \), and to estimate \( \lambda \) by either \( \hat{\lambda}_1 \) or \( \hat{\lambda}_2 \).

For now, let us mention only the theoretical performance attributable to this procedure. The effect of utilising (suitably estimated) \( g \) in place of \( h \) is
to raise the rate of convergence to $h_0$ from the version using $h$'s $n^{-1/10}$ to the result

$$\frac{PM}{h_0} = 1 + O_p(n^{-4/13})$$

(Park and Marron, 1990). (See Park, 1989, for updated smoothness requirements for this to hold). Between rates $1/10 = 0.1$ and $1/2 = 0.5$, $4/13 \approx 0.308$ is "about halfway".

5. IMPROVING MATTERS BY REINSTATING CROSS-TERMS

The $R(f'')$ estimate $S(\hat{f}''(.;g)) = R(\hat{f}''(.;g)) - R(K'')/(nh^5)$ can also be written

$$S(\hat{f}''(.;g)) = \frac{1}{n^2g^5} \sum_i \sum_{j \neq i} (K * K)^{(iv)} \left( \frac{X_i - X_j}{g} \right)$$

where the star denotes convolution and the superscript means the fourth derivative. (Notice that here, and throughout until Section 7, we happen to be using the same kernel $K$ for this estimation step as for the estimation of $f$). The omitted $i = j$ "cross-terms" make up $R(K'')/(nh^5)$ which, were they present, contribute a known positive amount to the bias; deletion of cross-terms is therefore an appealing idea, and one that brooks no argument when $g = h$ (driving points made by Falk, 1992). However, the bias due to the smoothing turns out to be negative, so the positive cross-term bias can be used to try to counteract the smoothing bias. We can do so by choosing the bandwidth $g$ to (approximately) cancel the two (or at least the cross-term with the leading term in the smoothing bias).

This trick is investigated theoretically for estimating integrated squared density derivatives in Jones and Sheather (1991) and incorporated into bandwidth selection in Sheather and Jones (1991). Sheather and Jones's method therefore differs from Park and Marron's (1990) by replacing the estimate $S(\hat{f}''(.;g))$ by its cross-terms-reinstated version

$$\frac{1}{n^2g^5} \sum_i \sum_j (K * K)^{(iv)} \left( \frac{X_i - X_j}{g} \right)$$

which is, of course, $R(\hat{f}''(.;g))$. The optimal $g$ differs from that in Section 4 too, and Jones and Sheather (1991) show that it should be a little larger again, being proportional to $n^{-1/7}$. By the analogous argument that follows from (8), we write $g_3(H)$, say, as

$$g_3(H) = C_3(K)D_3(f)H^{5/7}$$

(11)
for appropriate $C_3$ and $D_3$. As for PM in (7), SJ is defined as the value of $H$ that solves

$$H = \left\{ \frac{R(K)}{\sigma_K^2 R(\hat{f}''(.; g_2(H)))} \right\}^{1/5} n^{-1/5}. \tag{12}$$

Sheather and Jones (1991) chose to add a further estimation step. $D_3(f)$, which is a function of $R(f'')/R(f''')$, is also estimated by kernel methods, and the (optimal) bandwidths associated with this "first stage" pilot estimation are, in turn, estimated using a normal scale model. Putting the latter parametric assumption back a stage (a practice demanded by the theory of SJ) is also an option for PM, as suggested in Park and Marron (1990) and included in more recent implementations. However, in our wide simulation experience, these comparable versions of SJ and PM have (remarkably) consistently performed such that SJ does better than PM (if not always by a lot). Indeed, this behaviour (and that of cross-terms "in" versus cross-terms "out" in other techniques) has been so (close to) universally in favour of the cross-terms-in version that we are confident of the widespread advisability of this idea (Young, 1990, independently promotes the value of utilising the cross-term in another class of functional estimation problems; Gasser, Kneip and Köhler, 1991, also retain cross-terms in kernel regression bandwidth selection). We therefore assume a cross-terms-in version of all further methods to be described unless stated otherwise. Also, we do not bother to include PM in our reported simulation results at all, the reader being able to impute PM's performance as being just a little inferior to that reported for SJ. The closeness of PM and SJ for practical purposes will be indicated in Section 9.

Another, computational, advantage of SJ is that we have only extremely rarely found cases where equation (12) does not have a unique solution. PM is also fairly well behaved in this regard, although a spurious secondary solution has been observed a little more often. The driving force behind this difference is that SJ utilises an estimate of the positive quantity $R(f'')$ that is itself assuredly positive, while $S(\hat{f}(.; H))$ in PM can sometimes take negative values for small $H$. Broadly speaking, the algorithmic behaviour in terms of numbers of minima or solutions of equations of the "better" methods described here is also rather better than that of the $n^{-1/10}$ methods. LSCV, in particular, is well-known to suffer often from multiple local minima, a property investigated theoretically by Hall and Marron (1991b).

Of course, although SJ is based on a bias reduction trick, in doing so we had an eye on its variance behaviour also, and only recommend it because the overall MSE performance is enhanced by this bias operation. In fact, the relative error rate of convergence is improved just a little from PM's $n^{-4/13}$:

$$SJ/h_0 = 1 + O_p(n^{-5/14})$$
(5/14 ≈ 0.357). The constant multiplier of \( n^{-5/7} \) in the relative MSE of \( s_J \) can be much improved over that of \( p_m \), however; Sheather and Jones (1991) report a coefficient ratio of 0.503 when \( f \) and \( K \) are normal.

6. OPTIONS, OPTIONS

In developing \( p_m \) and \( s_J \), particular options have been taken (as yet apparently arbitrarily!) in a number of places; in this section, we will discuss some of these options, introducing one or two competing bandwidth selection methods as we go.

The methods of Sections 4 and 5 are based on AMISE rather than MISE itself. A direct attack on MISE is perhaps over-complicated, and the following “hybrid” approach is more appealing. It turns out that the asymptotic representation of integrated variance, \( R(K)/(nh) \), is a very good approximation to the finite sample variance. In fact, it is exactly equal to the integrated variance of \( \hat{f} \) plus \( n^{-1}\int (K_h * f)^2(x)dx \) which is of order \( n^{-1} \). The first term in (2) is not such a reliable estimate of the integrated squared bias, however (although it is certainly of some use!); see Marron and Wand (1992) for evidence of this. Therefore, Hall, Marron and Park (1992) suggest the approximate MISE function

\[
\int (K_h * f - f)^2(x)dx + \frac{R(K)}{nh}
\]

which can be estimated by

\[
SCV(h) = \int \{K_h * \hat{f}(.; g) - \hat{f}(.; g)\}^2(x)dx + \frac{R(K)}{nh}.
\]

Here, \( SCV \) stands for Smoothed Cross-Validation. See Hall, Marron and Park (1992) for more motivation for the method and the name, and Marron (1992) for discussion of the “smoothed bootstrapping” view of the technique.

Because of their asymptotic equivalence, the minimiser of \( SCV \) has the same asymptotic behaviour as the minimiser of the estimated AMISE

\[
\frac{1}{4}\sigma_h^4 h^4 R(\hat{f}''(.; g)) + \frac{R(K)}{nh},
\]

but has the practical appeal that it removes one level of asymptotic approximation. The \( O(n^{-1/10}) \) version of the minimiser of \( SCV \) using \( g = h \), and thus (rightly) replacing \( R(\hat{f}''(.; g)) \) by \( S(\hat{f}''(.; g)) \), is precisely Taylor’s (1989) bootstrap selector \( TSB \). The method we call \( SCV \) and investigate in our simulations is the \( O_p(n^{-5/14}) \) version of the minimiser of (13) arising from taking
to be optimal for the resulting bandwidth selector, writing \( g \) in terms only of \( K, f \) and \( n \) (and not as a function of \( h \)), estimating terms in \( f \) once more by kernel methods, and introducing a normal scale model only at the following “second stage” (this, except for the extra pilot kernel estimation step, is a cross-terms-in version of the method of Section 4 of Hall, Marron and Park, 1992). The SCV algorithm is also extremely closely related to that giving rise to Cao, Cuevas and González-Manteiga’s (1992) \( h_{BC} \), developed in Cao-Abad (1990).

Notice that for SCV we minimise an appropriate objective function, while for PM and SJ we chose to solve an equation related to a minimisation problem instead. The latter arises, in the case that \( g \) is taken to depend on \( h \), as a kind of approximate minimiser of estimated AMISE by ignoring \( g \)'s dependence on \( h \) and performing the (simple) minimisation over remaining \( h \) terms only (it is exact when \( g \) is independent of \( h \), of course). Minimisation is the only realistic version of this for SCV, but for AMISE-based methods — such as PM and SJ — minimisation and solving-the-equation are genuinely competing options. Interestingly, the values of \( g \) optimal in these cases differ slightly from one another (the optimal value for the equation solution is the one identical with best estimating \( R(f'') \)). Experience with both versions in simulations showed a distinctly better performance in general, however, by the methods based on (3) rather than (2), and that is why the versions of PM and SJ given in Sections 4 and 5 were adopted.

The dependence of \( g \) on \( h \) in (7) or (12), or indeed in a minimisation like that of (13), is another option to consider. In (7) and (12), we chose to use knowledge of the form \( h_\ast \propto n^{-1/5} \) and \( g_\ast \propto n^{-\alpha} \), for some appropriate \( \alpha \), to motivate taking \( g \propto h^{5\alpha} \). The analogous equation solving method when \( g \) was taken to be equal to \( h \) is due to Scott, Tapia and Thompson (1977) (again an \( n^{-1/10} \) method). But a much simpler direct plug-in approach — requiring no numerical work — results from leaving \( g \) as a function of \( n \) only, making no mention of \( h \), and estimating the coefficient of \( n^{-\alpha} \). We include a version of such a method in our simulations: again we use cross-terms-in estimation, do one further stage of kernel estimation of coefficients, finish with a normal scale model, and call the result DPI. When the distinction between \( g \) and \( h \) is not made here, we obtain the earliest of direct plug-in ideas, going back to Woodroofe (1970).

However, these two choices are particular instances of the general “bandwidth factorisation”

\[
g \propto n^p h^m.
\]

Best convergence rate achievement when solving equations requires \( p - \frac{1}{2} m = -\alpha \), so that the appropriate overall size of \( g \) is maintained at \( n^{-\alpha} \). Park (1989)
describes theoretically the impact of this (on constants only) when solving the equation (in the cross-terms-out case) and, in private communication, makes the interesting observation that the choice \( m = 1 \), so that \( g \propto n^{2/35} h \) in the cross-terms-in case, affords a coefficient which is independent of scale. Use of this factorisation in the solution of (12) is closely related to the innovative method of Gasser, Kneip and Köhler (1991) in the regression case. More recently, and with some influence from the work of Sheather and Jones (1991), Engel, Herrman and Gasser (1992) have investigated various versions of this methodology in the density estimation context. We therefore call the bandwidth described here EHG, noting that it appears to be most closely related to Engel, Herrmann and Gasser's (1992) \( h_3 \) rather than their preferred \( h_2 \) version; it is included in our simulation study. (For factorisation in the minimisation case, see Section 7).

There are two more important implementational questions to be considered. These are (i) how many stages of kernel-based estimation of pilot bandwidths are necessary prior to an eventual normal scale modelling, and (ii) how should one estimate that final scale parameter? Extensive experimental work continues on these questions. For the moment, we take a two-stage approach in the sense that we apply kernel estimates to estimate \( g \) as well as the objective function to do with \( h \) itself, but apply the normal scale idea to appropriate bandwidths associated with \( g \)'s functionals. We have found that this, generally speaking, works better than the single stage approach (use a normal scale model in specifying \( g \) itself) which is exclusively considered by Cao, Cuevas and González-Manteiga (1992). See Aldershof (1991) and Park and Marron (1992) for more on the question of how many estimation stages to take. See the last five paragraphs of Section 8 for discussion of appropriate scale estimation.

7. ROOT-N-CONVERGENT METHODS

The best relative error rate of convergence to \( h_0 \) of any method displayed thus far is SJ's \( n^{-5/14} \). Pushing the idea of better estimating MISE yet further allows us to reach what Hall and Marron (1991a) show to be the best possible convergence rate, that of \( n^{-1/2} \). Several methods achieve this by utilising higher order kernels within functional estimation. A higher order kernel, \( L \), of order \( k \), say, has zero moments \( \int x^j L(x) dx \) for \( j = 1, 2, \ldots, k - 1 \) and non-zero \( k \)'th moment \( \left( \int L(x) dx = 1 \right) \). Probability density kernels like the \( K \) we have been using until now are therefore of order 2, and higher order kernels must allow negativity over part of their support.
Hall, Marron and Park (1992) indicate that such an approach is possible in the SCV framework: simply employ a version of $\hat{f}(.; g)$ incorporating a higher order kernel $L$ together, of course, with an appropriately recalculated optimal $g$ in (13). It turns out that, provided a cross-terms-in version of things is adopted, one only need extend to having $L$ a fourth order kernel — the obvious "next choice" — to achieve $\sqrt{n}$-convergence. Slightly earlier, Hall, Sheather, Jones and Marron (1991) had described how to do the analogous trick in the totally asymptotic framework that most other methods employ. One extra element is then necessary, however. The next term (at least) in the asymptotic expansion of MISE, (2), which is of order $h^6$, must be added to it (and estimated). Alternatively — the version explicitly defined in Hall, Sheather, Jones and Marron (1991) — the corresponding extension of the asymptotic formula (3) for $h_0$, involving as it will a further term of order $n^{-3/5}$, can be utilised. Again, use a fourth order $L$ in estimating $R(f''')$ and a corresponding optimal value for $g$ (the $R(f''')$ term arising in the extra term can afford, it turns out, to be estimated less well). Call the resulting bandwidth HSJM; it does not appear in the simulations of Section 8, but is used in Section 9.

At much the same time, Chiu (1991a), taking an approach through the Fourier domain, paralleled both of these developments with what can be interpreted as versions of them using kernels of infinite order (and correspondingly more arbitrary "large" values for $g$).

The theoretical advantages of higher order kernels in this context do not, however, seem to carry through very well to practice; see Marron and Wand (1992) for similar experience for density estimation per se. For this reason, it is encouraging that the next $\sqrt{n}$-convergent method that we mention avoids the necessity to use higher order kernels. Instead, the method of Jones, Marron and Park (1991) returns to the bandwidth factorisation method (14) of Section 6, but in the context of minimising an objective function like SCV of (13). It turns out that, in contrast to the solve-the-equation case addressed earlier, using Factorisation Within Minimisation with the special choice of $m = -2$ i.e.

$$g \propto n^{-23/45} h^{-2}$$

(15)

so that overall $g \propto n^{-1/9}$, affords $\sqrt{n}$-convergence even using second order kernels! See also Kim, Park and Marron (1992). We have failed to get any intuitive feeling for why this peculiar choice of dependence on $h$ should have such an effect.

The resulting bandwidth choice, FWM, is included in our simulation study of the next section (the option of staying within the SCV context has been taken although, again, there are other possibilities). It is the only represen-
tative of this section that we consider there, simulations of Park and Turlach (1992) indicating that FWM performs rather better than HSJM at least. However, the full force of the theoretical advantages of this bandwidth selector are still not felt in practice. As with the other $n^{-1/2}$ methods, its performance can be very good indeed (in cases where $f$ has few features), but in other situations it has a tendency to oversmooth badly.

More recently, Chiu (1992) has amended his earlier approach to provide a method which has the novel properties of retaining $n^{-1/2}$ convergence while at the same time exhibiting good performance in practice. This is achieved by replacing his original method for, essentially, specifying $g$ with a method using least squares cross-validatory choice of the bandwidth associated with an infinite order kernel estimate of $f$. This fine contribution arose too late to be included in our simulation study, but Chiu (1992) provides useful simulation evidence that can easily be compared with ours. Loosely, Chiu's method appears to have practical performance in much the same ballpark as a number of the better ones we display.

On the theoretical side, some of these methods add a further $n^{-1}$ term to (relative) MSE, which is smoothing dependent, to the unavoidable term

$$\frac{4}{25} \left\{ \frac{R(f^{(iv)} f^{1/2})}{R(f'')} - 1 \right\} n^{-1}.$$

Fan and Marron (1992) have shown that the displayed formula is a lower bound on bandwidth selector performance, coefficient and all.

8. RESULTS OF A MONTE CARLO STUDY

In this section, we report some of the more important results of a major Monte Carlo study of these bandwidth selection methods. The target densities were the 15 normal mixture densities of Marron and Wand (1992), shown in Figure 1 (for formulae, see Marron and Wand's Table 1). Only finite normal mixtures were considered because these admit simple exact calculation of MISE. While the fact that all these densities are infinitely differentiable could be viewed as a limitation, it is seen that for reasonable sample sizes, a number of these do not behave at all as one would expect from the lessons learned through asymptotics in the infinitely differentiable case. This set of densities does a good job of covering the range from "very easy to estimate" (i.e. densities where one can expect to recover most of the features reasonably well with a data set of moderate size) to "very hard to estimate" (i.e. those where only huge data sets will give reasonable recovery of features). However.
it is important to note that they should not be considered as "representative" in that it is not sensible to average answers (and especially conclusions) over this set of densities. For such purposes we feel that this set of densities contains too many "hard to estimate" cases (note that in some other studies, e.g. Cao, Cuevas and González-Manteiga, 1992, and Park and Turlach, 1992, only "easy to estimate" densities are considered). Another important point about this set of densities is that it contains many members where the fixed bandwidth kernel estimator, treated in this paper, can be very substantially improved upon, say by a variable kernel method (Abramson, 1982) or a transformation-based method (Wand, Marron and Ruppert, 1991). However, we consider it appropriate to include such cases because these "high tech" methods are typically used after one has understood the need for them from studying simple kernel estimates. One still needs a decent bandwidth for this initial look.

* * * put Figure 1 about here * * *

In this study 500 samples of size \( n = 100 \) and \( n = 1000 \) were simulated from each density. A large number of bandwidth selection methods were tried, but only those that are well known and/or those that worked well are reported on explicitly here. These are the bandwidth selectors given appropriate acronyms (all but PM and HSJM) in the text above. Some of these will be given more attention than others, as appropriate. The Gaussian kernel was used as \( K \) throughout. Choice of "scale estimate" is discussed at the end of this section. All bandwidths, except ROT and DPI (which afford much simpler calculations), were chosen by grid search minimization (root finding for SJ), over a logarithmically spaced grid of 25 values from \( h_0/3 \) to \( 9h_0 \). The grid search (versus a usually faster iterative search method) is viewed as essential to handle, and also to understand the frequency and effect of, local minima (roots), which occurred rather frequently for some of the minimization methods. Indeed, better results for some methods are obtained by taking other than the global minimum (root) whenever multiple minima (roots) are present: for LSCV we always used the rightmost local minimum, while for BCV and TSB we obtained improved performance by choosing the minimum furthest to the left; for root-finders, we took the rightmost root. The reason for the logarithmic grid is discussed in Section 4 of Marron and Wand (1992). The range of bandwidths was wide enough to contain most of the bandwidths of interest, although in some cases there was no minimum (root), so the appropriate interval end point was taken as the result.

As a first step to understanding how these bandwidth selectors compare, consider Figure 2. This gives plots of kernel estimates of the densities of the
bandwidth estimates themselves on the log scale (logs are natural as indicated above, and base 3 was chosen because of the bandwidth grid used) for densities 7, 9 and 12. The five selection methods represented here are LSCV, BCV, ROT, SJ and FWM. Each density estimate shown is based on the 500 bandwidths obtained in the simulations. The bandwidth used for these estimates was the oversmoothed one of Terrell (1990) based on fixed variance. This is not unreasonable in view of the fact that each of the automatic bandwidths has a limiting normal distribution and so the distribution on the log scale will have little structure (but see Chiu, 1990, for more on limiting bandwidth distributions). The relative MISE, as a function of \( h \), is also included in each plot.

* * * put Figure 2 about here * * *

Densities 7 and 9 were selected because the relative performance of the selectors was representative of that for each of the "midrange densities", 5, 6, 7, 8 and 9 which we believe are the type people think about most often when considering use of a density estimate. Density number 12 is representative of what we saw for the "hard to estimate densities", 3, 4 and 10 to 15.

Note that in general the distribution of LSCV is centered correctly around \( h_0 \) in each plot, but it is much more spread than the other methods. On the other hand most of the other methods have at least some, and sometimes a lot, of rightward bias. This phenomenon of trading off variability for some bias appears very frequently for, and we speculate is an intrinsic part of, bandwidth selectors with fast rates of convergence. To compare bandwidth selectors, one approach is to plug each of the selected bandwidths into MISE, and average the values (the result of this is studied below). One can visually see the effect of this operation in these pictures: it is not surprising that LSCV is usually the worst in this sense for 7 and 9, because of its large variance, but is the best for 12, because its superior centering becomes more important.

These pictures tell an important part of the story, but it is by no means complete, because MISE (and also the bandwidth scale) is difficult to interpret (e.g. what have we learned from the statement "the MISE is half the size"?). A way of further understanding what is being seen in Figure 2 is given in Figure 3. Figure 3 shows what is happening for the resulting density estimates when using bandwidths appearing in various parts of the populations represented in Figure 2, in particular showing the estimates resulting from the data sets near certain quantiles of the bandwidth distributions. We have made many such plots, but the main ideas are conveyed by the ones appearing here. Note that for \( n = 100 \), the 10th percentile of the LSCV distribution represents gross undersmoothing, and the 90th percentile represents mild oversmoothing. On the other hand, the 10th percentile for SJ
is still undersmoothed, but not so badly, while the 90th percentile is again oversmoothed. Of course it is too much to expect any of the percentiles to be too close together, because there are limitations to “how much information is contained in the data”. However, these demonstrate a point we have seen in many such plots: LSCV is unacceptably variable in almost any sense. Also these plots give a feeling for which features of Figure 2 are important, and which are not.

** Figure 3 about here **

For \( n = 1000 \), in Figure 2, all selectors improve on their variability, which is to be expected. However, note that for LSCV this improvement is the least, which reflects its slow asymptotic rate of convergence, as discussed in Section 3. The \( n = 1000 \) part of Figure 3 shows the added variability is still important. While most selectors are often doing quite well at \( n = 1000 \), LSCV is still unacceptable.

The performance of BCV in Figure 2 is indicative of what we saw elsewhere: very erratic, with performance swinging from quite good to very poor. This is worst for \( n = 100 \). It performs quite well for density 7, but is severely oversmoothed for density 9. It has an interesting bimodal distribution for \( n = 1000 \) on density number 12. We have not been able to understand fully what drives the small sample performance of BCV. In view of this erratic behaviour and the fact that its large sample properties are poorer than most other methods, we also view this method as being generally unacceptable.

A striking feature of the performance of ROT is its very small spread in bandwidth distribution in all cases in Figure 2. This is because its only source of variability is the scale estimation. However, this bandwidth selector is also clearly unacceptable because it often has gross bias problems. Note that this bias does not improve with large \( n \), which fits with the above theory because this bandwidth is inconsistent. Figure 3 gives a feeling for the magnitude of this bias in terms of the resulting density estimators.

The overall behavior of SJ is perhaps the best we have seen. It is usually much less variable than LSCV, but is usually centered more accurately than the other methods. There are some interesting and somewhat surprising comparisons between SJ and LSCV for density 12. For \( n = 100 \) the distribution of LSCV is strongly bimodal, with the effect of each mode being shown in Figure 3 (take care that the layout of Figure 3(c) is rather different from that of Figures 3(a) and 3(b)). Note that the first mode is making an attempt to “estimate the spikes”, while the second mode is attempting to “ignore the spiky structure”, with a large bandwidth. The selector SJ consistently ignores the spikes, while LSCV is clearly undecided. The MISE curve in Figure 2 indicates that in the MISE sense, it is optimal to attempt to estimate
those spikes, although the curve rises very slowly to the right of the minimum indicating the MISE optimal bandwidth is not much better than some reasonably far to the right. Indeed, thinking visually about the effects of this in Figure 3, it is not clear that this is the “intuitively best bandwidth”. This effect becomes even stronger for \( n = 1000 \), where \( SJ \) is again clearly inferior to LSCV in the MISE sense, but the visual impression in Figure 3 is definitely in favor of \( SJ \) (i.e. no experienced data analyst would choose a bandwidth such as those chosen by LSCV, but that for \( SJ \) is quite believable). Again there is clear warning against tying lessons learned about smoothing parameter selection too closely to MISE (or to any other of the usual norms for measuring the difference between functions).

Next consider FWM. This method has the best asymptotic performance of those considered here. However, its simulation performance is disappointing. While it is very good in terms of spread of bandwidth distribution, it has too large a rightward bias. One explanation for this is based on the fact that FWM attempts to estimate MISE, while \( SJ \) approaches MISE only through the approximation AMISE. Wand and Marron (1992) show in their Section 4 that the approximation of MISE by AMISE is very poor on the right hand side, with MISE being much lower. Hence \( h_0 \) is typically larger than \( h^* \). Now since all of the modern bandwidth selection methods usually exhibit a bias towards the right, there is a cancellation of these two effects, which means that methods which target AMISE actually have some advantage. However, we do not feel this is the full explanation (based on trying other methods which also target AMISE).

Another way of studying the simulation results is to consider summaries, and compare across all 15 densities. As a starting point we summarized the overall MISE performance of a generic bandwidth selector, say \( \hat{h} \), through \( CRIT = \text{avg} \left( \frac{\text{MISE}(\hat{h})}{\text{MISE}(h_0)} \right) \). (Here, we recognise that it is arguable whether or not \( \text{MISE}(\hat{h}) \) is a meaningful quantity \textit{per se}; \( CRIT \) is, however, linked to the mean squared difference between \( \hat{h} \) and \( h_0 \) in a way which drives, for example, Remark 3.7 of Park & Marron, 1990). Line graphs summarizing the values of \( CRIT \) for each of the 15 densities, and both \( n = 100 \) and \( n = 1000 \), are shown in Figure 4. We learned from Figures 2 and 3 that equally important summaries of the bandwidth population are “centerpoint” and “spread”, so also included in Figure 4 are line graphs summarizing the means and standard deviations (on the same log base 3 scale used in Figure 2).

*** put Figure 4 about here ***

Figure 4 demonstrates the overall consistent performance of \( SJ \). For \( n = \)
100, it has the smallest CRIT for most densities, except that ROT is slightly better for a few. The reason ROT is better is pure good luck, because the factor of 0.9, chosen quite arbitrarily, works out to just what was needed to make the normal reference bandwidth close to $h_0$. Note that ROT fares badly for a number of other densities, where this ratio is quite different, and is never better than SJ for $n = 1000$. The selector FWM is slightly better than SJ for the “easy to estimate densities” 1,2 and 5. This seems to be due to the empirical fact that FWM is more strongly affected by using the normal reference distribution in choice of coefficients than is SJ; when the distribution is not too far from the normal, there is a beneficial effect, but otherwise the opposite. The selector LSCV is better than SJ for the “hardest to estimate” densities 10, 12, 14 and 15.

For $n = 1000$, most of the overall lessons are the same. Again SJ gives the most consistent performance. A possible exception to this rule is the “hard to estimate densities” where LSCV is much better in terms of CRIT. However this needs to be discounted for the reason indicated above in the discussion of the density 12 part of Figure 3. Again FWM is better for the “close to normal densities”. An interesting comparison is available between $n = 100$ and $n = 1000$: for the “easy and midrange densities” most bandwidth selectors (except ROT, but that is inconsistent) give better CRIT performance for $n = 1000$. However, for the “harder densities”, only LSCV gets better in the sense of CRIT, and the others actually get worse! There is not a contradiction, because CRIT is relative with respect to MISE($h_0$), which also gets smaller for larger $n$. However, it indicates that for these densities, the “asymptotics have not kicked in yet”. In particular, the asymptotics indicate that CRIT should be converging faster (and towards 0!) for SJ and FWM. We believe these densities may be viewed as modeling “unsmooth densities”, where the fast rates of convergence of SJ and FWM break down (although LSCV maintains about the same convergence properties).

When a given selector does not work well, it is clear from the mean and standard deviation plots why not. The superb centerpoint, but terrible spread properties of LSCV are clearly seen to be very consistent. The erratic performance of BCV is seen to be sometimes caused by gross bias and sometimes by excessive spread. As indicated in Figure 2, when ROT works well, it is because it is consistently less variable than the other methods, and when it doesn’t it loses due to big bias problems. Also, these indicate that FWM is consistently less variable than SJ, but has consistently worse bias.

Having looked at results for the first 5 bandwidths we especially wished to concentrate on, we now briefly report an overview of the Monte Carlo results for 4 further bandwidths, TSB, SCV, EHG and DPI. We constructed an analog of Figure 4 for these selectors, but have decided against taking the space for
this, because we felt the more interesting comparisons were between TSB and BCV, and between LSCV, EHG and SJ, as shown in Figure 5 below. Both DPI and SCV gave excellent performance on the "easy to estimate densities", but were too biased elsewhere. The bias of SCV was typically slightly worse, but its spread was usually very small. But in most cases both of these were either slightly, or else substantially, worse than SJ.

*** put Figure 5 about here ***

The bandwidth TSB, as implemented here, usually gives a performance quite similar to BCV, but always at least slightly worse. The poorer performance is always caused by slightly more bias, which we believe is due to the "lack of cancellation" in targeting MISE vs. AMISE, discussed above.

The selector EHG is in a crude sense in between LSCV and SJ. For $n = 100$, it shares both very large variability and also small bias with LSCV, and thus cannot be recommended. However, for $n = 1000$, its spread becomes much better, and it is quite comparable to SJ, except for density 10 where it gives much better MISE performance (but recall that this is not a particularly good basis of comparison for these spiky densities). While we cannot recommend EHG in general because of its performance at the smaller sample size, it is among the best we have considered for larger samples. Note that in their paper, Engel, Hermann and Gasser (1992) use a "stopped iterative approach" to the solution of the fixed point equation (their $h_2$), and report the results as fairly similar to SJ. Here we use the actual solution to the fixed point equation (more like their $h_3$), which we speculate gives a result which is less biased, but more variable. We have not implemented their stopped iteration method yet. Because of our grid search approach, we were able to notice a slight tendency towards occasional double roots (in roughly 1-2% of the cases), mostly for the easier to estimate densities.

Next we consider scale estimation, which is an important issue for SJ, DPI, FWM and SCV. Denote the usual sample standard deviation by SD, the interquartile range by IQR and the median absolute deviation from the median by MAD. The usual statistical folklore is that the SD is the most efficient "scale" for truly normal data, but IQR and MAD are more robust otherwise. However, the way in which these measures are being used here is quite different, and requires careful rethinking. This is illustrated in Figure 6, which shows the selector DPI based on each of these, on the scale estimator used in ROT (RO), as well as on two other scale measures D1 and SS developed in Janssen, Marron, Veraverbeke and Sarle (1992).

*** put Figure 6 about here ***

23
As one would expect, SD gives the best CRIT for the normal density, number 1, but is a disaster for the outlier density, number 5. Also as expected, IQR and MAD are less efficient at the normal, but much better for the outlier density. For the remaining densities, IQR and MAD work better than SD only for the very nonnormal unimodal densities, numbers 3 and 4, and SD works better for all of the others. This is explained by consideration of the theoretical standard deviation, interquartile range, and median absolute deviation for these densities. For numbers 3 and 4, one expects IQR and MAD to be relatively smaller than SD, and vice versa in the other cases. A smaller scale estimate results in a smaller value of DPI, which means less bias towards large bandwidths in the mean plots, hence a smaller value of CRIT. A consequence of this is that none of the classical scale estimates is always appropriate, and surprisingly that the robust methods are often inferior, in situations where the density has structure as simple as two distinct modes.

The problem is that all of these scale measures are measuring a "global type of scale", which can be quite inappropriate for the situation at hand. To understand this, consider the separated bimodal density, number 7. The notion of "scale" that is important for bandwidth selection is the width of each peak, but the classical scale measures instead are concerned more with the distance between peaks. This effect can made arbitrarily dramatic by moving the two peaks apart, as done in Janssen, Marron, Veraverbeke and Sarle (1992). In that paper, alternative measures of scale are also developed which are more "local" in character. The effects of two of these, D1 and SS, when used in DPI, are shown in Figure 6. These measures are usually more variable than the usual ones (a price paid for their "local" characteristics), but have much better bias properties. The better of these two is SS, and it is always somewhat better than all the classical scale measures, except for being mildly worse for the "easy to estimate densities" (because for these bandwidth bias is not such a crucial issue). An exception is that SS is much better than SD for number 5, because of the well known tendency for outliers to inflate SD. Note that SS is much better on densities 3, 4 and 7.

We have also made plots similar to Figure 6 to understand the effect of scale on SJ and FWM. For both of these selectors, the lessons were very similar to those for DPI. One important difference was that performance of SJ was less dependent on which scale measure was used. This appears to be related to the way in which the appropriate g relates to h: for SJ, the scale parameter is taken to the power 3/13 while for DPI and FWM, the relevant powers are 1 and 3, respectively.

Since the scale SS was usually slightly better, and sometimes much better, we used this as the scale measure in all the figures shown earlier in this section. Its definition is rather complicated: first, given a parameter J, define
the "minimal 2\beta spacing" \( RD1 = \min_{j=[n\beta]+1,...,n-[n\beta]} \{X_{(j+[n\beta])} - X_{(j-[n\beta])}\} \).

The first improved scale measure we mentioned above is \( D1 = RD1/\mu_1 \) where \( \mu_1 = 2\Phi^{-1}(\frac{1}{2} + \beta) - 4^{0.875}(n/25)^{-2/3} \). To improve on \( D1 \), let \( \kappa = 0.001943 + 0.2832\beta + 0.3803\beta^2 - 1.542\beta^3 \), for \( j = [n\beta] + 1, ..., n - [n\beta] \) define \( \bar{X}_{j,\kappa,\beta} = \text{Avg}(X_{(j+[n\beta])},...,X_{(j+[n\beta])}) \), and let \( C = 2\beta/(\beta + \kappa) \). Define the minimal smoothed 3rd difference \( RSD3 = \min_{j=[n\beta]+1,...,n-[n\beta]} \{X_{(j+[n\beta])} - C\bar{X}_{j,\kappa,\beta} + CX_{j,-\beta,-\kappa} - X_{(j-[n\beta])}\} \). We also need \( \hat{\omega}_1 = \left( \frac{\beta^2 \left( 1 - (\frac{\hat{\omega}_1}{\hat{\omega}_2})^2 \right)^{RD1}}{6\Phi(0)^2 RSD3} \right)^{1/2} \), and define \( D3 = \begin{cases} \hat{\omega}_1^{4/5} D1 & RSD3 > 0 \\ D1 & \text{otherwise} \end{cases} \). Finally, we set \( SS = \min(SD, D3) \).

See Janssen, Marron, Veraverbeke and Sarle (1992) for motivation of these definitions, and for the recommendation to use \( \beta = 2 \) which we follow here.

9. TWO REAL DATA EXAMPLES

Further support for some of the clearest conclusions of our work, together with further evidence that the contest between various bandwidth selection methods is not of academic interest only, is provided by applying the methods to a variety of datasets taken from the literature. Here, we briefly report on only two such examples. These are taken from a more detailed report of Sheather (1992) which considers a total of six data examples. (These six, in turn, comprise only a subset of our real data experiences).

In this section, we utilise six of the aforementioned bandwidth selectors. These are LSCV, BCV, ROT, PM, SJ and HSJM. The version of ROT used here is (3.29) of Silverman (1986), rather than (3.30). PM is implemented precisely as in Park and Marron's (1990) paper, while HSJM replaces the similarly performing FWM. The Gaussian kernel is used in all density estimates, and all are plotted on the same scale within each figure.

Example 1

The first of our two examples is a familiar one from the smoothing literature. It is the data on 107 eruption lengths (in minutes) of the Old Faithful Geyser. These data first appeared in Weisberg (1980) and are also in Silverman (1986). Separate plots of the kernel estimates obtained from each of the six automatic bandwidth selection methods are given in Figure 7. The bandwidths vary from 0.10 (for LSCV) to 0.63 (for ROT), and the resulting density estimates vary considerably too. That for LSCV appears to be undersmoothed in as much as several small modes extra to the two major ones identified by the other methods are present. Although SJ and PM differ
a little numerically, they give almost the same pictures; this similarity of quantitative conclusions between the two is a typical result in our practical experience. BCV provides a similar result too. ROT and HSJM each smooth rather more than SJ, PM and BCV, and, agreeing with Silverman's subjective choice of \( h = 0.25 \), it might be argued that they oversmooth somewhat. Notice, however, that any oversmoothing is not so severe as to destroy the bimodal nature of the estimate in this case.

* * * Figure 7 about here * * *

The value 0.10 for LSCV corresponds to a local minimum of the LSCV function. This is a case where a little rounding in the data has made LSCV tend to \(-\infty\) as \( h \to 0 \). This phenomenon — specific to least squares cross-validation and not shared to anything like the same extent by other methods — is, by now, well appreciated (Silverman, 1986, Chiu, 1991b).

Example 2

The second example we choose to discuss here is application of the above six automatic bandwidth selectors to a dataset of size 476 on the thickness of the paper on which certain Mexican stamps of 1872 were printed (Izenman and Sommer, 1988). Separate plots of five of the six resulting density estimates are displayed in Figure 8. The rounding in these data (measurements were given to the nearest micron) was such as to confound least squares cross-validation altogether: this time there was no local minimum away from \( h = 0 \), so no plot corresponding to LSCV is displayed. Here too there is a great difference between the trio of plots associated with SJ, PM and BCV, which are essentially identical; and those arrived at by ROT and HSJM. The latter both smooth away many of the modes indicated by the former. Of course, one cannot make definitive conclusions for real datasets because we do not know any "right answer". But we do feel quite strongly that the seven mode estimate favoured by SJ, PM and BCV is preferable to those involving greater smoothing. Reasons for this are given by Izenman and Sommer (1988). Those authors obtained plausible subject matter explanations for each of the modes apart from the smaller two towards the right-hand end of the plot. Moreover, a parametric test procedure (based on normal mixture theory) supported these same five modes, while a nonparametric test for multimodality favoured seven. We find it gratifying that the better bandwidth selection methods applied to these data do not smooth away what appears to be genuinely interesting structure, even if it might take a more sophisticated density estimate to tell us more about the status of the sixth and seventh (minor) modes.

* * * Figure 8 about here * * *
The selector BCV has performed very well for both of these two examples because we took the smallest local minimum of BCV(\(h\)). In each case, as well as BCV(\(h\)) \(\rightarrow -\infty\) as \(h \rightarrow \infty\), there is an alternative local minimum at a value of \(h\) which oversmoothes the data but which affords a considerably lower value of BCV(\(h\)); see Figures 2.3 and 4.3 of Sheather (1992). Examination of the BCV objective function itself thus might make this good choice of \(h\) rather surprising; in any case, we note that the simulation results imply that this smallest local minimum strategy will not always work well (because, we believe, such a smaller local minimum does not always appear).

10. RELATED BANDWIDTH SELECTION PROBLEMS

It may be useful to set the particular situation we are concerned with in a wider smoothing context. There are many smoothing parameter selection problems that differ from our chosen one in one or more ways, and an important point is that the principles applied here are usually applicable to them too: what is required is some handle on how a relevant loss or risk function relates to the unknown curve(s), and some hope of understanding how to estimate that unknown function reasonably well. (In the following, the references given are relevant to such an approach, and are not meant to be representative of the entire literature on smoothing parameter selection; indeed we largely avoid giving references to any of the myriad of papers dealing only with (modifications to) basic cross-validation ideas).

For example, similar general ideas are being developed for the important kernel regression case (Chiu, 1991c, Gasser, Kneip and Köhler, 1991, Härdle, Hall and Marron, 1992), nonparametric regression probably being an even more commonly met requirement in practice than density estimation (we deal with the latter partly because we consider it to be easier, and so lessons to do with bandwidth selection per se are more clearly learnt, and yet they should also be generalisable). Härdle, Hall and Marron (1988) is also very relevant in the kernel regression context. Spectral density estimation is another standard smoothing example (Park and Cho, 1991), as are hazard rate estimation and estimation of the intensity function of counting processes. Variations on the i.i.d. data situation are another way to go. Of importance here is the dependent data case (Györfi, Härdle, Sarda and Vieu, 1989). Censored data also comes to mind; Jones (1991b) looks at length biased data.

There are, of course, alternatives to kernel smoothing methods, and modifications to the basic “constant kernel” estimate designed to improve performance. On the latter, it is not clear whether variable kernels will submit to the current kind of approach, but semiparametric transformation methods
(Wand, Marron and Ruppert, 1991) are designed to allow constant kernel methods to carry over direct. The case of higher order kernels, although not dealt with explicitly here, causes no extra problems at least in principle (and has been explicitly considered in some of the papers cited earlier).

As we concentrate on global curve estimates, so others consider bandwidth selection for estimating \( f \) well locally (Müller, 1985, Sheather, 1986, Eubank and Schucany, 1992, Thombs and Sheather, 1992). If some particular aspect of the underlying distribution, quantified as a functional thereof, is really the object of prime interest, then the amount of smoothing done needs to be tailored to estimation of that functional (Hall, diCiccio and Romano, 1988, and Goldstein and Messer, 1992, give nice examples of this), and the bandwidth selection method adapted likewise. Density, or other curve, derivatives are sometimes of prime interest; Jones (1992) considers prospects for extension of our bandwidth selection ideas to such a context, as well as to two other situations mentioned in this section (see also Marron, 1986).

Finally, we have not even touched on bandwidth selection when boundary effects are taken into account, when we are dealing with the multivariate case (Wand and Jones, 1992), questions of obtaining confidence bounds, and so on, and so on.

11. A CONCLUDING PARAGRAPH

For all our considerable theoretical, simulation and practical experience with a wide range of automatic bandwidth selection methods, we are, of course, reluctant to be too strong in our conclusions about which methods to use in practice. Nonetheless, two particular conclusions that we believe are rather indisputable are that the bandwidths LSCV, i.e. least squares cross-validation, and ROT, the normal-based rule-of-thumb, both currently popular in practice, are not to be recommended for general application. We have discussed a variety of alternative methods, several of which prove to have rather better properties and therefore remain worthy of further consideration. Included in this latter class might be Park and Marron's (1990) PM, versions of Hall, Marron and Park's (1992) SCSV (including one due to Cao, Cuevas and González-Manteiga, 1992), Sheather and Jones's (1991) SJ, its relations DPI and EHG (Engel, Herrmann and Gasser, 1992), and the recent suggestion of Chiu (1992). But finally, if we were pressed to commend one single method for general purposes, our current preference is for SJ.
REFERENCES


FIGURE LEGENDS

FIG. 1. The 15 normal mixture target densities.

FIG. 2. Estimated distributions based on 500 simulations of the five bandwidth selection methods LSCV, BCV, ROT, SJ and FWM (on a log3 scale), for each of three target densities (numbers 7, 9 and 12) and for two sample sizes (100 and 1000), as indicated on the plots. The solid vertical line is at $h_0$ and the solid non-density curve denotes $MISE(h)/MISE(h_0)$.

FIG. 3. Density estimates for datasets corresponding to certain choices made by bandwidth selection methods: (a) for density 7 and $n = 100$, 10 density estimates near the 10th and 90th percentiles of LSCV, SJ and ROT distributions; (b) as in (a) except $n = 1000$; and (c) for density 12, 10 estimates near the 33rd and 67th percentile of LSCV’s distribution when $n = 100$ and near its median when $n = 1000$, 10 estimates near the median of SJ’s distribution for each of $n = 100$ and $n = 1000$, and 10 estimates near the median of ROT’s distribution when $n = 1000$.

FIG. 4. Values of GRIT and mean and standard deviation for the distributions of bandwidth selection methods of Figure 2 for all 15 target densities. The densities are ordered by value of GRIT for SJ. (a) $n = 100$ (b) $n = 1000$.

FIG. 5. As for Figure 4, except that ROT and FWM are replaced by TSB and EHG.

FIG. 6. Setup as for Figure 4, except that here we investigate the effect of scale estimation by looking at method DPI employing the six scale estimates SD, IQR, MAD, RO, D1 and SS.

FIG. 7. Kernel estimates obtained by bandwidth estimators ROT, HSJM, PM, SJ, LSCV and BCV, as indicated above the plots, for the Old Faithful geyser data.

FIG. 8. Kernel estimates obtained by bandwidth estimators ROT, HSJM, PM, SJ, LSCV and BCV, as indicated above the plots, for the Mexican stamp thickness data.
Figure 1

#1 Gaussian Density
#2 Skewed Unimodal Density
#3 Strongly Skewed Density

#4 Kurtotic Unimodal Density
#5 Outlier Density
#6 Bimodal Density

#7 Separated Bimodal Density
#8 Asym. Bimodal Density
#9 Trimodal Density

#10 Claw Density
#11 Double Claw Density
#12 Asym. Claw Density

#13 Asym. Db. Claw Density
#14 Smooth Comb Density
#15 Discrete Comb Density
Figure 2

#7 Separated Bimodal

- LSCV
- BCV
- ROT
- SJ
- FNM

MSE(h)/MSE(h_{min})

Log_3(n) - Log_3(h_{min})

n = 100

#7 Separated Bimodal

- LSCV
- BCV
- ROT
- SJ
- FNM

MSE(h)/MSE(h_{min})

Log_3(n) - Log_3(h_{min})

n = 1000

#9 Trimodal

- LSCV
- BCV
- ROT
- SJ
- FNM

MSE(h)/MSE(h_{min})

Log_3(n) - Log_3(h_{min})

n = 100

#9 Trimodal

- LSCV
- BCV
- ROT
- SJ
- FNM

MSE(h)/MSE(h_{min})

Log_3(n) - Log_3(h_{min})

n = 1000

#12 Asym. Claw

- LSCV
- BCV
- ROT
- SJ
- FNM

MSE(h)/MSE(h_{min})

Log_3(n) - Log_3(h_{min})

n = 100

#12 Asym. Claw

- LSCV
- BCV
- ROT
- SJ
- FNM

MSE(h)/MSE(h_{min})

Log_3(n) - Log_3(h_{min})

n = 1000
Figure 3a

67 Separated Bimodal

10 datasets, near the 10th percentile for $h_{\text{est}}$

$h_{\text{est}} = 0.2815$

$h_{\text{est}} = 0.2658$

$n = 100$

67 Separated Bimodal

10 datasets, near the 90th percentile for $h_{\text{est}}$

$h_{\text{est}} = 0.2515$

$h_{\text{est}} = 0.3356$

$n = 100$

67 Separated Bimodal

10 datasets, near the 10th percentile for $h_{\text{est}}$

$h_{\text{est}} = 0.2815$

$h_{\text{est}} = 0.2658$

$n = 100$

67 Separated Bimodal

10 datasets, near the 90th percentile for $h_{\text{est}}$

$h_{\text{est}} = 0.2515$

$h_{\text{est}} = 0.3356$

$n = 100$

67 Separated Bimodal

10 datasets, near the 10th percentile for $h_{\text{est}}$

$h_{\text{est}} = 0.2815$

$h_{\text{est}} = 0.5475$

$n = 100$

67 Separated Bimodal

10 datasets, near the 90th percentile for $h_{\text{est}}$

$h_{\text{est}} = 0.2515$

$h_{\text{est}} = 0.5975$

$n = 100$
Figure 3b

87 Separated Bimodal

10 datasets, near the 90th percentile for $h_{sev}$

- $h_{sev} = 0.1575$
- $h_{sev} = 0.1575$
- $n = 1000$''

10 datasets, near the 90th percentile for $h_{jel}$

- $h_{jel} = 0.1575$
- $h_{jel} = 0.1575$
- $n = 1000$''

10 datasets, near the 90th percentile for $h_{jel}$

- $h_{jel} = 0.1575$
- $h_{jel} = 0.1575$
- $n = 1000$''

10 datasets, near the 90th percentile for $h_{ntr}$

- $h_{ntr} = 0.1575$
- $h_{ntr} = 0.1575$
- $n = 1000$''

10 datasets, near the 90th percentile for $h_{ntr}$

- $h_{ntr} = 0.1575$
- $h_{ntr} = 0.1575$
- $n = 1000$''

10 datasets, near the 90th percentile for $h_{ntr}$

- $h_{ntr} = 0.1575$
- $h_{ntr} = 0.1575$
- $n = 1000$''
Figure 3c

#12 Asym. Claw

10 datasets, near the 33th percentile for \( h_{33} \)

\( h_{33} = 0.2015 \)

\( h_{35} = 0.1885 \)

\( n = 100 \)

#12 Asym. Claw

10 datasets, near the 50th percentile for \( h_{50} \)

\( h_{33} = 0.0678 \)

\( h_{35} = 0.0714 \)

\( n = 1000 \)

#12 Asym. Claw

10 datasets, near the 67th percentile for \( h_{67} \)

\( h_{33} = 0.2016 \)

\( h_{35} = 0.4914 \)

\( n = 100 \)

#12 Asym. Claw

10 datasets, near the 50th percentile for \( h_{50} \)

\( h_{33} = 0.0678 \)

\( h_{35} = 0.1350 \)

\( n = 1000 \)

#12 Asym. Claw

10 datasets, near the 50th percentile for \( h_{50} \)

\( h_{33} = 0.2016 \)

\( h_{35} = 0.4164 \)

\( n = 100 \)

#12 Asym. Claw

10 datasets, near the 50th percentile for \( h_{50} \)

\( h_{33} = 0.0678 \)

\( h_{35} = 0.2520 \)

\( n = 1000 \)
Summary of CRIT, n = 100
(ordered by CRIT for SJ)

Figure 4a

Summary of Means, n = 100
(ordered by CRIT for SJ)

Summary of SD's, n = 100
(ordered by CRIT for SJ)

LSCV
BCV
ROT
SJ
FWM

Density number

Standardized SDs

Standardized Means

log\(_{10}\)(CRIT)
Summary of CRIT, n = 1000  
(ordered by CRIT for SJ)

Figure 4b  Summary of Means, n = 1000  
(ordered by CRIT for SJ)

Summary of SD's, n = 1000  
(ordered by CRIT for SJ)

LSCV  
---

BCV  
---

ROT  
---

SJ  
-------------

FWM  
---
Summary of CRIT, n = 100
(ordered by CRIT for SJ)

Figure 5a

Summary of Means, n = 100
(ordered by CRIT for SJ)

Summary of SD's, n = 100
(ordered by CRIT for SJ)

LSCV
BCV
TSB
SJ
EHG
Figure 5b

Summary of CRIT, n = 1000
(ordered by CRIT for SJ)

Summary of Means, n = 1000
(ordered by CRIT for SJ)

Summary of SD's, n = 1000
(ordered by CRIT for SJ)

LSCV
BCV
TSB
SJ
EHG
Summary of CRIT, n = 100
(ordered by CRIT for DPSS)

Figure 6a
Summary of Means, n = 100
(ordered by CRIT for DPSS)

Summary of SD's, n = 100
(ordered by CRIT for DPSS)

DPSD
DPIQ
DPMA
DPRO
DPD1
DPSS